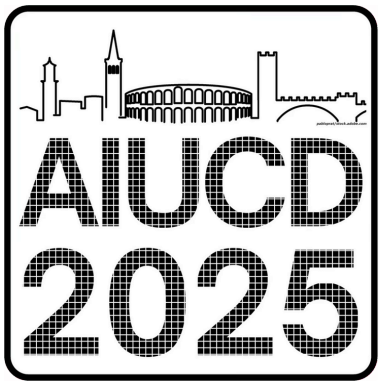


Atti del XIV Convegno Annuale

Diversità, Equità e Inclusione: Sfide e Opportunità per l'Informatica Umanistica nell'Era dell'Intelligenza Artificiale

Verona :: 11-13 giugno 2025

A cura di:
Simone Rebora • Marco Rospocher • Stefano Bazzaco

UNIVERSITÀ di VERONA
Dipartimento di LINGUE E LETTERATURE STRANIERE

daih

ASSOCIAZIONE per l'INFORMATICA UMANISTICA e la CULTURA DIGITALE

ISBN 978-88-942535-9-7

Copyright ©2025 AIUCD
Associazione per l'Informatica Umanistica e la Cultura Digitale

Il presente volume e tutti i contributi sono rilasciati sotto licenza Creative Commons Attribution ShareAlike 4.0 International license (CC-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). All other rights retained by the legal owners.

A cura di: Simone Rebora; Marco Rospocher; Stefano Bazzaco (2025). Diversity, Equity, and Inclusion: Challenges and Opportunities for Digital Humanities in the Age of Artificial Intelligence, Proceedings del XIV Convegno Annuale AIUCD, Verona 11-13 giugno 2025, Università di Verona.

Ultimo accesso agli URL in data 8 maggio 2025.

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

Il programma della conferenza AIUCD 2025 è disponibile online
https://aiucd2025.dlls.univr.it/detailed-schedule/
The AIUCD 2025 Conference Program is available online
https://aiucd2025.dlls.univr.it/en-gb/detailed-schedule/

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, effettuata dai membri del Comitato di Programma sotto la supervisione del Comitato Scientifico di AIUCD 2025.
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, carried out by the members of the Programme Committee under the supervision of the Scientific Committee of AIUCD 2025.

Gli atti del convegno AIUCD 2025 sono pubblicati come raccolta di contributi in formato PDF forniti direttamente dagli autori e dalle autrici. I file sono stati raccolti e assemblati senza interventi redazionali da parte dei curatori.
The proceedings of the AIUCD 2025 conference are published as a collection of PDF contributions provided directly by the authors. The files have been collected and compiled without editorial intervention by the editors..

Il logo di AIUCD 2025 include l'immagine "Verona Dark Line Simple Minimalist Skyline With White Background" di @pabloprat/stock.adobe.com, ottenuta tramite la licenza Adobe Stock dell'Università di Verona.
The AIUCD 2025 logo includes the image "Verona Dark Line Simple Minimalist Skyline With White Background" by @pabloprat/stock.adobe.com, used under the Adobe Stock license of the University of Verona.

Il background della copertina è stato creato con tecniche di AI generativa con lo strumento "Magic Media" disponibile su Canva, usando un prompt con il tema del convegno.
The background of the cover was created using generative AI techniques with the "Magic Media" tool available on Canva, using a prompt based on the conference theme.

Comitato Organizzatore / *Organizing Committee*

General Chairs
Simone Rebora (Università degli Studi di Verona)
Marco Rospocher (Università degli Studi di Verona)

Local Chair
Anna Cappellotto (Università degli Studi di Verona)

Registration Chair
Giorgia Pomarolli (Università degli Studi di Verona)

Proceedings Chair
Stefano Bazzaco (Università degli Studi di Verona)

Sponsorship Chair
Matteo Lissandrini (Università degli Studi di Verona)

Publicity Chair
Sabrina Piccinin (Università degli Studi di Verona)

Comitato Scientifico / *Scientific Committee*

Program Chairs
Simone Rebora (Università degli Studi di Verona)
Marco Rospocher (Università degli Studi di Verona)

Digital Humanities e inclusione / *Inclusive DH*
Stefano Bazzaco (Università degli Studi di Verona)
Massimo Salgaro (Università degli Studi di Verona)

Archivi ed Edizioni Digitali / *Archives and Digital Editions*
Elisa Cugliana (Cologne Center for eHumanities)
Christian D'Agata (Università di Catania)

Metodi Computazionali / *Computational Methods*
Rachele Sprugnoli (Università degli Studi di Parma)
Sara Tonelli (Fondazione Bruno Kessler)

Rappresentazione di Dati e Conoscenza / *Data and Knowledge Representation*
Francesco Mambrini (Università Cattolica del Sacro Cuore)
Elena Spadini (Universität Bern)

Preservazione della Memoria e del Patrimonio Digitale / *Preservation of Memory and Digital Cultural Heritage*
Monica Berti (Universität Leipzig)
Daria Spampinato (Istituto di Scienze e Tecnologie della Cognizione-CNR)

Comitato di programma / *Program committee*

Stefano Allegrezza (Università di Macerata); Laura Antonietti (Université de Versailles Saint Quentin en Yvelines (Université Paris-Saclay); Luigi Bambaci (École pratique des hautes études, PSL); Liborio P. Barbarino (Università di Catania); Nicola Barbuti (Università degli Studi di Bari Aldo Moro (Dipartimento di Ricerca e Innovazione Umanistica); Sofia Baroncini (Leibniz Institute of European History); Andrea Bellandi (Institute for Computational Linguistics (CNR); Mario A. Bochicchio (University of Bari, Dep.t of Computer Science); Andrea Bolioli (Independent researcher); Marco Bombieri (University of Verona); Paolo Bonora (Università di Bologna); Flavia Bruni (Università di Chieti-Pescara); Marina Buzzoni (Ca' Foscari University of Venice); Alberto Campagnolo (KU Leuven); Anna Cappellotto (Università di Verona); Emmanuela Carbé (Università di Siena); Vittore Casarosa (ISTI-CNR); Raffaele Cioffi (Università di Napoli Federico II); Fabio Ciotti (Università di Roma Tor Vergata); Vincenzo Colaprice (University of Turin); Giuseppe Consolo (Università degli studi di Napoli, Federico II); Elisa Conti (Università di Catania); Salvatore Cristofaro (CNR ISTC); Giulia D'Agostino (TU Darmstadt); Elisa D'Argenio (HUN-REN Hungarian Research Centre for Linguistics); Enrico Daga (The Open University); Stefano Dall'Aglio (Università Ca' Foscari Venezia); Marilena Daquino (University of Bologna); Mauro De Bari (University of Bari Aldo Moro); Angelo M. Del Grosso (CNR-ILC); Matteo Di Franco (Università di Napoli Federico II); Giorgio Maria Di Nunzio (University of Padua); Stefano Ferilli (University of Bari); Lorenzo Ferroni (Università degli Studi di Verona); Franz Fischer (Ca' Foscari Università Ca' Foscari); Greta H. Franzini (Eurac Research); Francesca Frontini (CNR-ILC); Daniele Fusi (VeDPH, Stuttgart University); Mariangela Giglio (University of Bologna); Tiago Luis Gil (University of Brasilia); Luca Giovannini (University of Potsdam); Milena Giuffrida (Università di Catania); Edmondo Grassi (Università Telematica San Raffaele Roma); Miryam Grasso (Università di Catania); Piergiovanna Grossi (Università di Verona); Fahad Khan (CNR-ILC); Michele Lacriola (Università di Siena); Maurizio Lana (Univ. del Piemonte Orientale); Federica Lazzerini (Università degli Studi di Torino); Eleonora Litta (Università Cattolica del Sacro Cuore, Milano); Dominique Longrée (ULiège); Diego Mantoan (University of Palermo); Anna Maria Marras (University of Turin); Cristina Marras (CNR); Pietro Mazzarisi (University of Trieste); Barbara McGillivray (King's College London); Federico V. Meschini (Tuscia University); Alessio Miaschi (Istituto di Linguistica Computazionale "A. Zampolli" (CNR-ILC), Pisa); Andrea Micheletti (University of Padua); Giulia Miglietta (Università del Salento); Paolo Monella (Università Kore di Enna); Johanna Monti (Università degli Studi di Napoli "L'Orientale"); Rossana Morriello (Università degli Studi di Firenze); Gloria Mugelli (ILC CNR); Serge Noiret (AIPH (Associazione Italiana di Public History) (European University Institute); Giuseppe Palazzolo (Università di Catania); Mafalda Papini (CNR-ILC); Enrico Pasini (UniTO/CNR-ILIESI); Giulia Pedonese (CNR (Istituto di Linguistica Computazionale "Antonio Zampolli"); Paola Peratello (Università Ca' Foscari Venezia); Federico Pianzola (University of Groningen); Chantal Pivetta (Lund University (Sweden)); Igor Pizzirusso (AIPH); Giulia Re (ILC-CNR); Giulia Renda (University of Bologna); Dario Rodighiero (University of Groningen); Roberto Rosselli Del Turco (Università di Torino); Enrica Salvatori (Università di Pisa); Emilio M. Sanfilippo (CNR); Eva Sassolini (CNR-ILC); Andrea Schimmenti (University of Bologna); Flavia Sciolette (CNR-ILC); Pietro Sichera (ILIESI-CNR); Daniele Silvi (Università di Roma 'Tor Vergata'); Giulia Speranza (University of Naples L Orientale); Francesco V. Stella (UNISI); Timothy Tambassi (Ca' Foscari University of Venice); Mirko Tavosanis (Università di Pisa); Francesca Tomasi (Università di Bologna); Simona Turbanti (University of Milano); Marco Venuti (Università di Catania); Gennaro Vessio (University of Bari Aldo Moro); Gabriele Vezzani (Università di Verona / RWTH Aachen University); Fabio Vitali (University of Bologna).

Enti organizzatori / *Organisers*

AIUCD; Università di Verona: Dipartimento di Lingue e Letterature Straniere; Digital Arena for Inclusive Humanities (DAIH).

Sommario

Prefazione

Simone Rebora, Marco Rospocher, Stefano Bazzaco I-II

Digital Humanities and Inclusion

PrevNet. A FAIR and inclusive resource for the study of preverbs in historical languages 2
Andrea Farina, Barbara McGillivray

Potential bias in AI: cultural representation and the marginalization of African art 11
Francesca Bignotti

Exploring data-driven narratives in Digital Humanities web-based projects: features and impact 18
Tommaso Battisti, Marilena Daquino

Il Glossario delle Infrastrutture di Ricerca (GIR) 24
Lucia Francalanci, Alessia Scognamiglio, Irene Falini, Pietro Restaneo, Giulia Pedonese, Alessia Spadi

Educational Impact of Storytelling and Data Visualisation in the Interpretation of Humanities Data 29
Giulia Renda, Marilena Daquino

IncluInstIT: Un nuovo corpus per lo studio di linguaggio inclusivo su Instagram 35
Irene Caiazzo, Giovanna Maria Dimitri, Liana Tronci

Per un'analisi della rivista Umanistica Digitale in ottica DEI 40
Rossana Morriello, Lucia Sardo

Carpe bias, quam minimum credula queries 46
Sabato Danzilli

Evaluating bias within an epistemological framework for AI-based research in the humanities 52
Sarah Oberbichler, Cindarella Petz

Semplificare la lettura dei manoscritti utilizzando tecnologie WEB interattive e interazioni «hover» 60
Giacomo Marchioro, Andrea Brugnoli, Francesca Carnazzi, Paolo Pellegrini, Edoardo Ferrarini

Accessibilità e inclusione per la documentazione del restauro: gli archivi del Centro Conservazione e Restauro «La Venaria Reale» 66
Stefania De Blasi, Edi Guerzoni, Chiara Pipino

Il patrimonio culturale digitale delle minoranze etniche: il progetto DIGICHer tra le comunità Sámi, ladine ed ebraiche in Europa 72
Matteo Cova, Eleonora De Longis

Fulfilling GEN-der AIms: do image-generating tools discriminate? An on-field study 78
Francesco Meledandri

From Bias Paralysis to Bias as a Category of Analysis. Introducing the Bias-Aware Framework 86
Mrinalini Luthra, Amber Zijlma

Supporting Children with Linguistic Vulnerabilities Through Advanced, Theory-Driven Technological Solutions: The TELMI Approach for Italian children with DLD and Children with Italian as L2 94
Arianna Compostella, Giulia Valcamonica, Mattia Gianotti, Matteo Secco, Silvia Silleresi, Fabrizio Arosio, Franca Garzotto, Maria Teresa Guasti

Tecnologie AI per la didattica 104
Gabriele Prosperi, Giulia Miglietta, Eleonora Miccoli, Mario Bochicchio

DEA - An Innovative Technological Tool for Personalized Linguistic Training for Italian Children with Developmental Dyslexia 110
Marta Tagliani, Maria Vender, Giulia Valcamonica, Giovanni Caleffi, Franca Garzotto, Denis Delfitto

Il ruolo delle Infrastrutture nella costruzione di un ambiente di ricerca inclusivo. Un modello di buone pratiche 118
Marta Caradonna, Nicola Giampietro, Roberta Bianca Luzietti, Monica Monachini, Valeria Quochi, Emiliano Degl'Innocenti, Alessia Spadi, Alessandra Caravale, Antonio D'Eredità, Paola Moscati, Giacomo Mancuso

Predicting Grammatical Cases in Slovenian Varieties in Italy: A Use Case from the LORIS 1.1 Language Assistant 124
David Bordon

Verso un futuro senza barriere: l'accessibilità dei documenti elettronici nell 'European Accessibility Act 129
Stefano Allegrezza

La gestione del nuovo sapere digitale contemporaneo. Scenari, criticità, sfide, prospettive 134
Nicola Barbuti

Archives and Digital Editions

Learner Corpus of Creative Writing: An interdisciplinary challenge 143
Ioanna Tyrou, Katerina Florou

Retrieval-Augmented Generation systems for enhanced access to digital archives 149
Michele Ciletti

Preserving Clarity: The MAGIC project approach to ancient manuscripts 156
Yahya Momtaz, Stefania Conte, Guido Russo

Digital Explorations of Historical Multilingual Practices. The Challenges of the HyperAzpilcueta Project 160
Manuela Bragagnolo, Marcus Pöckelmann, Polina Solonets, Andreas Wagner

Digitalizzazione di un fondo archivistico per la creazione di un centro di documentazione digitale 165
Dario Baldini

Risorgimento Digitale: Un progetto di hyperedizione per i testi risorgimentali. Le Noterelle di Abba come caso di studio 171
Vincent Mobilia

SpaceLat: La geografia della letteratura latina tardoantica 177
Riccardo Consolini

Verso l'edizione digitale del carteggio Canneti-Fiacchi 184
Chiara Manca, Fiammetta Sabba, Bianca Sorbara, Silvia Tripodi

L'edizione critica digitale della 'Scienza nuova' di Giambattista Vico in Scholarly Digital Edition 190
Alessia Scognamiglio, Roberto Evangelista, Manuela Sanna, Salvatore Prinzi, Stefano Veneroni, Chiara Aiola, Luca de Santis

L'edizione digitale del papiro P.Tor.Choach. 12 in collaborazione con il Museo Egizio di Torino 197
Chiara Senatore

Dal palcoscenico al digitale: modelli di data visualization per la valorizzazione dell'Archivio Teatro delle Marionette di Gianni e Cosetta Colla 202
Elena Radaelli

Metodologie computazionali per l'organizzazione di archivi nati digitalmente 208
Mariangela Giglio

Archivi digitali per la conservazione e valorizzazione del patrimonio culturale: il caso del Santuario della Madonna di Carufo 215
Caterina Ciccotti

Il progetto ArPeR. Per un Archivio dei periodici romaneschi 222
Martina Ludovisi

Un modello integrato per il Roman d'Alexandre del codice Correr 1493: annotazione linguistica e edizione critica digitale 228
Giacomo Costa, Simone Zenzaro

Descrivere la catastrofe: documentare la diegesi per la catalogazione di opere distopiche e post-apocalittiche 232
Luca Paolo Bruno, Valeria Stabile, Juan Scassa, Carmelo Caruso, Ludovica Pannitto

Un'applicazione pratica per l'edizione digitale di testi agiografici e calendariali 238
Luca Avellis

Embracing flexibility: new EVT features for critical editing, accessibility and inclusivity 244
Roberto Rosselli Del Turco, Davide Cucurnia, Marina Buzzoni

Riscoperte poliane: edizione digitale di un manoscritto inedito de Il Milione 251
Giulia Fabbris, Samuela Simion, Fabio Soncin

«Proximior perfectioni»: criticità e future prospettive del progetto Dante Limina 257
Elisabetta Tonello

TEI Encoding as a Unified Structure for Multilingual Digital Editions: The LeggoManzoni Case Study 264
Mariia Levchenko, Beatrice Nava, Ersilia Russo

Verso l'implementazione di un sistema di riconoscimento di allusioni al lessico dantesco nelle testimonianze del Lager: il caso d'uso in Voci dall'Inferno 270
Carla Congiu, Angelo Mario Del Grosso, Marina Riccucci

Il progetto RETI (REndering Texts and Images): metodologia e primi risultati 276
Chiara Barbero, Matteo Di Franco, Federica Lazzerini, Annamaria Persia

Edizione digitale ed autorialità plurima: quali sfide? 283
Stefania Tesser

Il corpus di prosa letteraria del progetto RIND (1830-1930). Assunti teorici e vincoli pratici 289
Stefano Ondelli, Pietro Mazzarisi

ZoneRW: verso un'integrazione con Kraken ed eScriptorium per il riconoscimento e la gestione avanzata delle regioni di interesse 297
Pietro Sichera, Angelo Mario Del Grosso, Laura Mazzagufo, Daria Spampinato

L'edizione digitale dei primi 16 Taccuini di Paolo Orsi 303
Giuseppina Monterosso, Andrea Bolioli, Elisa Bonacini, Gianmario Cattaneo, Dario Gonella, Anna Maria Marras, Salvatore Spina, Paola Venuti

Artificial intelligence vs human handwriting: annotating damaged manuscripts 308
Dumitru Scutelnic, Laura Gazzani, Paolo Pellegrini, Claudia Daffara

Modellazione, interoperabilità e riuso in DiScEPT 314
Tiziana Mancinelli, Hansmichael Hohenegger, Federico Boschetti, Angelo Mario Del Grosso, Eleonora De Longis, Gloria Mugelli

Navigating the Digital Transition: Lessons from a Hybrid Critical Edition Project 319
Elisa Bastianello, Reto Baumgartner

Human–LLM Synergy in Higher Education Publishing: Two ChatGPT Use Cases within Editorial Pipelines 327
Gianluca Pavani

Taming the Hydra: A Model for Textual Dynamics and Constellations of Goethe's Venetian Epigrams 334
Daniele Fusi, Matteo Zupancic, Franz Fischer, Claus Zittel

Computational Methods

The Influence of AI Tools on University Students' Writing Style: A Stylometric Analysis of Narrative Texts 343
Dimitris Bilianos, Katerina Florou

Usare i Large Language Model per l'analisi del testo narrativo: strategie di prompt engineering per il riconoscimento del discorso indiretto libero nella narrativa italiana 1830-1930 349
Aurora Argenzio, Fabio Ciotti, Anna Chiara Corradino

Historical GIS e metodologie digitali per una storia della copertura boschiva 357
Vincenzo Colaprice

Experiments on the Use of LLMs for the Translation of the Babylonian Talmud 363
Mafalda Papini, Davide Albanesi, David Dattilo, Emiliano Giovannetti, Simone Marchi

Metodi di allineamento testuale bilingue per un'edizione genetica digitale dei Mémoires di Carlo Goldoni 368
Matteo Zibardi

Eastern Law in Western Words: Analyzing Roman Legal Terminology in Medieval Charters 375
Tamás Kovács, Angelos Nikolaou, Johannes Laroche, Georg Vogeler

Il corpus del Digesto: approcci e metodi computazionali per la creazione di risorse linguistiche 379
Alessandra Cinini, Paola Marongiu, Eva Sassolini

Preliminary Results for the Explanation of Neural Network-based Handwriting Identification in Historical Manuscripts 386
Riccardo De Cesaris, Valerio Caravani, Arianna Pastorini, Serena Ammirati, Paolo Merialdo

From Documents to Data: Digital Technologies in the Study of Notarial Charters 392
Franziska Decker, Sandy Aoun, Giuseppe Consolo

Verso la svolta computazionale della critica dantesca 397
Fara Autiero, Vittorio Celotto, Gennaro Ferrante, Chiara Fusco, Sandra Gorla, Giuseppe Andrea Liberti, Mariangela Palomba, Serena Picarelli, Stefano Angelo Rizzo, Silvia Tripodi

Phylo-1-preview. Un modello T5-Base per l'emendazione dei testi antichi 404
Giuseppe Ferrara

«Glottolab: A Linguistic Adventure»: Lo sviluppo di un'attività gamificata incentrata sulla linguistica 411
Cecilia Cattaneo, Claudia Roberta Combei, Chiara Zanchi

Concordanze e NLP: idee, metodi e regole per l'applicazione alla lingua italiana 419
Pietro Sichera, Christian D'Agata, Giuseppe Palazzolo

Reverse Engineering Critical Apparatuses for HTR Ground Truth Creation: The Case of Kennicott's Collation of the Hebrew Bible 426
Luigi Bambaci, Nachum Dershowitz, Daniel Stökl Ben Ezra

Data and Knowledge Representation

Prototyping an Atlas of Early Modern English Drama: An Experiment on DraCor Data 435
Luca Giovannini, Andreas Wagner

ATLAS: A data model for describing FAIR Digital Humanities research outcomes 440
Chiara Martignano, Giorgia Rubin, Sebastiano Giacomini, Alessia Bardi, Marina Buzzoni, Marilena Daquino, Riccardo Del Gratta, Angelo Mario Del Grosso, Franz Fischer, Roberto Rosselli Del Turco, Francesca Tomasi

Cantautorato e Digital Humanities. Per una valorizzazione dell'opera di Fabrizio De André, Lucio Dalla, Gianmaria Testa 448
Marcello Ranieri

OWL Ontology on the European Integration Process between 1949 and 1979 453
Lorenzo Galvagno

LiITA, una Knowledge Base di risorse interconnesse per l'italiano 460
Eleonora Litta, Marco Passarotti, Paolo Brasolin, Valerio Basile, Cristina Bosco, Andrea Di Fabio

Dai limina a LiMINA: un database per i marginalia alla Commedia 466
Serena Malatesta, Beatrice Mosca

Dai Materiali Didattici alle Piattaforme FAIR: Costruire un'Infrastruttura di Training in H2IOSC 473
Giulia Pedonese, Francesca Frontini, Roberta Ottaviani, Federico Boschetti, Alessia Spadi, Lucia Francalanci, Alessia Scognamiglio, Pietro Restaneo, Antonina Chaban, Jana Striova, Laura Benassi

IlluminAI: un sistema di navigazione interattivo per i manoscritti miniati rinascimentali 478
Valeria Minisini, Giorgio Gosti, Bruno Fanini

Making Germanic Cultural Heritage accessible to students: a proposal for a case study 485
Chiara De Bastiani, Giulia Fabbris

/DH.arc Vocabularies: Making semantic artefacts more visible and accessible using SKOS 492
Laurent Fintoni

Modeling an Ontology for Heritage Science: Challenges and Key Strategies 499
Erica Scarpa, Riccardo Valente, Irene Rossi

Linked Open Data and IIIF for connecting manuscripts images with their transcriptions: a case study from the Veneranda Biblioteca Ambrosiana 505
Lorenza Talarico

Describing Monastic Iconography Using Semantic Data: A Preliminary Investigation 511
Sofia Baroncini, Francesco Mele

A Linguistic Knowledge Graph of Word Borrowings from Portuguese 519
Anas Fahad Khan, Ana Salgado

E.T and Visual Culture Ontology (ETVCO): Perspectives on Extraterrestrial Influence in Visual Heritage 524
Kaosaier Wusiman, Simone Casazza

Automating XML-TEI Encoding of Unpublished Correspondence: A Comparative Analysis of two LLM Approaches 531
Marco De Cristofaro, Daniel Zilio

Modelli e tecnologie integrate e innovative per una cittadinanza digitale equa e sostenibile 538
Cristina Marras, Vittoria Fabiani, Enrico Pasini, Lisa Reggiani, Pietro Sichera, Paolo Ongaro, Martina Rossi

A Case Study in Cultural Heritage: A System Linking Three Open Data Tools – Digital Philology for Dummies (DPhD), Edition Visualization Technology (EVT), and a Relational Database 544
Renato Caenaro, Chantal Pivetta

Modeling Intermediality and Interpretations in Contemporary Combinatory Literature: Revealing Il Giuoco dell'Oca by Edoardo Sanguineti 551
Enrica Bruno, Maria Francesca Bocchi, Francesca Tomasi

From Metadata to Storytelling: A Framework For 3D Cultural Heritage Visualization on RDF Data 558
Sebastian Barzaghi, Simona Colitti, Arianna Moretti, Giulia Renda

Between Text and Icon: Towards A Representational Model for Ekphrastic Relations 566
Maria Francesca Bocchi, Carlo Teo Pedretti, Fabio Vitali

Preservation of Memory and Digital Cultural Heritage

Life and Death of DH Projects: A Preliminary Investigation of Their Lifecycles in Italy 575
Erica Andreose, Giorgia Crosilla, Remo Grillo, Gianmarco Spinaci

Research on Street Art in the Digital Space 581
Aleksandra Tselikova

Motion Visualisation of Dancers' Performances 587
Giacomo Alliata, Loïc Serafin, Samy Mannane, Sarah Kenderdine

Entità in relazione: policies, soluzioni tecnologiche e modelli lessicali per un (eco)sistema informativo integrato 593
Herbert Natta, Michela Tardella, Eleonora Lattanzi, Gianluca Rossi, Roberta Maggi

Preserving and enhancing cultural heritage: the Digest project 600
Alessandra Cinini, Paola Marongiu, Eva Sassolini, Monica Monachini

The Staccioli Digital Archive: Using Knowledge Graphs to power digital art history catalogues and art exhibitions 608
Klaus Werner, Pietro Liuzzo, Alessandro Adamou

Fantàsime: Interactive Drama per la valorizzazione del Patrimonio Culturale 613
Maria Chiara Provenzano, Eleonora Miccoli, Mario A. Bochicchio

Soluzioni phygital e mediazione culturale: riflessioni digiteconomiche nell'era dell'IA 619
Nicola Barbuti, Mauro De Bari

MeMo: Una mappa letteraria digitale per la memoria del Mezzogiorno 627
Laura Giurdanella, Giuseppe Palazzolo, Bernardo De Luca, Fara Autiero, Marco Gatto, Sabatino Peluso, Concetta Maria Pagliuca, Andrea Schembari

The relationship between art and sound: An experiment on the engagement of the cultural tourist 633
Sara Benetti, Nicola Orio

Analisi RTI delle iscrizioni runiche del Leone del Pireo (Arsenale di Venezia) 639
Paola Peratello, Elisa Corrò

Torino anni Ottanta. Digitalizzazione del patrimonio documentario e ricostruzione virtuale delle mostre negli spazi pubblici e privati 646
Filippo Yahia Masri

Linguistica dei corpora e informatica umanistica per la valorizzazione plurilingue del patrimonio culturale: implementazione del progetto UniVOCIttà 651
Rita Gramellini, Valeria Zotti

«Il mio sommario dunque è tutto qui?» Per Franco Fortini 658
Emmanuela Carbé, Mariangela Giglio, Pietro Orlandi, Jacopo Maria Romano, Giulio Quaresima

Prefazione

AIUCD 2025, il XIV Convegno annuale dell'Associazione per l'Informatica Umanistica e la Cultura Digitale, è dedicato al tema "Diversità, Equità e Inclusione: Sfide e Opportunità per l'Informatica Umanistica nell'Era dell'Intelligenza Artificiale". Il convegno si configura come un'occasione privilegiata per riflettere su tematiche di cruciale rilevanza in un'epoca in cui l'integrazione tra tecnologia e discipline umanistiche appare sempre più necessaria. In un mondo in cui l'intelligenza artificiale sta ridefinendo i confini del sapere, diventa fondamentale interrogarsi su come guidare questa trasformazione secondo principi di inclusività ed equità. Il tema delle Digital Humanities in relazione all'inclusione richiama l'attenzione sull'importanza di una rappresentazione diversificata nel campo delle scienze umanistiche e sottolinea il potenziale degli strumenti digitali e computazionali nella democratizzazione dell'accesso al sapere e alla cultura.

Il Convegno AIUCD 2025 è organizzato dalla Digital Arena for Inclusive Humanities (DAIH), centro di ricerca interdisciplinare del Dipartimento di Lingue e Letterature Straniere dell'Università degli Studi di Verona. Istituito nel 2023 nell'ambito del progetto di eccellenza 2023–2027 "Inclusive Humanities. Prospettive di sviluppo nella ricerca e nella didattica delle lingue e letterature straniere", il centro rappresenta la naturale evoluzione del precedente progetto di eccellenza del Dipartimento, incentrato su "Le Digital Humanities applicate alle lingue e letterature straniere" (2018-2022). DAIH si configura come catalizzatore di collaborazioni interdisciplinari, con obiettivi di ricerca all'intersezione tra informatica – in particolare intelligenza artificiale – e discipline umanistiche, con un focus specifico sugli studi linguistici e letterari. Il centro promuove l'avanzamento della conoscenza scientifica e l'innovazione nell'ambito delle tecnologie digitali e dei metodi computazionali applicati agli studi umanistici, contribuendo alla costruzione di una società più equa, inclusiva e diversificata.

La call for papers del convegno AIUCD 2025 ha sollecitato contributi su cinque assi tematici, uno dei quali esplicitamente centrato sul tema dell'inclusione: Digital Humanities e inclusione, Archivi ed edizioni digitali, Metodi computazionali, Rappresentazione dei dati e della conoscenza, e Preservazione della memoria e del patrimonio digitale. Sono pervenute 126 proposte, di cui 26 presentate con il contributo di studiosi e studiose affiliati a istituzioni straniere, a testimonianza della crescente dimensione internazionale e interdisciplinare del convegno annuale.

Le proposte sono state valutate attraverso un processo di revisione a doppio cieco, gestito da un comitato di programma composto da 10 track chair e 94 revisori, oltre ai general chair. Ogni contributo è stato assegnato ad almeno tre revisori indipendenti, ottimizzando le preferenze espresse nella fase di bidding e il carico di lavoro complessivo. In totale, sono state raccolte 366 valutazioni, che hanno permesso ai track chair di ciascun asse tematico di formulare raccomandazioni informate ai general chair in merito all'accettazione e alla forma di presentazione dei contributi. A seguito del ritiro di sei proposte, le decisioni finali hanno portato alla seguente selezione:

- 65 contributi accettati per presentazione orale,
- 37 contributi accettati come poster,
- 18 contributi rifiutati.

Le proposte accettate sono così distribuite nelle varie track:

- Digital Humanities e inclusione: 21
- Archivi ed edizioni digitali: 32
- Metodi computazionali: 14
- Rappresentazione dei dati e della conoscenza: 21
- Preservazione della memoria e del patrimonio digitale: 14

Come nelle edizioni precedenti, il convegno è stato preceduto dal ciclo di incontri online "Aspettando AIUCD 2025", organizzato in collaborazione con l'iniziativa Digital Spritz del Dipartimento di Lingue e Letterature Straniere dell'Università di Verona, che ha ospitato gli interventi di Rachele Sprugnoli (Università Cattolica del Sacro Cuore), Susanna Allés-Torrent (University of Miami), Giulia Pedonese e Francesca Frontini (CNR-ILC).

Il programma del convegno è ulteriormente arricchito dalla partecipazione di due keynote speaker d'eccezione:

- Evelyn Gius (fortext lab, Institute of Linguistics and Literary Studies, Technical University of Darmstadt), con un intervento dal titolo "Measuring What Matters — or, The Temperature of Literary Texts",
- Viviana Patti (Dipartimento di Informatica, Università di Torino), con un intervento dal titolo "Absit iniuria verbis".

Concludiamo questa prefazione con un sentito ringraziamento a tutte le persone e gli enti che hanno reso possibile la realizzazione di AIUCD 2025. Un ringraziamento speciale va al Direttivo AIUCD, e in particolare alla presidente Marina Buzzoni, al segretario Paolo Monella e alla tesoriera Francesca Frontini, per il costante supporto istituzionale e la preziosa collaborazione in tutte le fasi preparatorie del convegno. Desideriamo esprimere la nostra gratitudine al Comitato di Programma, composto da 10 track chair e 94 revisori, per la disponibilità e l'accuratezza nella valutazione delle proposte, così come a tutte le persone che hanno risposto alla call for papers e partecipato al convegno, contribuendo a rendere il programma ricco e stimolante. Un grazie sentito va al Comitato Organizzatore, formato dai membri del consiglio direttivo DAIH, per l'instancabile impegno nella gestione di ogni fase dell'evento: dal coinvolgimento degli sponsor (Matteo Lissandrini), alla comunicazione e promozione dell'evento (Sabrina Piccinin), dalla gestione delle iscrizioni (Giorgia Pomarolli) al coordinamento delle attività logistiche locali (Anna Cappellotto). Ringraziamo inoltre il personale della segreteria amministrativa del Dipartimento di Lingue e Letterature Straniere e le studentesse dell'Università di Verona che, attraverso i loro tirocini, hanno offerto un prezioso supporto all'organizzazione. Infine, rinnoviamo la nostra gratitudine a tutti gli enti e le aziende che, in varie forme, hanno sostenuto AIUCD 2025 e ne hanno reso possibile la realizzazione.

Verona, maggio 2025

Simone Rebora, Marco Rospocher e Stefano Bazzaco

Digital Humanities and Inclusion

PrevNet. A FAIR and inclusive resource for the study of preverbs in historical languages

Andrea Farina[1], Barbara McGillivray[2]
[1] King's College London, United Kingdom – andrea.farina@kcl.ac.uk
[2] King's College London, United Kingdom – barbara.mcgillivray@kcl.ac.uk

ABSTRACT (ENGLISH)

The digitization of Classical texts and the development of corpus and computational methods has revolutionized the study of Ancient Greek and Latin. Despite their profound impact on verb semantics and syntax, preverbs (prefixes attaching to verbal bases, cf. It. *circum-* in *circumnavigare* 'sail around') lack a dedicated digital resource; *PrevNet* addresses this need with a focus on Ancient Greek and Latin preverbs and motion verbs. It combines digital humanities methodologies with insights from historical and cognitive linguistics, and socio-cultural studies to explore how these ancient languages conceptualized movement, space, and reality. Based on a manually annotated dataset of 35 texts (500,000+ tokens) from multiple genres and periods (8th century BCE – 2nd century CE), *PrevNet* systematically captures the meanings and functions of preverbed motion verbs, facilitating linguistic and cultural analysis. Designed as an inclusive FAIR resource, it serves diverse audiences across Classics, Linguistics, and Digital Humanities, while also addressing various educational levels, from high school to academia. The resource enables cross-linguistic comparison, promotes accessibility to digital tools, and advances the study of linguistic diversity and cultural thought in the ancient Mediterranean. *PrevNet* demonstrates the transformative potential of digital humanities to bridge gaps between disciplines, democratize knowledge, and support underrepresented audiences in research and education.

Keywords: preverbs; Classical languages; motion verbs; FAIR principles; online resources

ABSTRACT (ITALIANO)

PrevNet. Una risorsa "FAIR" e inclusiva per lo studio dei preverbi nelle lingue storiche

La digitalizzazione dei testi classici e lo sviluppo di metodi della linguistica computazionale e dei corpora hanno rivoluzionato lo studio del greco antico e del latino. Nonostante il loro profondo impatto sulla semantica e sulla sintassi dei verbi, i preverbi (prefissi che si attaccano alle basi verbali, es. It. *circum-* in *circumnavigare*) mancano di una risorsa digitale dedicata; *PrevNet* risponde a questa esigenza concentrandosi su preverbi e verbi di movimento in greco antico e latino. PrevNet combina metodologie delle digital humanities con approfondimenti di linguistica storica e cognitiva e studi socioculturali, per esplorare come queste lingue antiche concettualizzassero il movimento, lo spazio e la realtà. La risorsa si basa su un dataset annotato manualmente di 35 testi (oltre 500.000 tokens) provenienti da diversi generi e periodi (VIII secolo a.C. - II secolo d.C.). *PrevNet* cattura sistematicamente i significati e le funzioni dei verbi di movimento preverbati, facilitando analisi linguistiche e culturali. Progettato come risorsa FAIR inclusiva, si rivolge a un pubblico diversificato nei campi di studi classici, linguistica e digital humanities, affrontando inoltre vari livelli educativi, dalla scuola superiore all'università. La risorsa consente il confronto interlinguistico, promuove l'accessibilità agli strumenti digitali e avanza lo studio della diversità linguistica e del pensiero culturale nel Mediterraneo antico. *PrevNet* dimostra il potenziale trasformativo delle digital humanities nel colmare le lacune tra discipline, democratizzare la conoscenza e supportare pubblici sottorappresentati nella ricerca e nell'educazione.

Parole chiave: preverbi; lingue classiche; verbi di movimento; principi FAIR; risorse online

1. INTRODUCTION

Since Busa's (1974–1980) pioneering work, the digitization of Classical texts and the consequent creation of corpora have transformed the study of historical languages such as Ancient Greek and Latin and fostered the creation of new online resources. Linguistics, with its quantitative methods, has particularly thrived in this digital landscape. Tools and resources have been developed to facilitate morphological analysis (e.g., Crane, 1991; Kahane and Mueller, 2001; Passarotti, 2007a; Vatri & McGillivray, 2018), sentence structure (e.g., Passarotti, 2007b; Bamman & Crane, 2011; Celano, 2019; 2024; Gorman, 2020), and word meaning (e.g., Miller et al., 1990; Bizzoni et al., 2014; Bermúdez Sabel, 2019; Biagetti et al., 2021; McGilllivray et al., 2022). Efforts towards the harmonization and linking of many of these

linguistic resources for Latin have further been carried on by the ERC-funded project (2018–2023) "LiLa: Linking Latin" (Passarotti & Mambrini, 2021). Nonetheless, other areas of Classical studies have made, and continue to make, efforts to leverage digital tools to advance the discipline (e.g., Simon et al., 2012; Elliott, 2021 on geography; Parker et al., 2019 on papyrology; Sommerschield, 2020 on epigraphy; Holleran, 2021 on social history).

Despite this *mare magnum* of resources for Ancient Greek and Latin, all focus on general linguistic phenomena. Resources like the valency lexica in Passarotti et al. (2016) and McGillivray (2012; 2021), and treebanks (e.g., Bamman & Crane, 2011) provide broad coverage of verbal behavior but lack targeted analyses of specific linguistic phenomena and their impact on verbal meaning and syntax. Additionally, despite their invaluable insights, these resources often face diversity and inclusion challenges, appearing distant from real world applications, accessible only to specialists or linguists, and disconnected from broader cultural contexts. This restricts their potential to engage a wider audience, including those from diverse linguistic and cultural backgrounds, who could benefit from a more inclusive and applied approach to the study of ancient languages and their conceptualizations of movement.

This paper presents *PrevNet*, a new online resource dedicated to the study of preverbs in Ancient Greek and Latin, particularly in motion verbs. Preverbs are meaningful prefixes that attach to verbal bases and profoundly shape both the semantics and syntactic behaviors of verbs (e.g., *circum-* in It. *circumnavigare* 'sail around'). In contrast with existing lexica, *PrevNet* provides data about the specific meanings that preverbs have in the contexts in which they are used, alongside the linguistic changes they trigger in the verbs they are attached to. *PrevNet* combines digital and corpus methodologies with insights from historical and cognitive linguistics, geography, and socio-cultural studies. The resource is based on the annotation of a corpus of 35 texts (500,000+ words) from both languages, spanning different authors, periods (8th century BCE – 2nd century CE), and literary genres (see Section 4). Starting from linguistic analyses on preverbs, *PrevNet* captures diverse information on meanings of preverbs and preverbed motion verbs, facilitating a deeper exploration of how these ancient cultures conceptualized movement and reality. For instance, Ancient Greek's 45 preverbed forms of 'sail' versus Latin's 14 (Farina, 2021) reflect broader socio-cultural distinctions, such as Greek maritime expertise (Vázquez Hoys, 2006; Beaulieu, 2016) and Roman pastoral traditions (Jones & Sidwell, 1997: 185). *PrevNet* records many Ancient Greek preverbed forms of *pléō* 'sail', showing how useful they were in expressing subtle differences in meaning (e.g., *para-pléō* 'sail across'), with related examples. The resource interface allows users to understand that these concepts were either absent or expressed by preverbed forms of more generic motion verbs in Latin (e.g., *transeo* 'travel across', often used when crossing a river). These linguistic patterns align with broader Aquamotion verb studies (Lander et al., 2012), showing that cultures with stronger ties to the sea develop richer Aquamotion lexica (e.g., Bartolotta, 2020; cf. Section 2).

PrevNet also exemplifies a commitment to inclusion and accessibility within the Digital Humanities by creating a FAIR resource that bridges linguistic, socio-cultural, and geographical analyses of Ancient Greek and Latin preverbs. Designed to be accessible to diverse audiences in terms of background (e.g., Classics, Linguistics, Digital Humanities, Anthropology) and level (e.g., researchers in academia, university students, teachers/students in high schools), *PrevNet* challenges the traditional focus of linguistic tools, which is often exclusively aimed at domain experts, providing an intuitive resource for exploring the role of preverbs in shaping ancient Mediterranean thought and culture. This resource also promotes cross-linguistic comparison, fostering a more inclusive understanding of linguistic diversity across historical languages. Additionally, its scalable and reusable design ensures long-term accessibility for a wide range of communities, aligning with efforts to reduce educational disparities in access to digital resources.

PrevNet also bridges theoretical and empirical research, allowing both specialists and non-specialists to explore linguistic diversity and the socio-cultural contexts embedded within the systems of the ancient world. Being a scalable and reusable resource for advancing the study of preverbs across time, space, and linguistic boundaries, *PrevNet* demonstrates how Digital Humanities can empower underrepresented audiences, democratize knowledge, and enhance participation, meeting the core objectives of inclusivity and diversity in research, education, and public engagement.

2. LINGUISTIC CONTEXT

Preverbs are a fascinating linguistic category that has attracted significant scholarly attention (e.g., Booij and Van Kemenade, 2003). These prefixes not only possess semantic value themselves (e.g., Lat. *circum-* 'around'), but also modify the verbs they attach to, often reshaping their meaning and syntactic behavior. For example, while Lat. *navigo* 'sail' denotes general sailing, *circumnavigo* 'sail around' implies a specific

action requiring a mass of land as the object (see also It. *circumnavigare l'isola* 'sail around the island' vs. *navigare il mare* 'sail over the sea'). Preverbs also alter so-called aspectual properties, as seen in Lat. *edo* 'eat' versus *comedo* 'eat up' (Cuzzolin, 1995), where the latter denotes a completed action, paralleling similar constructions in German (*trinken* 'drink' vs. *austrinken* 'drink up'). Despite their prominence in many languages, preverbs have largely been studied through theoretical or small-scale analyses (e.g., Revuelta Puigdollers, 2014; 2016). A few quantitative studies derived from the growing availability of large digital corpus data (see Section 1) exist (e.g., McGillivray, 2009; 2013; 2014; Meini and McGillivray, 2010), highlighting a gap in diachronic and cross-linguistic perspectives.

Given the cross-linguistic diversity of preverbs and their prominence in many languages and language families (Indo-European, e.g. Latin, Sanskrit, Italian, English; Ugro-Finnic, e.g., Estonian, see Ackerman, 2003; Kartvelian, e.g. Georgian, see Harris, 2003), creating an online resource dedicated to their study has the potential to allow for broader cross-linguistic comparisons. This widespread phenomenon may highlight the cognitive and linguistic strategies for encoding nuanced spatial, temporal, and metaphorical relationships. Within this context, the study of motion verbs provides an ideal foundation. Universally present in languages due to their connection to everyday experiences, motion verbs offer insights into the cognitive and cultural processes underlying spatial conceptualization (e.g., Talmy, 1985; Feinmann, 2020). In her typology, Levin (1993) identifies seven categories of English motion verbs, ranging from inherently directed motion verbs (e.g., *go*, *arrive*) to manner-of-motion verbs (e.g., *run*, *float*), and motion using vehicles (e.g., *drive*, *sail*). This categorization underpins VerbNet (Kipper Schuler et al., 2009), an online resource for English verb classes. Similarly, Talmy (2000) explores typological patterns in motion verb encoding, while Bartolotta (2020) highlights the diversity of manner-of-motion verbs in Aymara, spoken in Bolivia, Peru, and Chile, which vividly describe nuanced ways of moving (e.g., *hithitha* 'walk dragging', *ccuycutha* 'walk head down'). Studying preverbs in motion verbs thus provides a window into how ancient and modern languages reflect cultural and cognitive strategies for encoding motion, enriching our understanding of linguistic diversity and its socio-cultural underpinnings.

3. CREATION OF THE RESOURCE

To build *PrevNet*, we started from Ancient Greek and Latin preverbs verbs of motion. A sample of 16 preverbs (e.g., Lat./AGr. *ab-/apo-* 'away', *cum-/sun-* 'together', *in-/en-* 'in') and eight motion verbs spanning different motion domains (e.g., Lat./AGr. *eo/eîmi* 'go', *navigo/pléō* 'sail', *volo/pétomai* 'fly', *curro/trékhō* 'run') has been selected for each language. The total number of verbs is 256, of which 205 are attested (e.g., *ab-* does not attach onto *venio*, *curro*, *navigo*, *gradior*, *no*, so verbs like **abvenio*, **abcurro*, etc. are not attested). Given the diachronic and comparative scope of this study, we have tried to create a balanced corpus according to language periods, literary genres and their importance within a specific language/culture. The latter factor led to a wider representation of a specific literary genre among all the text we possess (e.g., theatre in Greece, Untersteiner, 1984; the satire in Rome, Petersman, 1999). Moreover, both sub-corpora focus more on the Classical period of these languages, as the number of texts and the literary, historical, and cultural importance of this period cannot be ignored.

Overall, our corpus includes 35 texts for more than 500,000 word occurrences. The general rule governing our choice of texts was to build onto existing studies, especially considering that *PrevNet* aims to be linkable to other existing resources (see Section 1). For this reason, we chose texts that have already been considered in previous quantitative studies (e.g., Plautus' *Amphitruo* and *Mostellaria*, see McGillivray, 2009) and/or have been annotated in the Ancient Greek and Latin Dependency Treebank. All preverbed verbs (2,829 tokens) occurring in the corpus were manually annotated following 20 linguistic parameters including morphology, preverb, preverb meaning in context, verb class using VerbNet labels (see Section 2), literal/non-literal meaning of the resulting verb, meaning of the verb itself (Farina, 2024). The annotation was carried out using INCEpTION (Klie et al., 2018). Data was retrieved through computational methods, using a Python script specifically designed for retrieving data from INCEpTION. The resulting CSV file includes 62,238 cells of data, which constitute the foundation of *PrevNet*. To allow for greater accessibility and reuse of data (Farina et al. 2025), the dataset will be made available on Figshare as soon as *PrevNet* is launched.

4. INCLUSIVE DESIGN

When designing the resource, a major factor was to ensure that it could serve the needs of a broad range of users, well beyond the linguistics community which was the target user group of previous similar resources for ancient languages. To maximize usefulness for the broadest audience possible, we actively

contacted potential user groups and circulated a survey seeking their feedback. This not only helped promote the creation of the resource but also allowed users themselves to take part in the development process and boost its ultimate impact. The survey included various queries about respondents: career stage (e.g., high school students and teachers, undergraduates, postgraduates, PhD students, lecturers), fields of study (e.g., Classics, Linguistics, Digital Humanities, Cultural History), and their familiarity with preverbs and motion verbs (see Section 2). Respondents were also asked what type of content they would like to see populate *PrevNet* and how they would want to engage with it. In total, 92 participants across 16 different fields responded to the survey (see Fig. 1), with the majority coming from Classics (42.6%), Linguistics (21.7%), and Digital Humanities (16.3%). This broad feedback was invaluable in ensuring that *PrevNet* meets the needs and expectations of its varied user base. For instance, most of the respondents manifested their interest in references – intended as curated lists of books, articles, or other resources related to preverbs, motion verbs, and Ancient Greek/Latin linguistics – and case studies (see Section 5) – intended as detailed examinations of specific linguistic phenomena or examples, showing how preverbs and verbs are used in various contexts. Besides the survey, usability tests were conducted by the developers with a group of volunteers that helped refine the user experience. We have also planned workshops to introduce users to the resource and solicit further input about how to ensure *PrevNet* keeps evolving in relation to its users' needs.

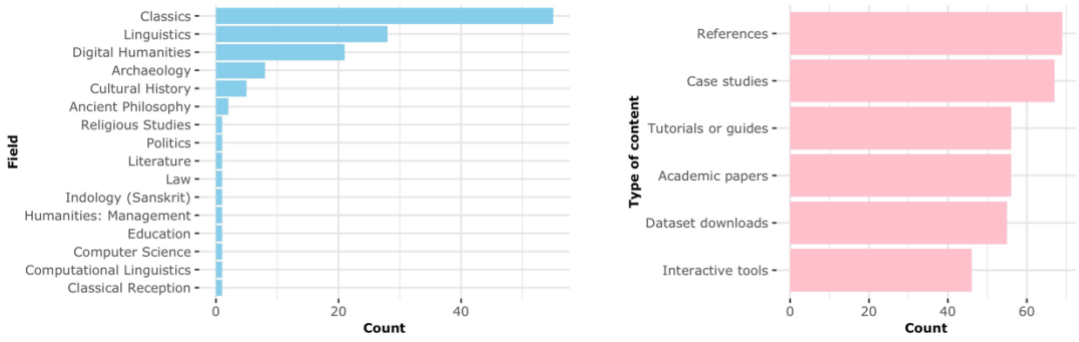


Figure 1. Results of the user survey question *What is your field of study?* (left) and *What type of content would you like to see in PrevNet?* (right)

The creation of *PrevNet* was a key component of the projects "Crossing Boundaries: Celebrating Linguistic Diversity at King's" (31st March 2025), funded by a Diversity & Inclusion Grant from the Faculty of Arts & Humanities at King's College London, and "From Classics to classrooms: enhancing linguistic understanding with digital resources" (February-December 2025), funded by a Public Engagement Small Grant Scheme at King's College London.

The first initiative brought together students from various departments within the Faculty of Arts & Humanities, such as Digital Humanities, and Languages, Literatures and Cultures. "Crossing Boundaries" was designed to draw on King's richly multicultural environment. The project organized a workshop with a group of students to explore linguistic diversity, specifically through the lens of how languages encode concepts of motion and spatial relationships across diverse cultures and historical periods. This allowed us to analyze the diversity of thought and perception across languages and cultures, fostering inclusivity and cross-cultural understanding. *PrevNet* allowed students to explore and compare motion verbs across different languages, showcasing how languages like English, Albanian, Chinese, Italian, and Latin uniquely represent movement, and revealing the cultural and linguistic underpinnings of these concepts. For instance, while English commonly uses verbs like *enter* or *exit* to express directional movement, other languages might convey the same ideas through subtle modifications to the verbs themselves (e.g., It. *circumnavigare* 'sail around') or display both options (e.g., Eng. *enter / go in(to)*). The workshop also encouraged collaboration across disciplines, with students from diverse linguistic and cultural backgrounds working together to compare motion verbs in their own language. Students also collaborated in the creation of "Paths of Motion", an art installation visually representing the diversity of linguistic approaches to movement. "Paths of Motion" was conceived as both a physical and digital artwork serving as a tangible and inspiring testimony of the project's commitment to celebrate linguistic diversity.

The project "From Classics to classrooms" was designed to make classical languages more accessible and engaging to secondary school students and teachers by leveraging *PrevNet* to explore linguistic diversity. Through interactive presentations and tailored teacher training sessions, the project fostered a new understanding of classical languages as living subjects that enrich modern linguistic perspectives.

Collaborations with schools, including Notting Hill & Ealing High School, ensured the resource's practicality and relevance. By equipping educators with innovative digital methods and involving students in hands-on linguistic analysis, the project not only revitalized interest in ancient languages but also empowered learners to appreciate the cross-linguistic connections between classical and contemporary languages. The effectiveness and impact of the project have been and will be evaluated through pre- and post-session questionnaires and follow-up focus groups, allowing for continuous refinement and ensuring long-term relevance.

5. *PREVNET*, A NEW NETWORK OF PREVERBS

PrevNet[1] is generated from linguistic annotations. It includes statistics associated with the linguistic parameters resulting from the manual annotation mentioned in Section 3. For instance, Fig. 2[2] shows the frequency of different verbal bases associated with the Lat. preverb *circum-* 'around' (left), alongside all the meanings associated with the resulting lemmas (*circumeo*, *circumvenio*, *circumvolo*) (right). All meanings are listed in the section "Examples" (bottom). Clicking on a specific meaning allows users to retrieve all annotated passages where any lemma appears with that meaning (Fig. 3).



Figure 2. *PrevNet* query for the preverb *circum-*. The resource shows interactive visualizations of the verbal bases occurring with *circum-*, providing their lemmas (left pie chart), and of the meanings associated with verbs preverbed with *circum-* (right pie chart), which are also given in the "Examples" section.

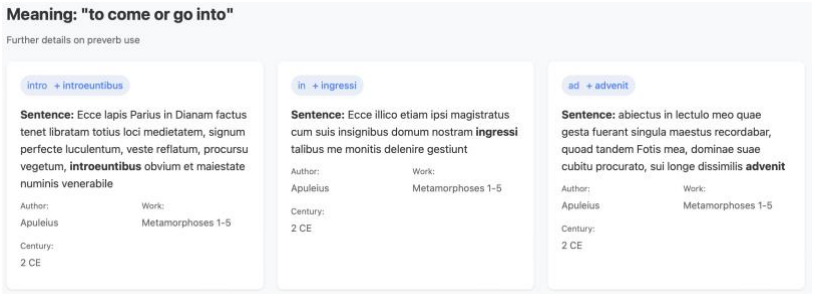


Figure 3. *PrevNet* query for the verb meaning "to come or go into". The resource lists all annotated passages where verbs with this meaning appear, regardless of their preverb.

Fig. 5 shows an example of the meanings of the preverb *cum-*, spanning from the prototypical 'together' to the grammaticalized[3] meaning 'completely', including negative implications in negative contexts (marked by 'malefactive'). The resource also shows the frequency of such meanings alongside a visualization to favor readability. All meanings are linked to their respective textual passages, making it possible to easily retrieve textual excerpts.

[1] *PrevNet* will be available at https://prevnet.sites.er.kcl.ac.uk. The resource is currently (April 2025) under construction, so for the moment access is restricted to developers only.
[2] Fig. 2 and Fig. 3 show screenshots from https://prevnet.sites.er.kcl.ac.uk (April 2025). Fig. 4 and Fig. 5 show mock-ups of *PrevNet*, as these functions are not yet available on *PrevNet* at the time of writing this article.
[3] Grammaticalization is the process by which a word – in this case, a preverb – with a concrete, specific meaning evolves into a grammatical element with a more abstract or functional role in language. For example, in the case of *cum-* it evolved from its original meaning 'together' to simply mark a completed situation (expressed with 'completely' in Fig. 2).

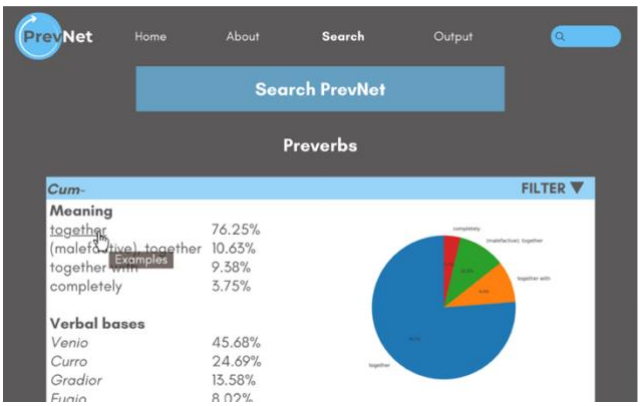


Figure 4. *PrevNet* mock-up query for the preverb *cum-*. The resource shows frequency percentages and visualizations, with the option of reading textual excerpts as examples

PrevNet also contains extra-linguistic information, such as places occurring with specific preverbed verbs. Using the World Historical Gazetteer (WHG) for geographical annotation, *PrevNet* cross-links to WHG for detailed place information. *PrevNet* offers the opportunity to visualize place names occurring with specific preverbs, linking them to the verbal bases they mostly co-occur with (e.g., the river Rhine always occurs with *transeo* 'cross', in Caesar's *De bello Gallico*), as well as to textual passages and directly to the WHG for deeper information about the history of the place itself and its visualization on a map (Fig. 5).



Figure 5. *PrevNet* mock-up query for the preverb *trans-*. The resource lists and ranks all the names of places occurring with the preverb *trans-*, with the option to view verbal bases connected to specific places, textual excerpts, or cross-linking to the World Historical Gazetteer

To enhance usability and in response to the users' demand (see Section 4), a section of comparative case studies is included in *PrevNet*. This way, users can unlock the potential of the resource, performing similar analyses for their own research and gaining insights not only into the development of metaphorical meanings but also into their socio-cultural implications. For instance, metaphorical meanings originating from preverbed basic motion verbs such as Lat. *eo/venio* and AGr. *eîmi/baínō* make for compelling comparative case studies (e.g., Lat. *evenio*, AGr. *sumbaínō*, *apobaínō* 'happen').
PrevNet adheres to the FAIR principles (Wilkinson et al., 2016). To make the resource findable and promote inclusion and interdisciplinarity, *PrevNet* is designed as a comprehensive repository containing information about preverbs and motion verbs across various languages, with metadata linking to academic papers on their linguistic and cultural significance, as also demanded by users (see Section 3). The resource is open access to eliminate barriers and interoperable with other existing resources (e.g., WHG), enabling seamless integration with broader digital infrastructures. Finally, the resource is reusable and designed to be expandable to additional (historical or modern) languages and verb classes over time. Open-access annotation guidelines are available for contributions and adaptations ensuring long-term utility and scalability (Farina, 2024).

ACKNOWLEDGEMENTS

Funding for this project has been received from the London Arts & Humanities Partnership (LAHP), the Faculty of Arts & Humanities at King's College London, and the Doctoral College and Centre for Research Staff Development at King's College London.

REFERENCES

Ackerman, F. (2003). Aspectual contrasts and lexeme derivation in Estonian: A realization-based morphological perspective. G. Booij, & J. van Marle (Eds.), *Yearbook of Morphology*, 13–31.

Bamman, D., & Crane, G. (2011). The Ancient Greek and Latin Dependency Treebanks. C. Sporleder, A. van den Bosch, & K. Zervanou (Eds.), *Language Technology for Cultural Heritage*, 79–98. https://doi.org/10.1007/978-3-642-20227-8_5

Bartolotta, A. (2020). On the typology of motion events in Aymara. *Atti Del Sodalizio Glottologico Milanese*, 1-24. https://doi.org/10.13130/1972-9901/15408

Beaulieu, M.-C. (2016). *The sea in the Greek imagination*. Philadelphia: University of Pennsylvania Press.

Bermúdez Sabel, H. (2019). Digital tools for semantic annotation: The WoPoss use case. *Bulletin de Linguistique et Des Sciences Du Langage*, 30, 12–37.

Biagetti, E., Zanchi, C., & Short, W. M. (2021). Toward the creation of WordNets for ancient Indo-European languages. *Proceedings of the 11th Global Wordnet Conference, University of South Africa (UNISA)*, *Global WordNet Association*, 258–266.

Bizzoni, Y., Boschetti, F., Diakoff, H., Del Gratta, R., Monachini, M., & Crane, G. (2014). The Making of Ancient Greek WordNet. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik (Iceland)*, 1140–1147.

Booij, G., & Van Kemenade, A. (2003). Preverbs: An introduction. G. Booij, & J. Van Marle (Eds.), *Yearbook of Morphology 2003*, 1–11. https://doi.org/10.1007/978-1-4020-1513-7_1

Busa, R. (1974-1980). *Index Thomisticus*. Stuttgart-Bad Cannstatt: Frommann-Holzboog.

Celano, G. G. A. (2019). The Dependency Treebanks for Ancient Greek and Latin. *Digital Classical Philology*, 279–298.

Celano, G. G. A. (2024). Opera Graeca Adnotata: Building a 34M+ Token Multilayer Corpus for Ancient Greek (Version 1). *arXiv*. https://doi.org/10.48550/ARXIV.2404.00739

Crane, G., Bamman, D., Cerrato, L., Jones, A., Mimno, D., Packel, A., Sculley, D., & Weaver, G. (2006). Beyond Digital Incunabula: Modeling the Next Generation of Digital Libraries. *Proceedings of the 10th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2006), Alicante (Spain), 2006*, 341–352.

Cuzzolin, P. (1995). A proposito di sub vos placo e della grammaticalizzazione delle preposizioni. *Archivio Glottologico Italiano*, 80(1–2), 12–143.

Elliott, T. (2021). The Pleiadic Gaze: Looking at Archaeology from the Perspective of a Digital Gazetteer. *Classical Archaeology in the Digital Age – The AIAC Presidential Panel*. https://doi.org/10.11588/PROPYLAEUM.708.C10612

Farina, A. (2021). *Aquamotion Verbs in Ancient Greek. A Study on pléō and its Compounds.* University of Pavia: MA Thesis.

Farina, A. (2024). *Guidelines for a linguistic annotation of preverbed verbs of motion*. King's College London. Figshare. https://doi.org/10.18742/25055573

Farina, A., Marongiu, P., Bru, M., & Borkowski, D. (2025). When data meets the past: data collection, sharing, and reuse in Ancient World studies. *Open Information Science*, 9, 1-20. https://doi.org/10.1515/opis-2025-0014

Feinmann, D. (2020). Language and Thought in the Motion Domain: Methodological Considerations and New Empirical Evidence. *Journal of Psycholinguistic Research*, 49(1), 1–29. https://doi.org/10.1007/s10936-019-09668-5

Gorman, V. (2020). Dependency Treebanks of Ancient Greek Prose. *Journal of Open Humanities Data*, 6(1), 1–3.

Harris, A. (2003). Preverbs and their origin in Georgian and Udi. G. Booij, & J. van Marle (Eds.), *Yearbook of Morphology*, 61–78.

Holleran, C. (2021). *Mapping Migration in Roman Iberia*. [https://mappingromanmigration.exeter.ac.uk/index.html]

Jones, P., & Sidwell, K. (1997). *The World of Rome: An Introduction to Roman Culture*. Cambridge: Cambridge University Press.

Kahane, A., & Mueller, M. (2001). *The Chicago Homer*. Web publication/site: University of Chicago Press.

Kipper Schuler, K., Korhonen, A., & Brown, S. (2009). VerbNet overview, extensions, mappings and applications. *Proceedings of NAACL HLT 2009: Tutorials, Boulder, Colorado, June 2009. Association for Computational Linguistics*, 13-14.

Klie, J.-C., Bugert, M., Boullosa, B., de Castilho, R. E., & Gurevych, I. (2018). The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation. *Proceedings of System Demonstrations of the 27th International Conference on Computational Linguistics (COLING 2018), Santa Fe, New Mexico, USA*, 5–9.

Lander, Y., Maisak, T. A., & Rakhilina, E. V. (2012). Verbs of aquamotion: Semantic domains and lexical systems. In M. Vulchanova & E. van der Zee (Eds.), *Motion Encoding in Language and Space*, 67–83.

McGillivray, B. (2009). *A computational approach to Latin verbs: New resources and methods*. University of Pisa: PhD Thesis.

McGillivray, B. (2012). A Valency Lexicon for Latin extracted from the Latin Dependency Treebank, Oxford Text Archive, http://hdl.handle.net/20.500.12024/2546.

McGillivray, B. (2013). Latin preverbs and verb argument structure: New insights from new methods. In E. van Gelderen, J. Barðdal, & M. Cennamo (Eds.), *Argument Structure in Flux. The Naples-Capri Papers*, 119–134.

McGillivray, B. (2014). *Methods in Latin computational linguistics*. Leiden: Brill.

McGillivray, B. (2021). Ancient Greek valency lexicon (AGVaLex). Figshare. Dataset. https://doi.org/10.6084/m9.figshare.14316251.v3

McGillivray, B., Kondakova, D., Burman, A., Dell'Oro, F., Bermúdez Sabel, H., Marongiu, P. & Márquez Cruz, M. (2022). A new corpus annotation framework for Latin diachronic lexical semantics. Journal of Latin Linguistics 21(1):47–105. 10.1515/joll-2022-2007

Meini, L., & McGillivray, B. (2010). Between semantics and syntax: Spatial verbs and prepositions in Latin. *Proceedings of the Space in Language Conference, Pisa (Italy), 8-10 October 2009*.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. J. (1990). Introduction to WordNet: An on-line lexical database. *International Journal of Lexicography*, 3(4), 235–244.

Parker, C.S., Parsons, S., Bandy, J., Chapman, C., Coppens, F., Seales, W.B. (2019). From Invisibility to Readability: Recovering the Ink of Herculaneum. *Public Library of Science*, PLOS ONE, 14(5). https://doi.org/10.1371/journal.pone.0215775

Passarotti, M. (2007a). LEMLAT. Uno strumento per la lemmatizzazione morfologica automatica del latino. F. Citti & T. el Vecchio (Eds.), *From Manuscript to Digital Text. Problems of Interpretation and Markup. Proceedings of the Colloquium (Bologna, June 12th 2003). Roma*, 107–128.

Passarotti, M. (2007b). Verso il Lessico Tomistico Biculturale. La treebank dell'Index Thomisticus. D. Fermia, & Raffaella Petrilli (Eds.), *Il filo del discorso. Intrecci testuali, articolazioni linguistiche, composizioni logiche. Atti del XIII Congresso Nazionale della Società di Filosofia del Linguaggio. Viterbo*, 187–205.

Passarotti M., González Saavedra B., Onambele C. (2016). Latin Vallex. A Treebank-based Semantic Valency Lexicon for Latin. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Portorož, Slovenia*, May 23-28, 2016, pp. 2599-2606.

Passarotti, M., & Mambrini, Francesco. (2021). Linking Latin: Interoperable Lexical Resources in the LiLa Project. E. Biagetti, C. Zanchi, & S. Luraghi (Eds.), *Building new resources for historical linguistics,* 103–124.

Petersman, H. (1999). The Language of Early Roman Satire: Its Function and Characteristics. *Proceedings of the British Academy*, 93, 289–310.

Revuelta Puigdollers, A. R. (2014). Some verbs prefixed by περι- in Ancient Greek. A. Bartolotta (Ed.), *The Greek verb. Morphology, Syntax and Semantics. Proceedings of the 8th International Meeting on Greek Linguistics*, 291–309.

Revuelta Puigdollers, A. R. (2016). A cognitive-functional study of the prefix circum-: Some non-prototypical cases. P. Poccetti (Ed.), *Latinitatis rationes: Descriptive and Historical Accounts for the Latin Language*, 127–146.

Simon, R., Barker, E., & Isaksen, L. (2012). Exploring Pelagios: A visual browser for geo-tagged datasets. *International Workshop on Supporting Users' Exploration of Digital Libraries, Paphos (Cyprus)*, 1–6.

Sommerschield, T. (2020). Restoring ancient texts using Machine Learning: A case-study on Greek and Latin epigraphy. *Papers of the British School at Rome*, 88, 387-388.

Talmy, L. (1985). Lexicalization patterns: Semantic structure in lexical forms. T. Shopen (Ed.), *Language Typology and Syntactic Description, III: Grammatical Categories and the Lexicon*, 57–149.

Talmy, L. (2000). *Toward a Cognitive Semantics. Vol. 1: Concept Structuring Systems*. Cambridge (MA): Mit Press.

Untersteiner, M. (1984). *Le origini della tragedia e del tragico: Dalla preistoria a Eschilo*. Milano: Cisalpino.
Vatri, A., & McGillivray, B. (2018). The Diorisis Ancient Greek Corpus. *Research Data Journal for the Humanities and Social Sciences*, 1–11.
Vázques Hoys, A.M. (2006). La importancia del agua en las civilizaciones antiguas: Grecia. *Tecnología del agua*, 276, 92-103.

Potential bias in AI: cultural representation and the marginalization of African art

Francesca Bignotti[1]
[1] University of Verona, Italy - francesca.bignotti@studenti.univr.it

ABSTRACT (ENGLISH)
The following paper explores the relationship between Artificial Intelligence (AI) and cultural representation, emphasizing how bias in Large Language Models (LLM) perpetuates the marginalization of African art. Building on the concept of "coloniality" and the presence of colonial structures in present-day knowledge systems, this study explores how contemporary AI systems trained on Eurocentric data sets extend and reinforce colonial hierarchies and rewrite cultural histories. The paper seeks to discuss the systemic lack of diversity and the reductionist approach towards representation of African art within digital platforms with the help of postcolonial theory, Digital Humanities and AI ethics. The approach used is a critical theoretical analysis and an interdisciplinary literature review, with a preliminary section that explores the possible application of the Bias Intelligence Quotient (BiQ) framework to the evaluation of cultural biases in the representation of African art. Thus, the paper seeks to provide mitigation techniques to reduce epistemological imbalance and encourage proper cultural inclusion of AI in promoting diversity and justice in the digital world.
Keywords: Artificial Intelligence (AI); Bias; Discrimination; Ethics of the Digital Humanities; Large Language Models (LLMs)

ABSTRACT (ITALIANO)
Possibili bias nell'IA: rappresentazione culturale e marginalizzazione dell'arte africana
Il presente articolo esplora la relazione tra Intelligenza Artificiale (IA) e rappresentazione culturale, sottolineando come i bias nei Modelli Linguistici di Grande Scala (LLM) perpetuino la marginalizzazione dell'arte africana. Basandosi sul concetto di "colonialità" e sulla presenza di strutture coloniali nei sistemi di conoscenza contemporanei, lo studio analizza come i sistemi di IA odierni, addestrati su dataset eurocentrici, estendano e rafforzino le gerarchie coloniali riscrivendo le storie culturali. L'articolo intende affrontare la sistemica mancanza di diversità e l'approccio riduzionista verso la rappresentazione dell'arte africana sulle piattaforme digitali, avvalendosi della teoria postcoloniale, delle Digital Humanities e dell'etica dell'IA. L'approccio adottato è un'analisi teorica critica e una revisione interdisciplinare della letteratura, con una sezione preliminare che esplora la possibile applicazione del framework Bias Intelligence Quotient (BiQ) alla valutazione dei bias culturali nella rappresentazione dell'arte africana. Pertanto, l'articolo mira a proporre tecniche di mitigazione per ridurre lo squilibrio epistemologico e a promuovere un'adeguata inclusione culturale dell'IA per favorire diversità e giustizia nel mondo digitale.
Parole chiave: Intelligenza Artificiale (IA); Bias; Discriminazione; Etica delle Digital Humanities; Modelli Linguistici di Grandi Dimensioni (LLM).

1. INTRODUCTION
In 2001, two art curators expressed in the catalogue for one of their exhibitions that those who are not exhibited do not exist (Hassan & Oguibe 2001), highlighting visibility's crucial role in gaining acknowledgement within society. Two decades later, this concept is still relevant, although in a different context; it is not only about galleries or public places but also about artificial intelligence (hereinafter AI). In the current world where AI has permeated almost every industry, it is not enough to simply be visible to be recognized as outstanding; it also means being embedded in the algorithms that drive our present society and future. Integrating AI into specific fields has created several opportunities (Nygren and Drimmer, 2023) but, at the same time, it has raised concerns about the existing biases and historical prejudices (Stahl & Eke, 2024). Colonialism left a long-term effect known as coloniality that still affects knowledge systems and cultural values to this day (Varshney et al., 2024).[1] The term coloniality can be

[1] It is crucial to make a distinction between data colonialism and data coloniality. Data colonialism can be defined as the use of data for colonial or imperialistic purposes whereby data from minorities is collected and used in a way that benefits colonial or imperialistic projects. Data coloniality is the structures and practices that ensure the subordination of these groups even when legal measures have been taken to address such issues, thus ensuring the continuity of control and dominance over data, including the ability to make decisions about its use.

defined as the structures of power that are a remnant of colonialism; it shows how the system of racial classification was employed to reinforce the colonial order with the emphasis on the subjugation of labour, which defines contemporary social, political and economic relations in the context of neocolonialism and systemic poverty (Arias, 2013). One example that poses challenges to a decolonial perspective is that AI favours already gathered and existing data, such as European sources, which are the most available (Nygren & Drimmer, 2023). This reliance on hierarchies poses a risk of perpetuating a perspective and goes against recent efforts to challenge Eurocentric views. By relying on distorted or incomplete representations, such models often reproduce narratives that may further alienate already underrepresented groups (Djanegara et al., 2024). Large language models (hereafter LLM) are now increasingly employed in academic research, creative writing, and professional sectors (Stöhr et al., 2024); however, they can also embed bias based on their data set and could unintentionally contribute to maintaining inequalities.[2] This theoretical analysis aims to understand how biases in LLMs can marginalize African art, perpetuating colonial epistemologies (Scheuerman et al., 2021) and Eurocentric theories. Furthermore, it incorporates Digital Humanities principles to offer practical measures to help reduce and mitigate these biases and create a more inclusive and equitable future.

2. METHODOLOGY

The paper's methodology is anchored on a theoretical approach whereby the research critically evaluates the current inter-disciplinary literature from postcolonial studies, critical theory, and recent contributions to AI and LLMs' ethics and cultural implications. Through these understandings, the study develops a conceptual framework to identify the epistemological and cultural processes of representing African artistic practices in AI data sets. In addition, a preliminary experimental section was developed to illustrate how cultural biases in LLMs can be evaluated.

3. BIAS AND THE MARGINALIZATION OF AFRICAN ART IN AI SYSTEMS

An essential strength of AI is that it can process large amounts of data and identify meaningful patterns; however, this capability highlights a potentially significant limitation. The data that fuels these models are not free from bias and often paint a skewed picture of non-Western societies (Tiribelli et al., 2024). AI and LLMs biases can be identified in different categories, affecting the system's precision and, essentially, the fairness of its outputs. These biases can be attributed to issues related to the training data, algorithm design or interactions between users and systems (Ferrara, 2023b). A clear example is the "Generative bias", which appears in systems that generate content like images, text or other output forms (Ferrara, 2023a). On the other hand, "Measurement bias" occurs when training data is collected or sampled unevenly, leading to the over or under-representation of certain social groups (Ferrara, 2023a). When AI models and LLMs are provided with a dataset which is prejudiced or incomplete, then the output generated will also be prejudiced and may amplify some cultures while reducing others (Chaka, 2022); in fact, discrimination based on culture, gender, ethnicity, occupation, or religion in AI is widely documented (Weidinger et al., 2021). When it comes to art history, for instance, LLMs have been trained on a mostly Western canon and, therefore, tend to favour European or American traditions (Mhlambi & Tiribelli, 2023) at the expense of non-Western art, predominantly African cultural heritage. This oppression is not just a question of data bias but a question of colonialism that, to this day, influences the process of categorization and valuation of culture across the globe (Baradaran, 2024). The exclusion of African art in AI models reflects the same colonialism (Tiribelli et al., 2024) that impacts many aspects of life, including cultural assets. Western art has long been privileged as the standard of artistic excellence, while African traditions have often been relegated to the margins or excluded from mainstream discourse altogether. These practices have contributed to the perception of African artistic practices not as part of a global artistic heritage but as a separate and inferior category confined to a distant and "primitive" past (Kasfir, 1992). LLMs trained within this colonial framework can replicate these exclusionary patterns (Kizhner et al., 2021), reinforcing that African art is not integral to global art history.
Indeed, despite efforts to mitigate biases in AI, responses generated by LLMs may reflect a generic tone when addressing African cultural heritage; in attempts to compensate for the historical

[2] Large Language Models are constructed through the formation of statistical-probabilistic connections between tokens, which represent the smallest textual units into which words are broken down during the model's training phase. For this reason, if an incomplete or decontextualized prompt is provided, the system will adapt it in the output using the terms that most frequently co-occur with those used by the user.

underrepresentation of African artistic traditions, these models often rely on generic statements, sometimes overly simplistic (Gururangan et al., 2018).

One reason for this knowledge gap is linguistic hegemony (Zeng & Yang, 2024), defined as one language's dominance over the other in social, cultural, and technological aspects. In technology and AI, languages such as English dominate data collection, algorithm development and content creation (Ferrara, 2023b). This results in a disparity whereby speakers of languages that are not dominant are underrepresented, their cultures are depicted less, and their opinions are shaped or not heard at all by the dominant linguistic group. This linguistic dominance affects the perception of people of different languages as well as the perception of their cultures on a global level. AI systems like LLMs are trained on data reflecting this linguistic hegemony, with datasets sourced mainly from the internet, where about two-thirds of the content is in English (El Louadi, 2024). As a result, the data from the developed countries is abundant, while the data from the developing countries, mainly from Africa, is scarce; for instance, out of the 54 African countries, datasets from Egypt were only used for machine learning models between 2015 and 2020 (El Louadi, 2024). This means that the African continent is portrayed through the representations of others; Africans are reduced to consumers of information about their culture instead of being producers of knowledge. Africa has a long way to go in adopting emerging technologies such as AI and meeting the increasing need for digital skills and digitization (Whaanga et al., 2015). In sub-Saharan Africa, only 6% of the population has broadband access, limiting both participation in the digital economy and representation in AI models. This lack of access, combined with the underrepresentation of Africa in digital datasets, naturally contributes to bias in AI systems (El Louadi, 2024).

4. BIAS MITIGATION FOR EQUITABLE AFRICAN ART REPRESENTATION IN AI

In Africa, digital marginalization, data marginalization, and algorithmic exclusions are major concerns that still define the current state of oppression and marginalization in the use of digital technologies. The digital technologies, big data, and algorithms used by organizations, corporations, institutions, and governments within different data jurisdictions contribute to this. Among the most significant problems are data ableism, data disablism, and data colonialism, which are evident when data generated, organized, captured, and manipulated by individuals and communities that are not in vogue is captured by big players (Chaka, 2022). This use of data simply maintains the existing structural inequalities where the value obtained from the data of powerless communities is captured for economic gain or power. Thus, data colonialism becomes one more mechanism to sustain such inequalities and is significantly damaging in the Global South, where the digital divide compounds the problems of accessibility and inclusion. As outlined, the underrepresentation of African cultural heritage in AI datasets presents a quantitative deficit, which is the lack of proportional data and a qualitative deficiency, where existing data fail to capture these artistic traditions' cultural richness fully. This dual bias results from systemic epistemological asymmetries that privilege the Eurocentric discourse, thus producing AI-generated distorted or oversimplified and generic outputs. This is because AI systems rely on probabilistic approximations, and without all the information, it tends to make content disconnected from specific knowledge. In these cases, models generate what are commonly referred to as hallucinated outputs (Banerjee et al., 2024), artificially fabricated and vague information presented as truth, often with a superficial and generic tone.[3] While these hallucinated results are difficult to eliminate entirely (Banerjee et al., 2024), some solutions can be implemented to mitigate such biased responses.

Addressing bias requires interventions targeting both data quantity and quality, starting with FAIR principles (Findable, Accessible, Interoperable, and Reusable) that provide a framework for organizing, structuring and curating datasets (Raza et al., 2024). These principles include practices such as detailed documentation of data provenance, clear descriptions of the processes used for collection and annotation and the mechanisms for updating the data regularly to capture changes in culture and technology (Huerta et al., 2023). The problem of bias in AI systems is not only a technical one but a social one, which requires a change in attitudes towards diversity (Schwartz et al., 2020) and international (Foka & Griffin, 2024) as well as interdisciplinary collaboration (Morris, 2020). In this manner, a collaborative and community-sensitive data validation process through collaborative annotations, contextualization, and enhancements ensures that cultural knowledge is dynamic and accurate. By integrating specific expertise into every phase of the AI workflow, systems can generate more nuanced outputs and avoid perpetuating hollow discourses that present artificial parity.

[3] Cambridge Dictionary, s.v. "hallucinate," (Accessed 27/12/2024), https://dictionary.cambridge.org/dictionary/english/hallucinate.

Including non-Eurocentric archives and African museum collections in AI training datasets (Costanza-Chock, 2020) is an essential measure for promoting diversity; therefore, traditional African artefacts, texts, and other digital resources from African institutions should be incorporated into the datasets. Digitizing artefacts and recording data histories preserved in African institutions ensures that datasets reflect diverse cultural perspectives, challenging the structural dominance of Western epistemologies. While digital content from Africa exists (Duarte, 2021), the continent's contribution to global data creation remains minimal. Africa's role in the global digital landscape is relatively marginal, with limited involvement in web domain registration and AI research (Ojanperä et al., 2017). Despite a broad internet user base, Africa lags in data infrastructure, with significantly fewer data centres than regions like the United States and Europe. Much of the data generated in Africa is often stored outside the continent, restricting its ability to leverage its digital potential fully (El Louadi, 2024). However, merely increasing data volume is insufficient; creating culturally specific data in collaboration with art historians (Nygren & Drimmer, 2023), cultural heritage experts and Digital Humanities scholars is equally crucial. This process ensures that AI systems generate outputs preserving African art's depth, diversity, and interpretative integrity rather than perpetuating hollow approximations (Foka & Griffin, 2024).

In addition to developing diverse data sets, it is also important to establish measures for real-time bias detection and mitigation in AI systems (Raji et al., 2020). Bias detection tools and fairness auditing protocols allow continuous monitoring of datasets and model outputs. These tools provide quantitative and qualitative scores to identify underrepresentation, stereotypes or hallucinated content, triggering immediate corrective actions. For example, outputs can be compared and cross-checked against annotations reviewed and maintained by field experts (Foka & Griffin, 2024) to identify recurring biases or Eurocentric tropes. To ensure transparency (Mitchell et al., 2019), there must be records of all the decisions that have gone into collecting data, annotating it and training the model.

A general critical decolonial approach has already been suggested to deal with datafication, algorithms, and digital citizenship from a combination of critical approaches integrated with the decolonial framework. This approach, named the Critical Southern Decolonial (CSD) framework, incorporates various theoretical lenses to paint an interconnected picture (Chaka, 2022). The framework supports the creation of new paradigms that can preserve the variety and plurality of the human and cultural experience and challenges the supremacy of the Western and American data and algorithm systems. Furthermore, the CSD framework also looks at the colonial dimensions of digital citizenship, where the latter is defined and practised in accordance with Eurocentric archetypes and thus does not acknowledge forms of citizenship, including Indigenous, nomadic, migrant or transnational citizenship. Instead, it proposes a new approach to digital citizenship that acknowledges diverse communities and their heterogeneity regardless of geographical location.

The increasing consciousness of epistemological inequalities and cultural erasure in datasets has led to new initiatives to fight for equality and properly represent non-Western societies. Projects such as Artificial Intelligence for Development in Africa (Siminyu et al., 2021) exemplify how digital technology can be a tool for cultural empowerment and epistemic justice. The AI4D Africa project, aimed at facilitating technological growth in the African continent, involves the creation of datasets that encompass the local languages, cultures, and contexts often excluded from global AI platforms. This project underscores how digital practices can evolve into tools for cultural equity, shifting the focus from data quantity to contextual quality and agency. Another significant project uniting the artistic context, AI, and its related biases and gaps was the An[0]ther {AI} in Art Summit (New Museum, 2019)[4]; which, for the first time, brought together 80 artists, scholars, curators, and technologists to address challenges and opportunities at the intersection of art and AI. Founded by Amir Baradaran, the summit was co-hosted by the Knight Foundation, Columbia University School of Engineering, New Museum, and NEW INC, exploring how AI is changing the art world and promoting equity (Baradaran, 2024). Thus, such projects prove the importance of a cross-disciplinary approach and ethical cooperation (Foka & Griffin, 2024) in utilizing digital tools to address bias. In a context where AI models often reinforce distorted representations, similar projects offer a way to imagine a future in which diversity is included and prized.

[4] Goldsmiths, University of London. (2019, May 10). An0ther AI in art: Decolonizing artificial intelligence (AI) and the future of art-making. ICCE Communication. Retrieved from https://sites.gold.ac.uk/iccecommunication/2019/05/10/an0ther-ai-in-art-decolonizing-artificial-intelligence-ai-and-the-future-of-art-making/ (Accessed 27/12/2024).

5. THEORETICAL ADAPTATION OF BIQ

Considering potential distortions and biases in the representation of African art within LLMs, this work proposes an adaptation of the Bias Intelligence Quotient (BiQ) (Narayan et al., 2024) within the context of the Comprehensive Bias Neutralization Framework (CBNF), to more precisely analyze the specific cultural biases inherent to this field of inquiry. The BiQ was originally developed as a metric capable of analyzing multiple aspects of bias, including not only the explicit content of responses but also the diversity of training data, linguistic tone, contextual sensitivity, and the model's ability to mitigate and update its outputs. This structure makes the framework particularly suitable for contexts such as African art and non-Western heritage, where epistemic flattening and Eurocentric interpretations risk being reproduced even by formally neutral models.

However, applying the BiQ in this context requires a targeted adjustment of its components. Specifically, it is proposed that the bias score be assigned a specific weight of 0.3, reflecting the centrality of explicit forms of generalization or simplification in the representation of African art. The penalty for low dataset diversity (P(d)) is understood as a fixed indicator linked to the quality and origin of the training data: in the original study, this is proposed as 0.3 for Latimer AI, which incorporates historical and cultural sources focused also on minority cultures, and 0.2 for ChatGPT, whose dataset is broader but less targeted in training.[5] Sentiment bias (s), calculated through automated semantic analysis using TextBlob, aims to detect polarization and is modulated through a coefficient λ equal to 0.5, due to its significant weight in the implicit perception of content. Context sensitivity is considered with a baseline value of 0.5 but is increased to 0.6 for prompts that directly involve the themes of colonialism, decolonization, or the restitution of African heritage.

In these cases, the model's ability to correctly interpret the historical and political framework is essential to avoid distortions on sensitive topics. In line with the original framework, it is also deemed important to evaluate the effectiveness of mitigation strategies (M) and adaptability (A). The former assesses the extent to which a model improves its responses when presented with more precise prompts or contextual terminology; the latter reflects flexibility in absorbing new trends, debates, and critical references, such as the growing role of the diaspora or practices of restitution concerning non-Western artifacts and art.

Following these assumptions and selected weights, a preliminary empirical exploration was later conducted to test the applicability of the framework to a limited number of sample queries related not only to African art tout court but also to its relationship with Western art and the themes of colonialism. The observed results, though not generalizable due to their exploratory nature, aligned with those reported in the original study, reinforcing the validity of the methodological proposal and suggesting that BiQ, if properly calibrated, can offer a solid basis for the critical analysis of bias in LLMs.

Nonetheless, it is important to emphasize the need for systematic and extended replications of this study, as discussed in Appendix 2 of the original report (Narayan et al., 2024), which outlines the potential to extend the BiQ methodology to a wide variety of prompts, allowing for broader and more robust analysis. Ongoing comparative analyses between different LLMs, even when conducted on thematically focused topics, could clearly demonstrate that the composition and origin of sources within training datasets play a decisive role in shaping the level of implicit and explicit bias. In particular, the inclusion of non-Western sources would emerge as a key strategy for designing equitable models, implying a reconfiguration of the very concept of a dataset, where cultural variety is no longer a supplementary element but a structural condition essential to ensuring ethical and epistemological quality.

6. CONCLUSIONS

This paper has highlighted how bias in AI models is not only the result of technical choices but is also a reflection of historical and cultural inequalities in the digital information universe. The predominance of a Eurocentric perspective in data sets limits our understanding of artistic diversity and reinforces narratives that may be perceived as discriminatory. To this end, these biases need to be addressed with a deliberate effort to ensure data diversity, the involvement of specialists from different fields, bias audits, and the adoption of decolonial approaches. Furthermore, a preliminary discussion considers how the BiQ framework could be adapted to assess cultural biases in the representation of African art, providing a conceptual foundation for future empirical research. The paper also suggests that African cultural heritage

[5] Latimer AI was explicitly designed to represent a wide range of cultural and social perspectives, including those that have been historically marginalized. Its training includes not only essays and publications, but also oral traditions and localized archives. Latimer AI is named after Howard Lewis Latimer, a scientist and inventor of African descent who lived at the turn of the 19th and 20th centuries.

and other marginalized artistic histories should be included in the digital ecosystem not as a secondary addition but as an intrinsic part of the art history discipline. In pursuit of a positive change in the digital environment, it is necessary to work on mitigating biases in datasets and algorithms and appreciating the diversity of the world's artistic traditions. Thus, as was noted at the beginning of this paper, visibility is crucial for existence, ensuring representation and cultural justice is an ethical and social imperative in a world increasingly mediated by AI.

BIBLIOGRAPHY

Arias, S. (2013). Coloniality and its preoccupations. Latin American Research Review, 48(3), 214–220. http://www.jstor.org/stable/43670103

Banerjee, S., Agarwal, A., & Singla, S. (2024). LLMs will always hallucinate, and we need to live with this. arXiv. https://doi.org/10.48550/arXiv.2409.05746

Baradaran, A. (2024). Towards a decolonial I in AI: Mapping the pervasive effects of artificial intelligence on the art ecosystem. AI & Society, 39(1), 7–19. https://doi.org/10.1007/s00146-023-01771-5

Chaka, C. (2022). Digital marginalization, data marginalization, and algorithmic exclusions: A critical Southern decolonial approach to datafication, algorithms, and digital citizenship from the Souths. Journal of e-Learning and Knowledge Society, 18(3), 83-95. https://doi.org/10.20368/1971-8829/1135678

Costanza-Chock, S. (2020). Design justice: Community-led practices to build the worlds we need. The MIT Press. https://doi.org/10.7551/mitpress/12255.001.0001

Djanegara, N. D. T., Zhang, D., Uz Zaman, H. B., Meinhardt, C., Watkins, G., Nwankwo, E., Wald, R., Kosoglu, R., Koyejo, S., & Elam, M. (2024). Exploring the impact of AI on Black Americans: Considerations for the Congressional Black Caucus's policy initiatives. Stanford Institute for Human-Centered AI. https://hai.stanford.edu/white-paper-exploring-impact-ai-black-americans-considerations-congressional-black-caucuss-policy

Duarte, C. (2021). Africa is going digital, Point of View. International Monetary Fund, Finance & Development, March 18–20. https://www.imf.org/en/Publications/fandd/issues/2021/03/africas-digital-future-after-COVID19-duarte

El Louadi, M. (2024). On the preservation of Africa's cultural heritage in the age of artificial intelligence. arXiv. https://doi.org/10.48550/arXiv.2403.06865

Ferrara, E. (2023a). Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. SSRN. http://dx.doi.org/10.2139/ssrn.4615421

Ferrara, E. (2023b). Should ChatGPT be biased? Challenges and risks of bias in large language models. arXiv. https://doi.org/10.48550/arXiv.2304.03738

Foka, A., & Griffin, G. (2024). AI, cultural heritage, and bias: Some key queries that arise from the use of GenAI. Heritage, 7, 6125–6136. https://doi.org/10.3390/heritage7110287

Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S. R., & Smith, N. A. (2018). Annotation artifacts in natural language inference data. arXiv. https://doi.org/10.48550/arXiv.1803.02324

Hassan, S. M., & Oguibe, O. (2001). Authentic / Ex-centric: Conceptualism in contemporary African art. Forum for African Arts, 4. https://www.proquest.com/scholarly-journals/authentic-ex-centric-at-venice-biennale-african/docview/220959008/se-2?accountid=14836

Huerta, E. A., Blaiszik, B., Brinson, L. C., Bouchard, K. E., Diaz, D., Doglioni, C., Duarte, J. M., Emani, M., Foster, I., Fox, G., Harris, P., Heinrich, L., Jha, S., Katz, D. S., Kindratenko, V., Kirkpatrick, C. R., Lassila-Perini, K., Madduri, R. K., Neubauer, M. S., Psomopoulos, F. E., Roy, A., Rübel, O., Zhao, Z., & Zhu, R. (2023). FAIR for AI: An interdisciplinary and international community building perspective. Scientific Data, 10(487). https://doi.org/10.1038/s41597-023-02298-6

Kasfir, S. L. (1992). African art and authenticity: A text with a shadow. African Arts, 25(2), 41–97. https://doi.org/10.2307/3337059

Kizhner, I., Terras, M., Rumyantsev, M., Khokhlova, V., Demeshkova, E., Rudov, I., & Afanasieva, J. (2021). Digital cultural colonialism: Measuring bias in aggregated digitized digitized content held in Google Arts and Culture. Digital Scholarship in the Humanities, 36(3), 607–640. Article fqaa055. https://doi.org/10.1093/llc/fqaa055

Mhlambi, S., & Tiribelli, S. (2023). Decolonizing AI ethics: Relational autonomy as a means to counter AI harms. Topoi, 42, 1–14. https://doi.org/10.1007/s11245-022-09874-2

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. Proceedings of the Conference on Fairness, Accountability, and Transparency, 220–229. https://doi.org/10.1145/3287560.328759

Morris, M. R. (2020). AI and accessibility. Communications of the ACM, 63(6), 35–37. https://doi.org/10.1145/3356727

Narayan, M., Pasmore, J., Sampaio, E., Raghavan, V., & Waters, G. (2024). Bias Neutralization Framework: Measuring Fairness in Large Language Models with Bias Intelligence Quotient (BiQ). arXiv. https://doi.org/10.48550/arXiv.2404.18276

Nygren, C., & Drimmer, S. (2023). Art history and AI: Ten axioms. International Journal for Digital Art History, 9, 5.02–5.13. https://doi.org/10.11588/dah.2023.9.90400

Ojanperä, S., Graham, M., Straumann, R., Sabbata, S. D., & Zook, M. (2017). Engagement in the knowledge economy: Regional patterns of content creation with a focus on Sub-Saharan Africa. Information Technologies & International Development, 13, 33–51. https://itidjournal.org/index.php/itid/article/view/1479.html

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 33–44. https://doi.org/10.48550/arXiv.2001.00973

Raza, S., Ghuge, S., Ding, C., Dolatabadi, E., & Pandya, D. (2024). FAIR enough: How can we develop and assess a FAIR-compliant dataset for large language models' training? arXiv. https://arxiv.org/abs/2401.11033

Scheuerman, M. K., Pape, M., & Hanna, A. (2021). Auto-essentialization: Gender in automated facial analysis as extended colonial project. Big Data & Society, 8. https://doi.org/10.1177/20539517211053712

Schwartz, R., Dodge, J., Smith, N. A., & Etzioni, O. (2020). Green AI. Communications of the ACM, 63(12), 54–63. https://doi.org/10.1145/3381831

Siminyu, K., Kalipe, G., Orlic, D., Abbott, J., Marivate, V., Freshia, S., Sibal, P., Neupane, B., Adelani, D. I., Taylor, A., Touré, A. L. I., Degila, K., Balogoun, M., Diop, T. I., David, D., Fourati, C., Haddad, H., & Naski, M. (2021). AI4D – African Language Program. arXiv. https://doi.org/10.48550/arXiv.2104.02516

Stahl, B. C., & Eke, D. (2024). The ethics of ChatGPT – Exploring the ethical issues of an emerging technology. International Journal of Information Management, 74. https://doi.org/10.1016/j.ijinfomgt.2023.102700

Stöhr, C., Ou, A. W., & Malmström, H. (2024). Perceptions and usage of AI chatbots among students in higher education across genders, academic levels and fields of study. Computers and Education: Artificial Intelligence, 7. https://doi.org/10.1016/j.caeai.2024.100259

Tiribelli, S., Pansoni, S., Frontoni, E., & Giovanola, B. (2024). Ethics of artificial intelligence for cultural heritage: Opportunities and challenges. IEEE Transactions on Technology and Society, 1(1), 298. https://doi.org/10.1109/TTS.2024.3432407

Varshney, K. R. (2024). Decolonial AI alignment: Openness, Viśeṣa-Dharma, and including excluded knowledges. arXiv. https://doi.org/10.48550/arXiv.2309.05030

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., Kenton, Z., Brown, S., Hawkins, W., Stepleton, T., Biles, C., Birhane, A., Haas, J., Rimell, L., Hendricks, L. A., Isaac, W., Legassick, S., Irving, G., & Gabriel, I. (2021). Ethical and social risks of harm from language models. CoRR, abs/2112.04359. https://doi.org/10.48550/arXiv.2112.04359

Whaanga, H., et al. (2015). He Matapihi Mā Mua, Mā Muri: The ethics, processes, and procedures associated with the digitization of indigenous knowledge—the Pei Jones collection. Cataloging & Classification Quarterly, 53(5-6), 520–547. https://doi.org/10.1080/01639374.2015.1009670

Zeng, J., & Yang, J. (2024). English language hegemony: Retrospect and prospect. Humanities and Social Sciences Communications, 11(317). https://doi.org/10.1057/s41599-024-02821-z

Exploring data-driven narratives in Digital Humanities web-based projects: features and impact

Tommaso Battisti[1], Marilena Daquino[2]
[1] University of Bologna, Italy - tommaso.battisti5@unibo.it
[2] University of Bologna, Italy - marilena.daquino2@unibo.it

ABSTRACT (ENGLISH)

Narrative approaches as an aid to information visualisation practices help to present complex information clearly and enable communication to a broader audience, serving as an interpretative framework with positive outcomes in the Cultural Heritage domain. Several studies discuss narrative design spaces for information visualisation, producing different taxonomies and classifications. Nonetheless, existing contributions mainly relate to data journalism or domain-agnostic projects, while surveys accounting for narrativity in the Digital Humanities are restricted to specific contexts and limited to the surface of these aspects. This work extends previous contributions by addressing narrative design choices in 186 web-based Digital Humanities projects that leverage information visualisation techniques. Projects are classified according to narrativity and narrative strategies, the humanities domain, and visualisation type. Findings show limited use of narrative approaches with a strong reliance on the use of storytelling tools and frequent homogeneous and similar solutions, while humanities domains show interesting and diverging patterns.

Keywords: Data-driven narratives; Information visualisation; Narrative design spaces; Survey

ABSTRACT (ITALIANO)

Esplorare le narrazioni guidate dai dati in progetti di Digital Humanities sul web: caratteristiche e impatto.
Gli approcci narrativi come supporto alle pratiche di visualizzazione dell'informazione aiutano a presentare informazioni complesse in modo chiaro e permettono la comunicazione a un pubblico più ampio, fungendo da quadro interpretativo con esiti positivi nel campo del Patrimonio Culturale. Diversi studi discutono aspetti di progettazione narrativa per la visualizzazione dell'informazione, producendo diverse tassonomie e classificazioni. Tuttavia, i contributi esistenti si riferiscono principalmente al campo del data journalism o a progetti senza relazione con domini specifici, mentre le indagini che tengono conto della narratività nelle Digital Humanities sono limitate a contesti precisi e affrontano solo superficialmente questi aspetti. Questo studio amplia i contributi precedenti analizzando le scelte di progettazione narrativa in 186 progetti di Digital Humanities che utilizzano tecniche di visualizzazione dell'informazione sul web. I progetti sono classificati in base alla narratività e alle strategie narrative, al dominio umanistico e al tipo di visualizzazione. I risultati mostrano un utilizzo limitato di approcci narrativi, con una forte dipendenza dall'uso di strumenti di storytelling e soluzioni frequentemente simili e omogenee, mentre i domini umanistici mostrano modelli interessanti e divergenti.

Parole chiave: Narrazioni guidate dai dati; Visualizzazione dell'informazione; Progettazione narrativa; Indagine

1. INTRODUCTION

Information visualisation offers significant opportunities to present and investigate complex information within the field of Digital Humanities (DH). Visual representations supported by computation tools and interaction are key for amplifying cognition (Card et al., 1999). In this context, the Web represents a powerful means to enhance data presentation and knowledge dissemination. Meanwhile, storytelling emerges as an additional aid to convey information more efficiently (Gershon & Page, 2001), with practical advantages for the Cultural Heritage domain, such as improving the presentation, exploration, and interpretation of complex information (Renda et al., 2023) and shaping stronger connections with the audiences, enhancing user experience and dissemination (Shan et al., 2022).
Several studies contributed to the definition of design spaces for data-driven narratives. Among these, (Segel & Heer, 2010) define different narrative genres, tactics, and structures, distinguishing author-driven and reader-driven approaches, with hybrid potentials in the middle. Similar contributions and further extensions followed (Hullman & Diakopoulos, 2011; Stolper et al., 2016; Roth, 2021; Zhao & Elmqvist, 2023). From a different perspective, (McKenna et al., 2017) defined and studied factors involved in the data-driven narratives' flow, capable of affecting the reading experience. Most of these studies

mainly surveyed diverse online sources, generally related to data journalism or more domain-agnostic works.
In the DH, various contributions propose solutions exploiting explorative and narrative approaches, which encompass generous interfaces (Whitelaw, 2015), workflows and platforms for the creation of Cultural Heritage visualisation-based stories (Liem et al., 2023), and authoring tools for linked open data-driven stories (Renda et al., 2023). Nonetheless, only a few surveys on DH projects exist and describe the application of information visualisation techniques in the field, while narrative aspects are even rarely considered. These are also limited to specific sub-contexts of applications such as Cultural Heritage collection interfaces (Windhager et al., 2019), text analysis support tasks (Jänicke et al., 2017), musicology works (Khulusi et al., 2020), and DH publications using visualisations as part of the research process (Panagiotidou et al., 2023). While having diverse scopes, interestingly, only (Windhager et al., 2019) address narrativity, limiting the analysis to an account of narrative approaches intended as "curated paths", without surveying deeper narrative design choices.
The purpose of this study is to broaden the scope of previous surveys encompassing the wider spectrum of the DH. In addition, it seeks to compensate for the lack of comprehensive reports on data-driven narrative design choices so far unexplored within the field.

2. METHODS

The present work analyses a corpus of 186 web-based DH projects leveraging information visualisation techniques with a particular focus on narrative strategies.
The corpus was developed by relying on diverse sources: 1) the existing classification by (Windhager et al., 2019) and related authors of interfaces, which were secondly surveyed to find additional material (33 projects); 2) the "projects" sections of the websites of the Italian (AIUCD)[1] and European (EADH)[2] DH associations, extending the inquiry to related research centres and institutions (104 projects); and 3) the "Best DH data visualisation" category of the Digital Humanities Awards[3] website (49 projects).
Selection criteria to include projects in the survey are the following: 1) must be web-based dissemination projects; 2) outcomes are potentially produced through visualisation tools or storytelling software but are not tools or software themselves; 3) free from access barriers like registration forms; 4) significantly reliant on visualisation techniques; and 5) accessible at the time of assessment.
The 186 projects composing the final corpus were classified through direct observation and interaction with the interfaces, evaluating only visual solutions and not accounting for any kind of textual description. This empirical approach minimises potential bias deriving from inconsistencies between documentation and available features of the systems during evaluation.
Our final classification framework reuses different taxonomies and schemas that were selected and adapted from literature. It was developed and refined using an iterative process inspired by (Segel & Heer, 2010) and consisting of testing different column configurations over samples of projects during classification until a suitable result was achieved, according to which we re-classified all instances. Even if two annotators evaluated the final dataset by reviewing a sample of instances, as recognised by (Segel & Heer, 2010), using a custom taxonomy over diverse design dimensions inevitably introduces some degree of subjectivity.
Features relevant to the current study include:
Narrativity. To identify projects incorporating information visualisation within narratives, whether linear author-driven or user-driven (Segel & Heer, 2010). Hybrid projects combine both narrative and non-narrative approaches.
Domain. Building on (Terras et al., 2016: 138) and (Schreibman et al., 2004), a categorical variable distinguishes the humanities field of the project, including: 1) history and archaeology 2) art and art history 3) language and literature, including linguistics, philology, narrative and literary studies 4) music and musicology 5) multimedia and performing arts 6) philosophy and religion, and 7) other, for other domains or collections with no unique focus.
Visualisation techniques. Adapted from (Windhager et al., 2019), a Boolean value indicates the presence of the following visualisation types: 1) plot 2) cluster or set 3) map, including a further identification for statistical symbol maps (i.e., data points are statistical charts) 4) network 5) hierarchical

[1] Associazione per l'Informatica Umanistica e la Cultura Digitale (AIUCD): https://www.aiucd.it/progetti/.

[2] European Association for Digital Humanities (EADH): https://eadh.org/projects.

[3] http://dhawards.org/.

diagram 6) treemap 7) word cloud 8) bars, any bar-based chart; 9) line chart 10) area chart 11) pie chart 12) 3D plot 13) proportional area 16) timeline, and 16) other, miscellaneous. Each column corresponds to its type and includes also the related stacked layouts and variations (e.g., a bar chart and a radial bar chart).

Narrative flow factors. A set of columns adapted from (McKenna et al., 2017) identifies relevant patterns in the design of narratives that involve: 1) story layout 2) visualisation role 3) story progression 4) navigation input 5) navigation progress, and 6) level of control over text, visualisations, and animated transitions. It is worth noticing that projects with multiple narratives can potentially report multiple design choices for the same flow factor. Moreover, while we use the term narrative since we refer to both authorial linear data stories and more user-directed narratives, the term "story" was used for factor names to match the original taxonomy.

The final dataset (Battisti, 2024) was analysed via a Jupyter Notebook (Battisti, 2025) to ensure reproducibility, while the following research questions guided our analysis:

- **RQ1**. How many projects use narrative approaches to information visualisation?
- **RQ2**. How can we characterise data-driven narratives in DH projects according to visualisation use and narrative flow factors?
- **RQ3**. How do humanities domains use narrativity?

3. RESULTS AND DISCUSSION

Overall, DH dissemination projects present only limited use of narrative approaches, the emergence of patterns linked to humanities domains, but also uncritical reuse of storytelling tools without additional customisation, impeding further valorisation of research material's unique characteristics.

RQ1. The great majority of the projects in our corpus are non-narrative (76,9%;143/186). Most of the remaining 23,1% of projects containing narratives rely on the complementary strengths of both narrative and non-narrative sections (12,9%; 24/186), while fully narrative projects are only 10,2% of the total (19/186). Narrativity is slightly higher than in (Windhager et al., 2019) survey on Cultural Heritage collection interfaces, but the count is still limited compared to the communication benefits that narrative approaches are claimed to bring to information visualisation (Gershon & Page, 2001; Renda et al., 2023), especially considering their role for dissemination.

RQ2. Considering visualisation use, the narrative context causes differences in visualisation frequency rankings. Most relevant, it widens the gap between maps—the most frequent solution, employed by 65,5% of the projects (122/186)—and networks, adopted by 37,6% of projects overall (70/186). In projects using at least a narrative element, the prevalence of maps rises to 87%, and the adoption of networks falls to 18,6%. The latter's underrepresentation may be an echo of limitations related to readability in dense configurations and connections' inadequacies in representing data measures compared to other visual variables (Koponen & Hildén, 2019). Nonetheless, narrative approaches and guided explorations could serve as an aid to enhance network exploration and analysis (Li et al., 2023).

Moreover, while most non-narrative projects (54,6%; 78/143) employ multiple visualisations, the majority of fully narrative projects rely on a single technique (57,9%; 11/19). Due to the presence of at least two sections (one narrative and one non-narrative) almost all hybrid projects use multiple visualisations (87,5%; 21/24). Concerning flow factors, narratives in DH dissemination projects show a particular preference. They are almost always linear and author-driven (95,4%; 41/43), with a preference for slideshow layouts (44,2%; 19/43) followed by document layouts (37,2%; 16/43), two projects with more narrative sections using different layouts, and six using hybrids or more complex formats. Despite slideshows being more frequent, document layouts account for more visualisations. Among the few exceptions, we found timelines (five occurrences in slideshows and three in documents) and maps (18 in slideshows and 10 in documents). The latter is also the most used visualisation in slideshow-document hybrid layouts.

Narrative progress is assigned to a button or click approaches 44,2% of the time (19/43) while 32,6% use the scroll (14/43) reflecting the division between slideshow and document layouts. 20,9% use multiple techniques (9/43), while only one uses a slider approach. Narrative progression is communicated by the visualisation itself in 25,6% of projects (11/43), followed by textual elements (11,6%; 5/43), multiple techniques (7%; 3/43) and only one by dots. Interestingly, most projects do not communicate user position within the narrative (53,5%; 23/43). The problem affects mainly document layouts that rely on the webpage scrollbar as an implicit signifier for user position. However, since a webpage may not reflect

the actual length of a narrative, the practice is in contrast with the idea that signifiers and feedback should ideally match user expectations to avoid misinterpretations during interaction (Norman, 2013).
Regarding the control of narrative elements, discrete and continuous text control is equally employed, with a project using them both. Differently, the majority of control type over visualisations is continuous (72,1%; 31/43), while discrete control affects most animations (65,1%; 28/43). 23,3% of projects (10/43) have no animation control.
In most narratives, visualisations and texts play an equal role in telling stories (53,5%; 23/43). 18,6% of narratives entrust visualisations with the main role (8/43), while only 11,6% of projects use them as support for text (5/43). The remaining projects use multiple solutions. In relation to the visualisation, maps and networks are mostly used with an equal role (71%; 22/31 and 62,5%; 5/8 respectively) and only 1/11 of timelines have a subordinate role. It is worth noticing that we excluded six projects containing both multiple roles and visualisations that impeded correct tracking between the two variables.
Relevant patterns also include 84.2% of slideshows with equal role visualisations (16/19), which correspond to 69.6% of equal role visualisations used in slideshows (16/23); 50% of document layout narratives with a subordinate role for visualisations (8/16), corresponding to the use of all the subordinated visualisations within document narratives (8/8); and all the other layouts without category providing visualisations with a driving role (2/2), which correspond to 40% of main-role visualisations used in other non-categorised layouts. Other patterns involving the layout include 94.7% of slideshows using button/click input (18/19) corresponding to the same amount also by shifting perspective; 87.5% of documents using scroll input (14/16) corresponding to the use of all scroll inputs within document layouts (14/14); and all document-slideshow hybrids using hybrid scroll-click navigation (4/4).
The most common control patterns are discrete text, continuous visualisation, and continuous animation control (37,2%; 16/43), which is always used by slideshows; and continuous text, continuous visualisation, and no animation control (20,9%; 9/43), which is always used by document layouts.
Association rules generated through the Apriori algorithm[4] also underline the connection between document layout, scroll input, continuous text control, and slideshows, as well as discrete text control, button input, and visualisation's equal role.
Overall, the preference for single visualisation techniques in fully narrative projects, the higher number of slideshows and their use of most maps and timelines, as well as more complex patterns involving narrative flow factors witness the frequent use of storytelling tools such as StoryMapJS[5] or TimelineJS[6]. Such tools typically rely on a single visualisation technique and are employed for most narratives using slideshows (66,7%; 12/18)[7]. On the one hand, these tools offer a technically convenient way to develop narratives for dissemination projects. However, they almost always lack further customisation, failing to adapt to the specificities of the research material and producing always similar and homogeneous results that inhibit the valorisation of the uniqueness of the working data.
RQ3. Overall, two domains accommodate most projects: "language and literature" (39,8%; 74/186) and "history and archaeology" (34,9%; 65/186). Given the greatly higher number of non-narrative projects, all domains present a preference for non-narrative approaches. However, by shifting perspective, while "language and literature" is the most popular domain among non-narrative projects (43,3%; 62/143) followed by "history and archaeology" (30,8%; 44/143), the latter category is the most present within narrative projects (48,8%; 21/43), with "language and literature" following at 27,9% (12/43). The gap enlarges considering only fully narrative projects, where "history and archaeology" count 57,9% of projects (11/19) and "language and literature" only 21,1% (4/19).
The narrative characterisation of the domains also comes with related and already discussed patterns. Therefore, while all domains use multiple visualisations—except for "music and musicology" which includes one single- and one multi-visualisation project— "history and archaeology" shows a more balanced approach (33/65 multi-visualisation to 32/65 single-visualisation). Noticeably it also accounts for most narrative projects, and a higher use of maps, including nearly half of all map occurrences in the dataset (46,7%; 57/122).

[4] The documentation of the Python algorithm used during the analysis can be accessed at https://rasbt.github.io/mlxtend/api_subpackages/mlxtend.frequent_patterns/.

[5] https://storymap.knightlab.com/.

[6] https://timeline.knightlab.com/.

[7] The number refers to narrative projects only. Nonetheless, these tools are often used also in non-narrative contexts to present items according to relevant metadata.

The divergent preferences corresponding to the two domains reflect distinct dissemination approaches, suggesting that disciplinary characteristics shape visualisation strategies and project design.
Limitations. Given the dimension of our corpus and the reliance on selected sources, we claim to have derived representative conclusions about the impact of narrativity in DH web-based dissemination projects. Nonetheless, the analysis of flow factors and visualisation use in narrative contexts suffer from the limited dimension of the narrative subset, which leads to interesting consideration but no further generalisable results. A similar line of reasoning involves the analysis of domains and their narrative characterisation since it was possible to derive conclusions only for the most frequent categories. Furthermore, the analysis focuses on projects rather than on single narrative sections leading to the impossibility of precisely understanding whether specific visualisations are used in narrative or non-narrative contexts within hybrid projects. Finally, also the presence of multiple narrative sections within single projects led to the impossibility of tracking visualisation roles when occurring both multiple visualisations and roles.

4. CONCLUSIONS

The present study analysed narrative practices in DH web-based dissemination projects leveraging information visualisation techniques, assessing the number of narrative approaches, how these vary according to the humanities domain, the use of visualisations in different narrative contexts, and narrative design patterns involving narrative flow factors. We extended both the context and the focus of previous humanities-related studies that almost never and only partially surveyed narrative aspects related to information visualisation techniques. The work also serves as a point of departure for further critical reflection on the role of storytelling tools and the relation between technical needs for simple and reusable solutions, and custom possibilities to enhance and valorise research material.

ACKNOWLEDGEMENTS

Project funded by the European Union – NextGenerationEU under the National Recovery and Resilience Plan (NRRP), Investment I.4.1 - Borse PNRR Patrimonio Culturale.

REFERENCES

Battisti, T. (2024). Classification of web-based Digital Humanities projects leveraging information visualisation techniques [Dataset]. Zenodo. https://doi.org/10.5281/ZENODO.14192758

Battisti, T. (2025). Data-driven insights into web-based Digital Humanities projects leveraging information visualization techniques: An analysis of the 'Classification of web-based Digital Humanities projects leveraging information visualisation techniques' dataset [Computer software]. Zenodo. https://doi.org/10.5281/ZENODO.14824189

Card, S. K., Mackinlay, J., & Shneiderman, B. (1999). Readings in information visualization: using vision to think. Morgan Kaufmann.

Gershon, N., & Page, W. (2001). What storytelling can do for information visualization. Communications of the ACM, 44(8), 31–37. https://doi.org/10.1145/381641.381653

Hullman, J., & Diakopoulos, N. (2011). Visualization Rhetoric: Framing Effects in Narrative Visualization. IEEE Transactions on Visualization and Computer Graphics, 17(12), 2231–2240. https://doi.org/10.1109/TVCG.2011.255

Jänicke, S., Franzini, G., Cheema, M. F., & Scheuermann, G. (2017). Visual Text Analysis in Digital Humanities. Computer Graphics Forum, 36(6), 226–250. https://doi.org/10.1111/cgf.12873

Khulusi, R., Kusnick, J., Meinecke, C., Gillmann, C., Focht, J., & Jänicke, S. (2020). A Survey on Visualizations for Musical Data. Computer Graphics Forum, 39(6), 82–110. https://doi.org/10.1111/cgf.13905

Koponen, J., & Hildén, J. (2019). Data visualization handbook (first edition). Otavan kirjapaino.

Li, W., Schöttler, S., Scott-Brown, J., Wang, Y., Chen, S., Qu, H., & Bach, B. (2023). NetworkNarratives: Data Tours for Visual Network Exploration and Analysis. Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, 1–15. https://doi.org/10.1145/3544548.3581452

Liem, J., Kusnick, J., Beck, S., Windhager, F., & Mayr, E. (2023). A Workflow Approach to Visualization-Based Storytelling with Cultural Heritage Data. 2023 IEEE 8th Workshop on Visualization for the Digital Humanities (VIS4DH), 13–17. https://doi.org/10.1109/VIS4DH60378.2023.00008

McKenna, S., Henry Riche, N., Lee, B., Boy, J., & Meyer, M. (2017). Visual Narrative Flow: Exploring Factors Shaping Data Visualization Story Reading Experiences. Computer Graphics Forum, 36(3), 377–387. https://doi.org/10.1111/cgf.13195

Norman, D. A. (2013). The design of everyday things (Revised and expanded ed). Basic Books.

Panagiotidou, G., Lamqaddam, H., Poblome, J., Brosens, K., Verbert, K., & Vande Moere, A. (2023). Communicating Uncertainty in Digital Humanities Visualization Research. IEEE Transactions on Visualization and Computer Graphics, 29(1), 635–645. https://doi.org/10.1109/TVCG.2022.3209436

Renda, G., Daquino, M., & Presutti, V. (2023). Melody: A Platform for Linked Open Data Visualisation and Curated Storytelling. Proceedings of the 34th ACM Conference on Hypertext and Social Media, 1–8. https://doi.org/10.1145/3603163.3609035

Roth, R. E. (2021). Cartographic Design as Visual Storytelling: Synthesis and Review of Map-Based Narratives, Genres, and Tropes. The Cartographic Journal, 58(1), 83–114. https://doi.org/10.1080/00087041.2019.1633103

Schreibman, S., Siemens, R. G., & Unsworth, J. (Eds.). (2004). A companion to digital humanities. Blackwell Pub.

Segel, E., & Heer, J. (2010). Narrative Visualization: Telling Stories with Data. IEEE Transactions on Visualization and Computer Graphics, 16(6), 1139–1148. https://doi.org/10.1109/TVCG.2010.179

Shan, X., Wang, D., & Li, J. (2022). Research on Data Storytelling Strategies for Cultural Heritage Transmission and Dissemination. In C. Stephanidis, M. Antona, S. Ntoa, & G. Salvendy (Eds.), HCI International 2022 – Late Breaking Posters (pp. 344–353). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-19679-9_43

Stolper, C. D., Lee, B., Riche, N. H., & Stasko, J. (2016). Emerging and recurring data-driven storytelling techniques: Analysis of a curated collection of recent stories. (2016).

Terras, M., Nyhan, J., & Vanhoutte, E. (2016). Defining Digital Humanities: A Reader. Routledge.

Whitelaw, M. (2015). Generous Interfaces for Digital Cultural Collections. Digital Humanities Quarterly, 9(1), 1–16.

Windhager, F., Federico, P., Schreder, G., Glinka, K., Dork, M., Miksch, S., & Mayr, E. (2019). Visualization of Cultural Heritage Collection Data: State of the Art and Future Challenges. IEEE Transactions on Visualization and Computer Graphics, 25(6), 2311–2330. https://doi.org/10.1109/TVCG.2018.2830759

Zhao, Z., & Elmqvist, N. (2023). The Stories We Tell About Data: Surveying Data-Driven Storytelling Using Visualization. IEEE Computer Graphics and Applications, 43(4), 97–110. https://doi.org/10.1109/MCG.2023.3269850

Il Glossario delle Infrastrutture di Ricerca (GIR)

Lucia Francalanci[1], Alessia Scognamiglio[2], Irene Falini[3], Pietro Restaneo[4], Giulia Pedonese[5]

[1] Istituto Opera del Vocabolario Italiano (CNR-OVI), Italia, lucia.francalanci@cnr.it
[2] Istituto per la Storia del Pensiero Filosofico e Scientifico Moderno (CNR-ISPF), Italia, alessia.scognamiglio@cnr.it
[3] Istituto Opera del Vocabolario Italiano (CNR-OVI), Italia, irene.falini@cnr.it
[4] Istituto per il Lessico Intellettuale Europeo e Storia delle Idee (CNR-ILIESI), Italia, pietro.restneo@cnr.it
[5] Istituto di Linguistica Computazionale "Antonio Zampolli" (CNR-ILC), Italia, giulia.pedonese@cnr.it

ABSTRACT (ITALIANO)

Il *Glossario delle Infrastrutture di Ricerca* (GIR) è un'iniziativa che si inserisce nell'ambito del progetto PNRR "H2IOSC - Humanities and Cultural Heritage Italian Open Science Cloud" del CNR, che mira a creare una federazione di Infrastrutture di Ricerca (IR) coinvolgendo i nodi italiani di quattro infrastrutture che fanno parte della *roadmap* dell'*European Strategy Forum on Research Infrastructure* nel settore dell'innovazione sociale e culturale: CLARIN-IT, DARIAH-IT, E-RIHS.it, OPERAS.it. Il GIR nasce dall'esigenza di avere una terminologia di riferimento all'interno del progetto H2IOSC. Nella redazione dei documenti legati al progetto è emersa la difficoltà legata alla mancanza di una terminologia univoca, condivisa e non ambigua, anche nell'ottica di una piena interoperabilità dei risultati della ricerca. Da qui la necessità di definire i termini che appartengono all'ambito della progettazione e implementazione delle IR, sciogliere i numerosi acronimi presenti, e documentare questo particolare linguaggio tecnico e specialistico. A partire da una rassegna critica sui più recenti studi teorici e applicati nel campo della terminologia della lingua italiana, del lessico delle IR, del PNRR e della progettazione europea, è stata individuata una metodologia per la costruzione del *Glossario* che prevede l'identificazione di un *corpus* di riferimento poi usato per la redazione delle voci. La struttura delle voci è stata progettata prendendo come riferimento alcuni vocabolari nati direttamente in rete e prevede, per ciascun lemma, l'indicazione della categoria grammaticale, la marca d'uso, le accezioni che il termine assume nel contesto delle IR, un campo "note" per eventuali integrazioni e le voci correlate, se presenti. Il GIR rappresenta un'iniziativa preziosa per promuovere la condivisione della conoscenza e la collaborazione tra le Infrastrutture di Ricerca in Italia.

Parole chiave: glossario; Infrastrutture di Ricerca; terminologia.

ABSTRACT (ENGLISH)

The Glossary of Research Infrastructures (GIR). The proposal for a *Glossary of Research Infrastructures* (GIR) was conceived within the framework of the PNRR project 'H2IOSC - Humanities and Cultural Heritage Italian Open Science Cloud'. H2IOSC aims at creating a federation of Research Infrastructures (RI) among the Italian nodes of 4 infrastructures that are part of the roadmap of the *European Strategy Forum on Research Infrastructure,* in the field of social and cultural innovation: CLARIN-IT, DARIAH-IT, E-RIHS.it, OPERAS.it. The GIR has been developed to answer the need for a common reference terminology within the H2IOSC project. While drafting the H2IOSC documents, a difficulty emerged related to the lack of unambiguous shared terminology, that would contribute to achieving the full interoperability of the research results. Hence the need to define the terms belonging to the field of IR project management and implementation; to explain the numerous acronyms present; and to document such a technical and specialized language. Our paper will start from a critical review of the most recent theoretical and applied studies in the field of terminology of the Italian language, such as the RI lexicon and the PNRR and European planning. At first, we defined the methodology for the construction of the Glossary, which envisages the identification of a reference *corpus.* The *corpus* was then used to draft the entries. The structure of the entries was designed by taking as a reference some vocabularies born directly on the web. The grid provides, for each entry, an indication of the grammatical category, the brand of use, the meaning or meanings that the term takes on in the context of the RI, 'notes' field for possible additions and any related entries. The GIR is a valuable initiative to promote knowledge sharing and collaboration between Research Infrastructures in Italy.

Keywords: glossary; Research Infrastructures; terminology.

1. IL PROGETTO E GLI OBIETTIVI

Il *Glossario delle Infrastrutture di Ricerca* (GIR) è un'iniziativa che si inserisce nell'ambito del progetto PNRR "H2IOSC - Humanities and Cultural Heritage Italian Open Science Cloud" (finanziato dall'Unione

europea NextGenerationEU - PNRR M4C2 - Codice progetto IR0000029 - CUP B63C22000730005)[1] del Consiglio Nazionale delle Ricerche, che mira a creare una federazione di Infrastrutture di Ricerca (IR), coinvolgendo i nodi italiani di quattro infrastrutture che fanno parte della *roadmap* dell'*European Strategy Forum on Research Infrastructure* (ESFRI) nel settore dell'innovazione sociale e culturale: CLARIN-IT, DARIAH-IT, E-RIHS.it., OPERAS.it.[2] Un'Infrastruttura di ricerca (IR) comprende un insieme di strutture, risorse e servizi correlati tra loro, utilizzati dalla comunità scientifica - e non solo - per condurre ricerche di alto livello nei rispettivi campi e promuovere l'innovazione nei territori.[3] Il Forum Strategico Europeo per le Infrastrutture di Ricerca (ESFRI), organo consultivo del Consiglio dell'Unione Europea per le IR, sottolinea l'importanza strategica che queste rappresentano per lo Spazio Europeo della Ricerca (ERA) e per l'innovazione dei territori dove sono incardinate.
Il gruppo di lavoro è composto da ricercatrici e ricercatori, tecnologhe e tecnologi afferenti al Dipartimento di Scienze Umane e Sociali e Patrimonio Culturale (DSU) del Consiglio Nazionale delle Ricerche, coinvolti come parte attiva in H2IOSC.[4]
Il GIR nasce dall'esigenza di avere a disposizione una terminologia di riferimento all'interno del progetto H2IOSC. Nella redazione dei documenti è emersa, infatti, fin da subito una difficoltà profonda legata alla mancanza di una terminologia univoca, condivisa e non ambigua, soprattutto nell'ottica di una piena interoperabilità dei risultati della ricerca. Non tutti i ricercatori e i tecnologi coinvolti nel progetto hanno una buona padronanza della terminologia specialistica, e la corretta comprensione dei testi, insieme ad una efficace comunicazione tra le IR, è imprescindibile e può essere ottenuta soltanto ricorrendo a lemmi di uso più comune o esplicitando il significato del lessico tecnico-scientifico. Ad oggi mancano risorse lessicografiche a cui poter fare riferimento per la comprensione di questo linguaggio, e gli stessi vocabolari specialistici non sempre registrano le accezioni tecniche che alcuni lemmi possono assumere nell'ambito delle Infrastrutture di Ricerca.
Da qui la necessità e l'obiettivo di costruire un glossario per definire i termini che appartengono a questo specifico ambito e sciogliere i numerosi acronimi presenti, in modo da documentare un linguaggio tecnico e scientifico che risulta ampiamente condiviso all'interno delle IR, e che è presente anche in altri ambiti istituzionali, come ad esempio quello accademico.

2. METODOLOGIA

A partire da una rassegna critica riguardo i più recenti studi teorici e applicati nel campo della terminologia della lingua italiana,[5] del lessico delle IR,[6] del PNRR[7] e della progettazione europea, è stata individuata una metodologia per la costruzione del *Glossario* che ha previsto prima di tutto l'identificazione di un *corpus* di riferimento da utilizzare per la redazione delle voci. Il *corpus* è composto da testi settoriali in lingua italiana e inglese e comprende piani nazionali, linee guida, capitolati e altra documentazione tecnica.[8]
L'interrogazione del *corpus* avviene tramite TIGRO (Gestore Ricerche del Tesoro Italiano delle Origini), un *software* per l'analisi lessicografica e linguistica di *corpora* testuali, che si basa su GATTO (Gestione degli Archivi Testuali del Tesoro delle Origini), il programma che gestisce la versione elettronica dei *corpora* dell'Opera del Vocabolario Italiano (OVI).[9] Nello specifico TIGRO consente di fare ricerche all'interno dei testi, di visualizzare i contesti e di mostrare le statistiche di frequenza delle forme. La documentazione richiesta dal PNRR è perlopiù in inglese, ma poiché ad essa si devono necessariamente affiancare anche

[1] https://www.h2iosc.cnr.it/ (cons. 08/04/2025). Cfr. Degl'Innocenti *et al.,* 2023.

[2] Common Language Resource and Technology Infrastructure (CLARIN), Digital Research Infrastructures for the Arts and Humanities (DARIAH), European Research Infrastructure for Heritage Science (E-RIHS) e Open Scholarly Communication in the European Research Area for Social Sciences and Humanities (OPERAS).

[3] Cfr. Regolamento (CE) n. 723/2009 del Consiglio del 25 giugno 2009, art. 2, lett. a).

[4] Al *Glossario delle Infrastrutture di Ricerca* collaborano, oltre a chi scrive, Antonina Chaban (Istituto Nazionale di Ottica, CNR-INO), Roberta Ottaviani (Istituto di Linguistica Computazionale "Antonio Zampolli", CNR-ILC), Salvatore Prinzi (Istituto per la Storia del Pensiero Filosofico e Scientifico Moderno, CNR-ISPF), Alessia Spadi (Istituto Opera del Vocabolario Italiano, CNR-OVI).

[5] Si veda ad esempio Chiocchetti & Ralli, 2022; Grimaldi & Zanola, 2021.

[6] Cfr. Liburdi, Marras & Russo, 2019.

[7] Si veda in particolare Cozzio, 2022; Vellutino, 2022.

[8] In particolare, sono inclusi tra i documenti ufficiali AgID, 2006; PNRR, 2021; PNIR, 2021-2027; PNSA, 2021-2027; PNR, 2021-2027.

[9] http://www.ovi.cnr.it/ (cons. 08/04/2025).

documenti in italiano, è emersa la necessità di progettare e costruire un *corpus* bilingue. Di conseguenza, una delle criticità che il gruppo di lavoro sta affrontando è la lemmatizzazione multilingue, motivo per cui il *corpus* non possiede ancora annotazioni morfosintattiche (*Part of Speech Tagging*).
Il lemmario è attualmente composto dai tecnicismi più frequenti nel *corpus*, dai termini rappresentativi di ciascuna unità di lavoro del progetto (*work package* o WP), da altri lemmi non frequenti nella documentazione scritta ma particolarmente caratteristici del linguaggio delle Infrastrutture di Ricerca e ricorrenti nella comunicazione orale legata a tale ambito.
Il gruppo di lavoro è consapevole che, ad accezione dei termini più attestati, la selezione delle voci può sembrare arbitraria; essa, tuttavia, risponde alle esigenze comunicative espresse dalle ricercatrici e dai ricercatori impegnati nei progetti infrastrutturali. Date queste premesse, il lemmario del GIR può considerarsi un contenitore sempre aperto ad accogliere voci che di volta in volta si attestano nell'uso.
La struttura delle voci è stata progettata prendendo come riferimento alcuni vocabolari nati direttamente in rete, come il *Tesoro della Lingua Italiana delle Origini* (TLIO), il dizionario storico dell'italiano antico elaborato dall'OVI.[10] La scheda lessicografica prevede, per ciascun lemma, l'indicazione della categoria grammaticale, la marca d'uso,[11] le accezioni che il termine assume nel contesto delle IR, un campo "note" per eventuali integrazioni, che possono essere di varia natura, e le voci correlate, se presenti. Le voci risultano così corredate da un set completo di metadati descrittivi. Le Fig. 1 e 2 illustrano alcuni esempi di schede del GIR:

FAIR

CATEGORIA GRAMMATICALE: acronimo usato come aggettivo invariabile.

MARCA D'USO: istruzione; Pubblica Amministrazione; ricerca.

DEFINIZIONE

Insieme dei requisiti che i dati e i risultati della ricerca devono rispettare per aderire a un modello di scienza aperta, ovvero essere *findable* 'rintracciabili' (devono essere reperibili, ossia descritti con metadati appropriati e identificativi univoci e persistenti), *accessible* 'accessibili' (devono essere resi accessibili ad altri e persistenti nel tempo), *interoperable* 'interoperabili' (devono poter essere integrati e utilizzati con altri dati o strumenti e in diversi contesti) e *reusable* 'riutilizzabili' (devono essere riutilizzabili e dunque ben descritti e documentati).

NOTE

L'acronimo inglese FAIR (*Findable*, *Accessible*, *Interoperable*, *Reusable*) viene coniato nel 2014, in occasione di un workshop Lorentz a Leida, nei Paesi Bassi. Negli anni successivi, un gruppo di esperti lavora a una lista di principi guida utili a garantire un uso ottimale dei dati della ricerca e nel 2016 vengono pubblicati ufficialmente sulla rivista «Nature» i 15 Principi FAIR (*FAIR Guiding Principles for scientific data management and stewardship*).

VOCI CORRELATE: *fairificazione*, *fairificare*.

Figura 1. Scheda della voce FAIR

[10] http://tlio.ovi.cnr.it/TLIO/ (cons. 08/04/2025).

[11] Posto che tutti i vocaboli del GIR sono legati a un uso marcatamente tecnico-specialistico, la marca d'uso indica gli ambiti specifici in cui ciascun termine risulta attestato, indipendentemente dall'accezione che assume nel contesto delle IR. La difficoltà nel rintracciare tutti i contesti d'uso dei tecnicismi sommata all'eterogeneità dei settori scientifici dei redattori rende inevitabile il rischio che l'indicazione della marca d'uso possa essere talvolta lacunosa.

INFRASTRUTTURA DI RICERCA (IR)

CATEGORIA GRAMMATICALE: locuzione sostantivale femminile.

MARCA D'USO: istruzione; Pubblica Amministrazione; ricerca.

DEFINIZIONE

Organizzazione che comprende strutture, risorse e servizi collegati che possono essere usati dalla comunità scientifica per condurre ricerche di alta qualità e promuovere l'innovazione nei rispettivi settori. Le Infrastrutture di Ricerca includono impianti, laboratori di ricerca, banche dati, strumenti e tecnologie scientifiche, risorse informatiche e reti di comunicazione. Possono essere utilizzate, oltre che per la ricerca, anche per l'istruzione e per i servizi pubblici e possono trovarsi in un'unica sede, essere distribuite in più sedi o essere virtuali.

NOTE

Si fa qui riferimento alla definizione proposta dalla Commissione europea (CE) nel Regolamento (CE) N. 723/2009 del Consiglio del 25 giugno 2009, art. 2, e a quella del Regolamento (UE) n. 1291/2013 del Parlamento Europeo e del Consiglio dell'11 dicembre 2013, art. 2, che istituisce il programma quadro di ricerca e innovazione (2014-2020).

Figura 2. Scheda della voce INFRASTRUTTURA DI RICERCA (IR)

3. LAVORI FUTURI

I lavori futuri prevedono l'ampliamento della bibliografia e la pubblicazione online sia del *Glossario delle Infrastrutture di Ricerca* sia del *corpus* di riferimento. Il passaggio successivo consisterà nella mappatura dei metadati attraverso SKOS (S*imple Knowledge Organisation System*), un linguaggio formale creato per rappresentare glossari, classificazioni, tassonomie e qualsiasi tipo di vocabolario strutturato. SKOS fornisce uno standard per la descrizione delle relazioni fra i sistemi di organizzazione della conoscenza nel quadro del web semantico. L'adozione di tale linguaggio aumenta a nostro avviso il valore del *Glossario*, trasformandolo in una risorsa di conoscenza interoperabile, accessibile e riutilizzabile.
Auspichiamo che il GIR e la banca dati di riferimento per la sua redazione possano rappresentare uno strumento utile per la comunità scientifica che ruota attorno alle IR, oltre che fungere da base per vari studi di tipo linguistico.

RINGRAZIAMENTI

Progetto H2IOSC - Humanities and cultural Heritage Italian Open Science Cloud finanziato dall'Unione europea NextGenerationEU - Piano Nazionale di Ripresa e Resilienza (PNRR) - Missione 4 "Istruzione e Ricerca" Componente 2 "Dalla ricerca all'impresa" Linea di Investimento 3.1 "Fondo per la realizzazione di un sistema integrato di infrastrutture di ricerca e innovazione" Azione 3.1.1 "Creazione di nuove IR o potenziamento di quelle esistenti che concorrono agli obiettivi di Eccellenza Scientifica di Horizon Europe e costituzione di reti" - Codice progetto IR0000029 - CUP B63C22000730005. Soggetto attuatore CNR.

BIBLIOGRAFIA

Agenzia per l'Italia Digitale (AgID). 2006. Linee Guida recanti regole tecniche per l'apertura dei dati e il riutilizzo dell'informazione del settore pubblico. Art. 12 D.Lgs. n. 36/2006. https://www.agid.gov.it/sites/default/files/repository_files/lg-open-data_v.1.0_1.pdf.

Chiocchetti, E., & Ralli, N. (a cura di). 2022. Risorse e strumenti per l'elaborazione e la diffusione della terminologia in Italia. Bolzano: Eurac Research.

Cozzio, M., (a cura di). 2022. Conoscere il PNRR. 150 parole chiave per capire regole, strumenti e funzionamento del Piano Nazionale di Ripresa e Resilienza. Santarcangelo di Romagna (RN): Maggioli Editore.

Degl'Innocenti, E., Monachini, M., Bucciero, A., Pasini, E., Fanini, B. & Frontini, F. (2023). H2IOSC: Humanities and Heritage Open Science Cloud. In Carbé, E., Lo Piccolo, G., Valenti, A., Stella, F. (a cura di), La memoria digitale: forme del testo e organizzazione della conoscenza. Atti del XII Convegno Annuale AIUCD (pp. 63-64). https://iris.unive.it/retrieve/0f226d38-e332-418b-9b14-d5558d1a0d9d/AIUCD2023.pdf

Grimaldi, C., & Zanola, M. T. (a cura di). 2021. Terminologie e vocabolari. Lessici specialistici e tesauri, glossari e dizionari. Firenze: Florence University Press.

Liburdi, A., Marras, C., & Russo, A., (2019). Infrastrutture, terminologie e policy per la ricerca umanistica: note per un confronto interdisciplinare. Conferenza GARR 2018 - Data (R)evolution - Selected Papers, 82-86. DOI: 10.26314/GARR-Conf18-proceedings.

Piano Nazionale di Ripresa e Resilienza (PNRR) 2021. https://www.governo.it/sites/governo.it/files/PNRR.pdf

Piano Nazionale Infrastrutture di Ricerca (PNIR) 2021-2027. https://www.mur.gov.it/sites/default/files/2021-10/Decreto%20Ministeriale%20n.1082%20del%2010-09-2021%20-%20PNIR%202021%20-%202027.pdf.

Piano Nazionale per la Scienza Aperta (PNSA) 2021-2027. https://www.mur.gov.it/sites/default/files/2022-06/Piano_Nazionale_per_la_Scienza_Aperta.pdf

Programma nazionale per la ricerca (PNR) 2021-2027. https://www.mur.gov.it/sites/default/files/2021-01/Pnr2021-27.pdf.

Vellutino, D. (2022). Neologismi istituzionali per la ripresa e resilienza dell'Italia. Analisi tipologico-strutturale dei nuovi termini del PNRR. AIDAinformazioni, 3-4, luglio-dicembre, 113-14.

Educational Impact of Storytelling and Data Visualisation in the Interpretation of Humanities Data

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

ABSTRACT (ENGLISH)
The paper explores the role of storytelling in Digital Humanities education, focusing on how students engage with data to create meaningful narratives. Using MELODY, a web-based tool for user-generated data storytelling, students integrated textual and visual elements to interpret and communicate music heritage. This study emphasises the pedagogical importance of fostering user-driven storytelling, arguing that Humanists benefit from approaches that restore complexity and dignity rather than oversimplifying knowledge. Insights from student-created data stories illustrate how narrative-driven tools can bridge the gap between technical skills and interpretative thinking, without the need to hide complexity.
Keywords: Digital Humanities; storytelling; data visualisation; pedagogy

ABSTRACT (ITALIANO)
Impatto educativo dello storytelling e della visualizzazione dei dati nell'interpretazione dei dati umanistici.
L'articolo esplora il ruolo dello storytelling nell'educazione alle Digital Humanities, concentrandosi su come gli studenti interagiscono con i dati per creare narrazioni significative. Utilizzando MELODY, uno strumento web-based per lo storytelling di dati generato dagli utenti, gli studenti hanno integrato elementi testuali e visivi per interpretare e comunicare il patrimonio musicale. Questo studio sottolinea l'importanza pedagogica di promuovere uno storytelling guidato dagli utenti, sostenendo che gli studiosi delle Humanities traggono beneficio da approcci che restituiscono complessità e dignità, piuttosto che semplificare eccessivamente il sapere. Le storie di dati create dagli studenti illustrano come gli strumenti narrativi possano colmare il divario tra competenze tecniche e pensiero interpretativo, senza la necessità di nascondere la complessità.
Parole chiave: Digital Humanities; storytelling; visualizzazione di dati; pedagogia

1. INTRODUCTION
Storytelling has been a cornerstone of human culture, shaping how we share, understand, and preserve knowledge. In Digital Humanities (DH), the narrative form offers a powerful method for interpreting and disseminating data, particularly for students in their first steps into complex datasets. Data visualisation plays a pivotal role in this process, serving not merely as a tool for presentation or summarisation but as a medium for exploration and critical interpretation. Visualisation in the Humanities must transcend its roots in empirical disciplines to embrace ambiguity and the interpretive nature of humanistic inquiry. This involves acknowledging the situated, partial, and constructed character of knowledge (Drucker, 2011). Interpretation and subjectivity are central to humanities research, setting it apart from the precision-driven methodologies often associated with the sciences. Rather than striving for objectivity, the goal is to make interpretive perspectives explicit, fostering transparency and critical engagement (van den Berg et al., 2018). Tools like MELODY (Renda et al., 2023) enable students to create narratives that combine text and visuals, emphasising human interpretation. By contextualising and transforming data into meaningful narratives, MELODY supports the creation of user-driven storytelling that restores complexity and critical inquiry. In this article we analyse student projects created and published online using MELODY[1] to explore the educational value of integrating storytelling into Digital Humanities introductory courses. Our analysis focuses on the interplay of text and visualisations that structure the narrative flow and contextualise data insights. Findings highlight how narrative-driven tools encourage the curation of presentational aspects of data stories, fostering critical engagement and creativity while supporting the development of interpretative skills.

2. LITERATURE REVIEW
Semantic web technologies offer powerful tools for querying and structuring data. However, their technical complexity often creates a barrier for humanities students, whose academic backgrounds may not align

[1] https://melody-data.github.io/stories/ (last visited 08/01/2025)

with the precision-driven nature of these tools. Traditional pedagogical approaches to teach semantic web technologies frequently prioritise technical skill acquisition, leaving limited space for creativity or interpretive exploration.
Storytelling has emerged as an effective way to contextualise the use of semantic web technologies in the humanities. It provides a narrative framework that enables students to bridge the gap between technical expertise and critical thinking, fostering a deeper engagement with the data they work with (McDowell & Turk, 2024). Project-based approaches to storytelling allow students to connect personally with their work (D'Ignazio & Bhargava, 2018; Bowler et al., 2017), using semantic web queries as the foundation for building interpretive and meaningful narratives (Evenstein Sigalov & Nachmias, 2023).
In this context, data visualisation becomes a key component of storytelling, helping students to transform complex, query-derived datasets into accessible and engaging outputs (Gershon & Page, 2001). Tools like MELODY (Renda et al., 2023) facilitate this process by combining semantic web querying with narrative-driven visualisation, allowing students to explore and present their data in creative and interpretive ways. This integration of storytelling into semantic web education not only demystifies the technologies but also enhances data literacy by emphasising the ethical and interpretive dimensions of data communication (McDowell, 2021; McDowell & Turk, 2024).
The Humanities' emphasis on subjectivity and situated knowledge aligns naturally with these narrative approaches (Boyd Davis et al., 2021; Drucker, 2011). By embedding storytelling into the use of semantic web tools, students are encouraged to make their interpretive perspectives explicit, countering the objectivity often associated with technological methodologies (Vance et al., 2022; Allington et al., 2016; van den Berg et al., 2018). This pedagogical model underscores the potential of semantic web technologies to support not just technical skill acquisition but also the interpretive richness characteristic of humanities research. However, to the best of our knowledge, empirical studies analysing and providing evidence for aforementioned theses are missing in the literature.

3. METHODOLOGY

This study is conducted within a master's program in Language and Communication, specifically during an introductory course on Semantic Web technologies for the Humanities. Students were tasked with creating data stories centered on music heritage[2]. These projects required students to apply SPARQL querying techniques while integrating their findings into cohesive narratives. The goal of course is to combine technical data exploration with interpretative storytelling to address research questions effectively. The MELODY platform[3] was selected as the primary tool to support this process. MELODY is a user-friendly platform designed for data storytelling, offering a what-you-see-is-what-you-get (WYSIWYG) interface that allows the integration of SPARQL queries and data visualisations. It enables users to create narratives enriched with charts, maps, tables, and text search functionalities (Renda et al., 2023). So created stories can be published for free on the web, and the backbone of stories, including queries, types of charts, and texts can be retrieved as JSON files in a dedicated repository[4].
Data collection involved two main sources. First, the students' data stories were harvested from the repository and analysed to evaluate the use of storytelling elements, including the frequency and distribution of textual and visual components, as well as the creative integration of MELODY's features. Secondly, qualitative feedback from students of the aforementioned course were collected after the exam (Renda & Daquino, 2024), and provided us with insights into their experiences, highlighting challenges encountered, skills developed, and reflections on the pedagogical value of storytelling.

4. FINDINGS

Fig. 1 presents the sequential distribution of element types across 51 data stories, with each row representing a unique story and the x-axis denoting the position of elements in the narrative. Four element categories are visualised: text (green), counters (red), data visualisation (blue), and text search (orange). The heatmap reveals that text elements are predominantly placed at the beginning of most stories, suggesting their foundational role in setting the narrative context. These text elements are often followed by data visualisations or counts, which provide supporting data to develop the narrative further. Notably, only 3 out of the 51 stories do not include any text elements, underscoring the critical importance

[2] https://melody-data.github.io/stories/ (last visited 08/01/2025)

[3] https://projects.dharc.unibo.it/melody/ (last visited 08/01/2025)

[4] https://github.com/melody-data/stories (last visited 08/01/2025)

of text as a narrative tool. Among the 48 stories that do include text, 38 contain additional text elements beyond the initial one, indicating its recurring role in elaborating, transitioning, or concluding the narrative (notice that Fig. 1 includes the latter stories only for the sake of the analysis).



Figure 1. Sequential distribution of text, visualisation, count, and search elements across data stories.

Building on these observations, we analysed the length of text elements to understand their variability and narrative role. Fig. 2 shows that text lengths are right-skewed, with shorter texts dominating the distribution. The most frequent range of text lengths is between 174 and 311 characters, representing over 70 instances. Medium-length texts (311–1131 characters) are present but less frequent, with a gradual decline in occurrence as text length increases. Very long texts, exceeding 1400 characters, are rare, with only a few instances surpassing 2000 characters. This distribution highlights a narrative strategy that favours brevity, possibly to ensure clarity and maintain audience engagement. The prevalence of shorter texts may reflect their use in providing quick explanations or transitions, while medium and long texts are reserved for more elaborate contextualisation or storytelling needs.

Given the importance of text as a closing element in data stories, we further analysed the 12 stories where text appears in the final position. Fig. reveals a slight negative correlation between story length (number of elements) and the length of the concluding text. As stories become longer, the length of the final text element tends to decrease. While most data points cluster around text lengths of 600 to 700 characters for shorter stories, longer stories exhibit a broader range of shorter final text lengths, with several falling below 400 characters. This trend may reflect a narrative strategy where longer stories, already rich in data and visualisations, require more concise conclusions to avoid redundancy and maintain focus.

These findings align with student feedback (Renda & Daquino, 2024), which emphasised the importance of text in providing context and guiding the narrative flow. For example, students reported using text to complement data retrieved from queries and to explain their findings, ensuring that complex information was accessible and meaningful. This feedback underscores the observed patterns, highlighting how text and visualisations work together to create cohesive and engaging data stories.

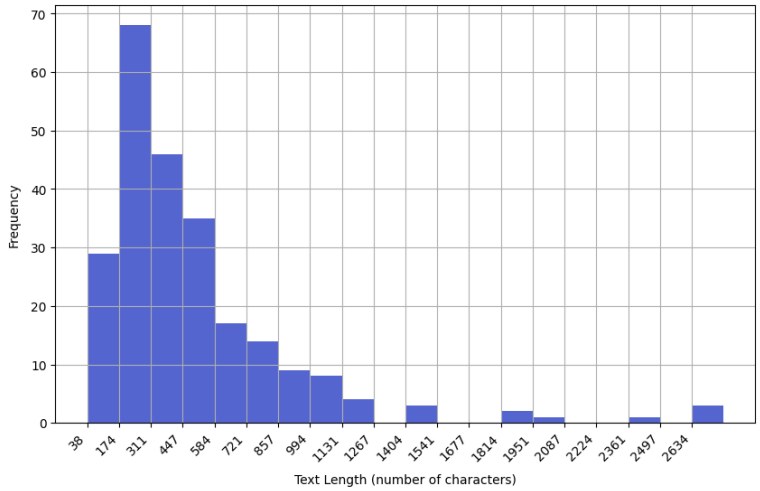


Figure 2. Distribution of text lengths in data stories.

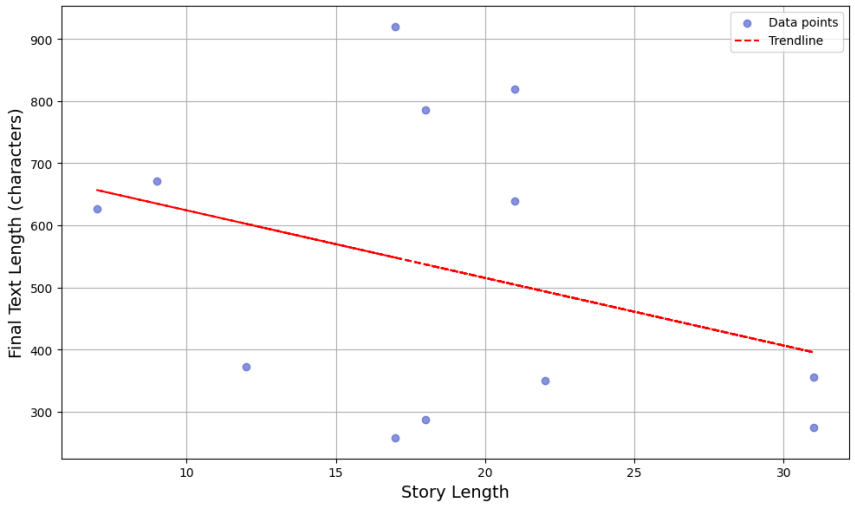


Figure 3. Negative correlation between story length and concluding text length.

5. DISCUSSION

The findings of this analysis, combined with feedback from students, reveal key insights into how textual and visual elements interact to create engaging and effective narratives, the challenges faced in the storytelling process, and the educational value of such projects.

The results highlight the central role of text in framing data stories, with most narratives beginning with a textual element to provide context or set the stage. Students consistently noted that text was used to complement data retrieved from queries, often incorporating additional information from external sources to provide a comprehensive overview of the topic. Text also played a pivotal role in explaining and contextualising data, ensuring that complex information was easily understood by the audience. By combining SPARQL queries with narrative-driven visualisation, MELODY facilitated the integration of semantic web technologies into interpretative storytelling, bridging the gap between technical expertise and humanistic inquiry.

In addition to text, data visualisations emerged as a key component in structuring and enhancing narratives. Students highlighted that visual representations not only helped to present insights but also provided a framework for the overall story. Visualisations were described as a starting point that guided the structure and linearity of the article, emphasising their importance in organising and delivering the narrative. Visualisations served not only as tools for presenting insights but also as interpretative devices

that guided students in structuring their narratives and making complex data accessible to diverse audiences.

A notable theme in the feedback was the emphasis on empathy and audience-centered storytelling. Students expressed the importance of tailoring content to meet the needs of the audience, focusing on relevance and clarity. For example, they considered how the materials could be made more useful for readers, which also inspired creativity in the process. This aligns with best practices in storytelling, where understanding the audience's perspective is critical for creating impactful narratives that resonate.

The project provided significant educational and creative value, as reflected in student feedback. Participants reported gaining new skills, deepening their understanding of data storytelling, and exploring intersections between data, technology, and specific themes such as music. For many, the process of building a data story was an iterative and rewarding learning experience that enhanced both technical and creative abilities. A tool like MELODY was particularly praised for its ability to uncover insights and facilitate engaging storytelling. The iterative storytelling process supported by MELODY enabled students to manage and narrate complex datasets effectively, demonstrating the value of narrative frameworks in simplifying and contextualising intricate information.

Students also recognised the value of storytelling elements in making data narratives more memorable. By incorporating a mix of text and visuals, they were able to craft stories that effectively conveyed key messages while maintaining reader engagement. This iterative process of crafting narratives, refining visuals, and integrating textual elements allowed students to develop an advanced understanding of how to communicate complex information.

While the feedback was largely positive, it also pointed to some challenges in the storytelling process. Students noted the effort required to iteratively refine their stories as they progressed through successive queries. Balancing textual and visual elements to achieve clarity and engagement was another recurring theme. Despite these challenges, the process was seen as a valuable learning opportunity that allowed students to experiment with the Semantic Web. This underscores the pedagogical potential of narrative-driven tools to foster data literacy and interpretative skills, making them invaluable for modern humanities education.

6. CONCLUSION

This study highlights the pedagogical value of narrative-driven tools like MELODY in fostering critical and interpretative skills within Digital Humanities introductory courses. By combining text and visual elements, students were able to craft engaging and effective narratives, enabling them to explore and communicate complex datasets creatively. The central role of text in structuring and contextualising data stories is deemed pivotal by students, and visualisations demonstrated having a complementary function in enhancing narrative coherence and audience engagement.

The iterative process of creating these stories not only fosters students' technical proficiency in semantic web technologies but also emphasises the importance of balancing textual and visual components to convey meaning effectively. While challenges related to refining narratives and managing incomplete data sources were noted, the process was universally regarded as an enriching educational experience that bridged technical learning with humanistic inquiry.

Overall, this research reaffirms the value of storytelling as a framework for managing complexity, fostering empathy, and restoring interpretive richness to data communication in Digital Humanities. Further studies might focus on refining user interfaces and interaction design in storytelling tools to lower barriers for non-technical users while promoting creative engagement with data.

ACKNOWLEDGEMENTS

This work was partially funded by Project PE 0000020 CHANGES - CUP J33C22002850006, NRP Mission 4 Component 2 Investment 1.3, funded by the European Union - NextGenerationEU.

REFERENCES

Allington, D., Brouillette, S., & Golumbia, D. (2016). *Neoliberal Tools (and Archives): A Political History of Digital Humanities*. Los Angeles Review of Books. https://lareviewofbooks.org/article/neoliberal-tools-archives-political-history-digital-humanities

Bowler, L., Acker, A., Jeng, W., & Chi, Y. (2017). "It lives all around us": Aspects of data literacy in teen's lives. *Proceedings of the Association for Information Science and Technology*, *54*(1), Article 1. https://doi.org/10.1002/pra2.2017.14505401004

Boyd Davis, S., Vane, O., & Kräutli, F. (2021). Can I believe what I see? Data visualization and trust in the humanities. *Interdisciplinary Science Reviews*, *46*(4), 522–546. https://doi.org/10.1080/03080188.2021.1872874

D'Ignazio, C., & Bhargava, R. (2018). Creative Data Literacy: A Constructionist Approach to Teaching Information Visualization. *Digital Humanities Quarterly*, *12*(4). https://dspace.mit.edu/handle/1721.1/123473

Drucker, J. (2011). Humanities Approaches to Graphical Display. *Digital Humanities Quarterly*, *5*(1). https://www.digitalhumanities.org/dhq/vol/5/1/000091/000091.html

Evenstein Sigalov, S., & Nachmias, R. (2023). Investigating the potential of the semantic web for education: Exploring Wikidata as a learning platform. *Education and Information Technologies*, *28*(10), 12565–12614. https://doi.org/10.1007/s10639-023-11664-1

Gershon, N., & Page, W. (2001). What storytelling can do for information visualization. *Communications of the ACM*, *44*(8), 31–37. https://doi.org/10.1145/381641.381653

McDowell, K. (2021). Storytelling wisdom: Story, information, and DIKW. *Journal of the Association for Information Science and Technology*, *72*(10), Article 10. https://doi.org/10.1002/asi.24466

McDowell, K., & Turk, M. J. (2024). Teaching data storytelling as data literacy. *Information and Learning Sciences*, *125*(5/6), Article 5/6. https://doi.org/10.1108/ILS-06-2023-0068

Renda, G., & Daquino, M. (2024). *MELODY Data Stories and Anonymous Student Feedback Dataset* [Dataset]. Zenodo. https://doi.org/10.5281/zenodo.14499992

Renda, G., Daquino, M., & Presutti, V. (2023). Melody: A Platform for Linked Open Data Visualisation and Curated Storytelling. *Proceedings of the 34th ACM Conference on Hypertext and Social Media*, 1–8. https://doi.org/10.1145/3603163.3609035

van den Berg, H., Betti, A., Castermans, T., Koopman, R., Speckmann, B., Verbeek, K., van der Werf, T., Wang, S., & Westenberg, M. A. (2018). A Philosophical Perspective on Visualization for Digital Humanities. *3Rd Workshop on Visualization for the Digital Humanities*. https://dare.uva.nl/search?identifier=8b14d5f6-956b-4947-9c56-85e478466c92

Vance, E. A., Glimp, D. R., Pieplow, N. D., Garrity, J. M., & Melbourne, B. A. (2022). Integrating the Humanities Into Data Science Education. *Statistics Education Research Journal*, *21*(2), Article 2. https://doi.org/10.52041/serj.v21i2.42

IncluInstIT: Un nuovo corpus per lo studio di linguaggio inclusivo su Instagram

Irene Caiazzo[1], Giovanna Maria Dimitri[2], Liana Tronci[3]
[1] Università per Stranieri di Siena, Italia – i.caiazzo@dottorandi.unistrasi.it
[2] Università degli Studi di Siena, Italia - giovanna.dimitri@unisi.it
[3] Università per Stranieri di Siena, Italia - tronci@unistrasi.it

ABSTRACT (ITALIANO)
Il contributo presenta una ricerca, tuttora in corso, di raccolta e analisi del corpus IncluInstIT, composto da forme di linguaggio inclusivo di genere presenti nella piattaforma social Instagram. Il corpus è stato elaborato attraverso metodologie semiautomatiche sia durante la fase di iniziale di raccolta dei post per la creazione del dataset, che nel successivo trattamento dei dati scaricati per selezionare i post pertinenti alla ricerca e rendere tutti i contenuti testuali - sia delle didascalie sia quelli contenuti nelle immagini - analizzabili da software di linguistica dei corpora.
Lo scopo del progetto è, da un lato, indagare come un luogo di dialogo digitale quale è Instagram possa essere un contenitore di usi linguistici innovativi, che si suppone non si trovino in tale quantità e varietà in altre piattaforme né in altre tipologie testuali (per esempio articoli di giornale o produzioni letterarie, ma neanche in produzioni orali) e, dall'altro, creare una metodologia di raccolta dati per costruire dataset relativi a casi di studio di linguaggio inclusivo che possa essere replicabile, anche per altre piattaforme e per altre lingue.
Parole chiave: linguaggio inclusivo; sessismo linguistico; linguistica dei corpora; italiano digitale

ABSTRACT (ENGLISH)
IncluInstIT: A New Corpus for Studying Inclusive Language on Instagram
The paper presents an ongoing research project focused on the collection and analysis of the IncluInstIT corpus, which comprises instances of gender-inclusive language found on the social media platform Instagram. The corpus was developed using semi-automatic methodologies both during the initial phase of post collection for dataset creation and in the subsequent processing of downloaded data to select posts relevant to the research. These methods also ensured that all textual content—whether captions or text embedded in images—was rendered analyzable by corpus linguistics software.
The project purses two primary goals: first, to investigate how a digital space for dialogue like Instagram can serve as a repository for innovative linguistic practices, which are hypothesized to occur in a quantity and variety not typically found on other platforms or in other textual genres (e.g., newspaper articles, literary productions, or even oral discourse); and second, to develop a data collection methodology for constructing datasets on instances of inclusive language that can be replicated for other platforms and languages.
Keywords: inclusive language; linguistic sexism; corpus linguistics; digital Italian

1. STATO DELL'ARTE
Il linguaggio inclusivo di genere - che può essere definito come l'uso di strategie linguistiche per rappresentare, attraverso la lingua, le diverse identità di genere - è ormai stato ampiamente trattato nei suoi aspetti teorici, sia per quanto riguarda la rappresentazione del femminile (a partire almeno da Sabatini 1986, 1987 e Violi 1986 fino a lavori più recenti come Giusti 2022, Fusco 2024, Robustelli 2024), sia con riferimento al tema del superamento del binarismo di genere e alla rappresentazione delle identità non binarie (per una panoramica sulla teorizzazione del linguaggio inclusivo in italiano cfr. Gheno 2022, Formato e Somma 2023, Formato 2024). Accanto alla riflessione teorica, quest'ultimo aspetto ha avuto anche dei risvolti pratici, con le proposte di implementazione delle forme di espressione linguistica finalizzate alla neutralizzazione del genere attraverso l'uso di segni quali asterisco (*), chiocciola (@), trattino basso (_), *-x*, o di possibili suoni della lingua quali *-u* o schwa, anche distinto per singolare (*-ə*) e plurale (*-3*) (cf. Comandini 2021 per una panoramica). Al momento, però, sono ancora pochi gli studi che analizzano come queste discussioni si riflettano nella pratica quotidiana della lingua. Comandini (2021) ha creato il primo corpus italiano (CoGeNSI) annotando manualmente testi prodotti su pagine Facebook queer al fine di analizzare le strategie di neutralizzazione del genere nelle loro regolarità e irregolarità e dare una prima tassonomia delle modalità d'uso di queste strategie dividendole in (a) riferimento alla prima

persona, (b) riferimento ad altre persone, (c) uso politico, ovvero l'uso che alcune persone o organizzazioni fanno di queste strategie per posizionarsi rispetto alle tematiche di inclusione.
Facendo tesoro dello studio di Comandini (2021) abbiamo deciso di creare un corpus coerente dal punto di vista del tipo di interrogazione ma diverso nelle forme e nei metodi di raccolta. Innanzitutto, è diversa la scelta del social media di riferimento: abbiamo scelto Instagram invece di Facebook in quanto, sebbene sia nato come piattaforma di condivisione di immagine, presenta anch'esso una componente linguistica - la didascalia, da sempre presente nella struttura del post - e ad oggi presenta contenuti non solo fotografici e autobiografici, ma anche di carattere divulgativo, anche nell'ambito del linguaggio inclusivo.[1] Un aspetto che rende di interesse la piattaforma per il nostro argomento è la sua demografia, più giovane rispetto a Facebook, e tendenzialmente più interessata e aperta a queste tematiche.[2] L'ultima ragione che ha spinto a questa scelta sono le modalità di raccolta dei post offerte dall'API utilizzata, *instaloader*[3]: la libreria infatti permetteva di non dover concentrare l'attenzione su pagine specifiche - come fatto nello studio di Comandini (2021) ma ha dato la possibilità di avere un ampio numero di autori di post, avendo scelto come filtro di selezione la presenza di hashtag e quindi potendo fare un'analisi che raccogliesse il maggior numero di utenti possibile.

2. MODALITÀ DI RACCOLTA

La raccolta dei post è stata eseguita seguendo le indicazioni contenute in Di Cristofaro (2023), dove si individua l'API *instaloader* come strumento di download dei post. Questa API consente di ottenere post attraverso diversi *target* di ricerca che nella loro forma base sono: *profile*, *hashtag* e *location* a cui si possono poi aggiungere ulteriori specificatori come *stories, tagged, comments* etc. (Di Cristofaro 2023, 208). Dunque è stato creato un profilo Instagram apposito per la raccolta dati (@accountphd), in quanto per utilizzare l'API era necessario essere loggati in un profilo e la modalità selezionata è stata la ricerca per hashtag e sono stati scaricati post che contenessero uno dei 21 hashtag selezionati inerenti all'argomento in esame: #femminileinclusivo, #femminilesovraesteso, #femminiliprofessionali, #generelinguistico, #italianoinclusivo, #linguaggioampio, #linguaggiodigenere, #linguaggioinclusivo, #linguaggioneutro, #linguaggiononsessista, #linguaggiosessista, #linguainclusiva, #linguaneutra, #linguanonbinaria, #linguasessista, #maschileinclusivo, #maschileneutro, #maschilenonmarcato, #maschilesovraesteso, #schwa, #sessismolinguistico.
La scelta degli hashtag da usare è stata fatta tenendo conto dei possibili hashtag legati alla tematica presa in esame e verificando l'esistenza di post che li contenessero sulla piattaforma. Oltre ai 21 hashtag elencati, ne sono stati testati altri che non hanno portato al download di alcun post ad esempio #linguaggiogenderizzato, #linguanonsessista, #linguaggiononbinario. Come principio generale per ogni hashtag scaricato che conteneva la parola 'maschile' si è cercato anche il corrispondente con 'femminile' allo stesso modo in cui si sono cercati sia hashtag con 'lingua' che con 'linguaggio'. Per necessità date dalle possibilità materiali e dalle modalità di processing che si sarebbero andate a utilizzare, sono stati esclusi hashtag che avrebbero previsto un numero particolarmente alto di post e dei quali un'alta percentuale non sarebbe stata strettamente attinente alla tematica (ad es. #asterisco, #chiocciola, #ministra, #femminile). Si può vedere l'hashtag #schwa come un'eccezione a questo principio, visto che presenta circa 8400 post, si è tuttavia deciso di conservarlo, considerando che l'uso di questo termine in italiano è strettamente legato al dibattito sul linguaggio inclusivo. Di conseguenza, il successivo filtraggio per lingua avrebbe eliminato i post non pertinenti, mentre quelli rimanenti sarebbero stati rilevanti per il corpus.
Per ogni post la libreria ha scaricato diversi file contenenti tutti i dati ottenibili di default dai post: un file JSON contenente i metadati (tra cui codice univoco del post, il suo url, profilo che ha pubblicato il post, data e ora, numero di "mi piace" e di commenti), un file .txt contenente la didascalia del post, un numero di file JPG o PNG corrispondente al numero di immagini presenti nel post e un file MP4 per i post contenenti video.

[1] La presenza di account che si occupano di divulgazione è stata già notata in diversi lavori, tra cui Bagaglini 2022 e 2024, quest'ultimo con riferimento diretto alla divulgazione di contenuti inerenti alla linguistica.
[2] Dati sull'utenza di Instagram e Facebook si possono trovare ai link: https://www.statista.com/statistics/248769/age-distribution-of-worldwide-instagram-users/ (cons: 10/01/2025)
https://www.statista.com/statistics/376128/facebook-global-user-age-distribution/ (cons: 10/01/2025)
Invece, riguardo alle tendenze delle diverse generazioni rispetto alle tematiche inerenti alle comunità queer: https://www.ipsos.com/it-it/pride-month-2024-genz-propensa-identificarsi-lgbt (cons: 10/01/2025)
[3] https://github.com/instaloader/instaloader (cons: 10/01/2025)

Questo download è durato dal 27/05/2024 al 13/07/2024 ha raccolto file senza limiti cronologici, quindi dal 2011, anno di creazione del social, al giorno stesso dell'ultimo download. La fase di raccolta si è conclusa con l'ottenimento di 12905 post scaricati che sono andati a comporre il dataset su cui si costruirà il corpus. Il numero di post è stato determinato in base al numero di file JSON, mentre i numeri degli altri tipi di file variano; sono stati scaricati 12928 documenti in formato .txt (la differenza tra numero di post e di file .txt, quindi didascalie, si spiega tenendo conto della presenza di didascalie modificate dagli autori dei post successivamente alla loro pubblicazione e che sono state ugualmente scaricate dal programma con la dicitura .old), 25105 in formato JPG, 61 in formato PNG, 2 immagini in formato WEBP (in seguito convertito in JPG per essere aperto dal computer) e 1207 file video in formato MP4.

3. PRE-PROCESSING DEL DATASET

Le modalità della prima fase presentavano dei limiti che hanno fatto sì che fosse necessario svolgere diverse fasi di ripulitura dei dati prima che questi potessero essere analizzati con software di linguistica dei corpora. Questi limiti consistevano principalmente nella presenza di post duplicati, fatto causato dalla presenza di due o più hashtag tra i selezionati e quindi che l'API ha scaricato più di una volta, e di post in lingue diverse dall'italiano.

In questa fase sono stati utilizzati solo i file .txt dei post, i quali sono stati sottoposti a processi sia automatici che manuali per la rimozione dei file non desiderati all'interno del dataset (ad esempio utilizzo del programma *remove duplicate lines*[4] e della API *langdetect*[5]) e che hanno fatto diminuire il numero di post a 4834. Di conseguenza sono diminuiti anche gli altri file a: 11628 JPG, 40 PNG, 331 MP4 e 1 WEBP. La presenza di post duplicati si spiega per la tendenza di inserire più di un hashtag nel post per segnalare l'argomento di cui si sta parlando, caratteristica tipica della piattaforma come si evince da una lettura dei metadati dei post scaricati in cui ci sono anche informazioni sugli hashtag utilizzati. La presenza di lingue straniere invece ha più di una ragione: in alcuni casi è dovuta alla presenza delle stesse parole anche in altre lingue ad esempio con gli hashtag #schwa e #linguaneutra (in questo secondo caso si è rilevato che tutti i post scaricati erano in lingua portoghese e non italiana e quindi sono stati rimossi), in altri invece il post tratta di lingua italiana ma lo fa in una lingua straniera a scopo spesso divulgativo, sono state trovate infatti post di alcune pagine di italiano a stranieri o che diffondono più in generale la cultura italiana all'estero.

4. FASE ATTUALE: GESTIRE LA MULTIMODALITÀ

La fase in corso ha lo scopo di recuperare il contenuto testuale presente all'interno delle immagini che, insieme alle didascalie, compongono i post che comporranno il corpus. Sebbene la ricerca non voglia prendere una svolta multimodale, e quindi più semiotica che strettamente linguistica, si ritiene che tenere conto dei contenuti linguistici nelle immagini sia necessario per tenere conto della caratteristica fondante della piattaforma di riferimento, ovvero il ruolo centrale che ha l'immagine - che sia una foto o che riporti una scritta - nella totalità del post.

Il primo passo per ottenere questo contenuto testuale è stato l'estrazione attraverso l'OCR *tesseract*[6] che ha permesso di creare un file .txt per ogni immagine che gli è stata sottoposta. Questi file sono stati successivamente divisi in file vuoti e file che presentavano contenuto ed è seguito il controllo manuale dei testi creati, per controllare fossero corretti.

Le principali problematiche di questa fase sono state: la difficoltà del software di recuperare il testo in immagini che spesso presentavano molto rumore di sottofondo (ad es. testo sovrapposto ad altre immagini) oppure con font non convenzionali o scrittura a mano e creare delle categorie per distinguere quali contenuti testuali fossero di interesse per la ricerca e quali no.

5. FASI FUTURE E OBIETTIVI DI ANALISI

Le fasi successive della ricerca si svilupperanno in diverse direzioni al fine di indagare in modo completo e in funzione dei vari parametri il corpus creato.

In primo luogo, si svolgerà un'analisi quantitativa dei testi presenti nel corpus, per verificare le tendenze in funzione della frequenza delle forme contenute e delle loro possibili correlazioni, al fine di comprendere quali temi entrano in gioco su Instagram quando si discute di linguaggio inclusivo e in quali modi se ne

[4] https://codebeautify.org/remove-duplicate-lines (cons: 10/01/2025)

[5] https://github.com/Mimino666/langdetect (cons: 10/01/2025)

[6] https://github.com/tesseract-ocr/tesseract (cons: 10/01/2025)

discute. Successivamente, avverrà una marcatura delle forme che presentano strategie di linguaggio inclusivo, al fine di indagarne gli usi e metterli in relazione con altri studi (per es. Comandini 2021). La disponibilità di post che coprono un arco temporale ampio permetterà anche di studiare se e come queste strategie sono state utilizzate nel passato e come vengono usate oggi, se ci sono stati mutamenti nel loro uso e, ad esempio, se le preferenze nell'uso di una o di un'altra forma sono cambiate.
A fianco a questo lavoro diretto sui testi, avverrà anche un'analisi dei metadati raccolti, così da studiare dati come la cronologia dei post, le tendenze nell'uso di hashtag, il numero e il tipo di utenti che hanno pubblicato questi post.
Un'ulteriore linea di ricerca, per ora solo ipotizzata, è l'analisi dei commenti presenti in questi post. Infatti grazie ai metadati già scaricati è possibile vedere quali post hanno ricevuto commenti e in che quantità, tramite la libreria già utilizzata nella raccolta dei post, *instaloader,* si potrebbero scaricare anche i relativi commenti e quindi vedere come anche in questo genere di testo web si usano strategie inclusive e, soprattutto, un tipo di analisi qualitativa potrebbe far emergere le reazioni e le opinioni di un maggior numero di utenti rispetto alla tematica.

BIBLIOGRAFIA

Bagaglini, V. (2022). «LA COMUNICAZIONE SCIENTIFICA SUI SOCIAL NETWORK: UN'ANALISI DELLA SCRITTURA DIVULGATIVA SU TWITTER, FACEBOOK E INSTAGRAM». Italiano LinguaDue 13 (2):310-35. https://doi.org/10.54103/2037-3597/17141.

- 2024. «La Divulgazione Della Linguistica Su Instagram: "La Polifonia Del Discorso Specialistico"». Lingue E Culture Dei Media 7 (1-2):38-67. https://doi.org/10.54103/2532-1803/22392.

Comandini, G. (2021). «Salve a tuttə, tutt*, tuttu, tuttx e tutt@: l'uso delle strategie di neutralizzazione di genere nella comunità queer online. : Indagine su un corpus di italiano scritto informale sul web.» Testo e Senso, fasc. 23 (23 dicembre 2021): 43–64.

Di Cristofaro, M. (2023). Corpus Approaches to Language in Social Media. 1ª ed. New York: Routledge,. https://doi.org/10.4324/9781003225218.

Formato, F., Somma A. L. (2023). «Gender Inclusive Language in Italy: A Sociolinguistic Overview» The Journal of Mediterranean and European Linguistic Anthropology, Vol. 5(1) (2023): 22–40. DOI: 10.47298/jomela/v5-i1-a3 jomela.pub

Formato, F. (2024). Feminism, corpus-assisted research, and language inclusivity. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781009236379

Gheno, V. (2022) Femminili singolari: il femminismo è nelle parole*.* Quarta edizione. Saggi pop 60. Firenze: Effequ.

Giusti, G. (2022). «Inclusività nella lingua italiana: come e perché. Fondamenti teorici e proposte operative», DEP*.* Deportate, esuli, profughe n. 48 / 2022: 1-19.

Ipsos. 2024. Pride Month 2024: la GenZ è la più propensa a identificarsi come LGBT+
https://www.ipsos.com/it-it/pride-month-2024-genz-propensa-identificarsi-lgbt

instaloader
https://github.com/instaloader/instaloader

langdetect
https://github.com/Mimino666/langdetect

remove duplicate lines
https://codebeautify.org/remove-duplicate-lines

Robustelli, C. (2024) «Sessismo Linguistico». Tindara A. et al. (Eds.), Promuovere la parità di genere nelle istituzioni pubbliche e private, 2024, 171–91.

Sabatini, A. (1986). Raccomandazioni per un uso non sessista della lingua italiana. Per la scuola e l'editoria scolastica. Roma: Presidenza del Consiglio dei ministri.

- (1987) Il sessismo nella lingua italiana. Roma: Presidenza del Consiglio dei ministri.

Statista (2024a). Distribution of Facebook users worldwide as of April 2024, by age and gender
https://www.statista.com/statistics/376128/facebook-global-user-age-distribution/

- (2024b). Distribution of Instagram users worldwide as of April 2024, by age and gender
https://www.statista.com/statistics/248769/age-distribution-of-worldwide-instagram-users/

tesseract
https://github.com/tesseract-ocr/tesseract

Violi, P. (1986). L’infinito singolare. Considerazioni sulle differenze sessuali nel linguaggio. Verona: Essedue, 1986.

Per un'analisi della rivista *Umanistica Digitale* in ottica DEI

Rossana Morriello[1], Lucia Sardo[2],
[1] Università degli Studi di Firenze, Italia – rossana.morriello@unifi.it
[2] Alma Mater Studiorum Università di Bologna, Italia – lucia.sardo@unibo.it

ABSTRACT (ITALIANO)
L'intervento presenta il risultato di una analisi sugli articoli della rivista Umanistica Digitale dalla sua fondazione a oggi in un'ottica DEI, considerando che questo è il tema del convegno che sarà organizzato nel 2025. L'analisi ha rilevato una grande ricchezza di tematiche trattate ma una scarsa presenza di articoli dedicati alle tematiche DEI; inoltre, l'utilizzo di parole chiave non uniformate rende difficoltosa la ricerca di testi su argomenti simili e riduce la *discoverability* degli articoli. Il contributo propone l'uso di un vocabolario strutturato per la corretta disseminazione delle tematiche DEI.
Parole chiave: DEI, Umanistica Digitale, keywords, text mining

ABSTRACT (ENGLISH)
For an analysis of Umanistica Digitale journal from a DEI perspective
The paper presents the result of an analysis on the articles of the journal Umanistica Digitale from its foundation to the present day from a DEI perspective, considering that this is the theme of the conference to be organised in 2025. The analysis revealed a great richness of topics treated but a low presence of articles dedicated to DEI topics; furthermore, the use of non-uniform keywords makes it difficult to search for texts on similar topics and reduces the discoverability of the articles. The contribution proposes the use of a structured vocabulary for proper dissemination of DEI topics.
Keywords: DEI, Umanistica Digitale, keywords, text mining

1. INTRODUZIONE
Le keywords svolgono diverse funzioni importanti nell'era dei dati e delle applicazioni dell'intelligenza artificiale, tra cui innanzitutto permettere la *discoverability* e la *findability* delle risorse attraverso i motori di ricerca e nei sistemi di information retrieval delle banche dati. Sono quindi alla base delle strategie di Search Engine Optimization (SEO) per diverse tipologie di risorse. Nell'ambito della ricerca accademica e delle banche dati scientifiche le keywords favoriscono quindi il recupero, la disseminazione e la condivisione dei risultati che è il fine ultimo della ricerca scientifica. Consentono di mettere in contatto ricercatori che si occupano di tematiche affini poiché tramite le parole chiave si possono cercare articoli di argomento simile a ciò che è di interesse, ed eventualmente citarli, contribuendo alla determinazione degli indicatori bibliometrici. La presenza di keywords favorisce l'applicazione di strumenti di data analysis e information retrieval avanzati, sempre più basati sull'intelligenza artificiale.
Per queste ragioni gli editori nelle loro banche dati, siano esse citazionali, testuali, bibliografiche, attribuiscono grande importanza a questa forma di indicizzazione semantica che consente loro di organizzare efficacemente i contenuti e di raggrupparli in cluster per agevolare innumerevoli forme di data analysis, oltre alla *discoverability*, e di aggiornare costantemente l'organizzazione dei contenuti digitali sulla base dell'evoluzione della ricerca scientifica.[1] Gli articoli contenuti nelle principali banche dati scientifiche sono infatti dotati di varie keywords, o key phrases, come sarebbbe più opportuno chiamarle poiché spesso si tratta di frasi composte da più parole (Hartley & Kostoff, 2003), insieme al titolo e all'abstract dell'articolo. In misura crescente, questi elementi paratestuali, e in particolare le key phrases, vengono utilizzate per l'assegnazione della peer review a determinati revisori, tramite il matching tra le keywords di un articolo e le keywords con le quali il revisore ha sintetizzato le sue competenze, come spesso avviene sulle piattaforme editoriali. Per motivi analoghi le keywords sono fondamentali nei progetti di ricerca che si sottopongono per il finanziamento e che richiedono revisione paritaria.
Per quanto riguarda il nostro macrotema, ovvero diversità, equità e inclusione (DEI) le keywords sono tra i metadati che consentono di ricondurre le pubblicazioni agli ambiti DEI, poiché permettono di costruire, insieme agli altri elementi paratestuali, il contesto e il significato necessari al lettore per accedere consapevolmente all'informazione su questi temi. Nell'acronimo DEI, con "diversità" si intende una serie di

[1] Per esempio, è quanto dichiara esplicitamente Elsevier nella banca dati SciVal, che offre soluzioni di research intelligence e data visualization sui dati di Scopus, a cui faremo riferimento più avanti in questo contributo.

differenze socio-economiche e culturali potenzialmente svantaggiose quali disabilità, età, genere, orientamento sessuale, status socio-economico, lingua, pratiche di alfabetizzazione, background educativo, localizzazione geografica, razza, etnia, cultura e religione (Raju 2020; Morriello & Sardo 2023), e l''inclusione" insieme all'"equità" sono aspetti essenziali e integrali ai processi volti all'incremento della diversità nel mondo dell'informazione, della ricerca e dell'educazione (Raju 2020). Implicita nell'acronimo è l'azione svolta per superare le barriere e migliorare la diversità. Dunque, non si tratta solo di mappare il soddisfacimento dei criteri di diversità, equità inclusione in un settore quale può essere il settore o i settori disciplinari di copertura di una rivista scientifica, ma anche e soprattutto di agire per potenziare la diversità. Con tale obiettivo, costruttivo e non certamente critico, proponiamo il nostro contributo.

2. OBIETTIVI E METODOLOGIA

L'intervento presenta un'analisi degli articoli della rivista *Umanistica Digitale* dalla sua fondazione a oggi, per rilevare quanto e come le tematiche riconducibili a diversità, equità e inclusione siano presenti e in che modo vengano indicizzate tramite le keywords.
L'analisi è stata condotta utilizzando le informazioni e i dati presenti sul sito della rivista stessa. Nello specifico sono stati presi in considerazione gli autori e le parole chiave inserite negli articoli, al fine di rilevare le keywords legate ai temi DEI. Inoltre, sono stati rilevati i dati relativi al genere fra coloro che hanno presentato contributi alla rivista.
La raccolta ed elaborazione dei dati è stata effettuata utilizzando il software open source OpenRefine, con una successiva integrazione e verifica manuale nel caso in cui sia reso necessario. A questa fase è seguita l'analisi di quanto emerso dai dati raccolti, con una riflessione particolare sull'uso delle keywords. La sezione finale dell'intervento introduce una proposta per un ripensamento nell'uso delle parole chiave, al fine di facilitare il reperimento di documenti rilevanti e il posizionamento degli studiosi/e nei rispettivi ambiti disciplinari.
I dati analizzati con OpenRefine sono poi stati confrontati con i dati di Scopus e di SciVal. Da settembre 2021, infatti, la rivista *Umanistica Digitale* è indicizzata nella banca dati citazionale Scopus di Elsevier e di conseguenza nella banca dati SciVal, dello stesso editore, che attinge ai dati di Scopus per offrire soluzioni di research intelligence e data visualization.
L'obiettivo di questo lavoro, come sopra anticipato, non è certo una critica delle manchevolezze ma mira piuttosto ad agevolare il rafforzamento della *findability* degli articoli, in particolare in prospettiva DEI, e rientra in un più ampio lavoro di ricerca in corso su alcune riviste scientifiche italiane di ambito umanistico.
Sono stati analizzati tutti gli articoli presenti sulla rivista *Umanistica Digitale* dal primo numero del 2017 al numero 18 del 2024. In totale si tratta di 205 articoli, nei quali sono comprese le introduzioni e le recensioni per quanto riguarda gli autori,[2] e nel caso in cui fossero presenti, le parole chiave significative.[3]

3. RISULTATI

Un primo dato da cui partire riguarda gli autori in una prospettiva di genere. In totale sono stati individuati 327 autori. Di questi, 162 sono donne e 165 uomini. L'analisi degli autori vede perciò una sostanziale parità fra i generi se si considerano tutti i coautori dell'articolo, mentre se si considera solo il primo (o l'unico) autore di un paper, le percentuali vedono le donne con una maggior presenza (54% circa).
Come sovente accade, sia nelle riviste sia nelle banche dati di altro genere, si è riscontrata anche in questo caso una mancanza di uniformità nel modo in cui vengono indicate le istituzioni di affiliazione, per cui non è stato possibile elaborare in questo senso i dati. Si evidenzia la necessità di trovare degli identificatori anche per gli enti per favorire la ricerca e l'individuazione univoca dei medesimi con maggior facilità. Le banche dati commerciali lavorano a uniformare le affiliazioni e presentano dati più precisi a riguardo, tuttavia, il dato non risultava particolarmente rilevante per le finalità del lavoro e si è deciso che fosse sufficiente individuare le affiliazioni alle università e non ai dipartimenti. Anche in questo caso si sono uniformate le modalità citazionali, prima utilizzando OpenRefine e quindi con una ulteriore verifica di eventuali match non riscontrati dal software.

[2] Secondo i dati di Scopus gli articoli rappresentano comunque il 95,9% delle tipologie in *Umanistica Digitale*.
[3] Sono state ovviamente escluse le parole chiave "recensione" o "review."

La maggior parte degli autori ha una affiliazione accademica o legata a centri e fondazioni accademiche; molti lavorano per istituzioni legate alla ricerca (CNR, CNRS, ad esempio); solo pochi lavorano in biblioteche o archivi o in istituti di istruzione superiore; ancora minore è il numero di autori che lavorano per aziende private, amministrazioni statali o locali. Alcuni si qualificano come ricercatori indipendenti.
Gli articoli presenti hanno da 1 a 11 autori. La maggior parte degli autori è responsabile di un solo paper pubblicato in *Umanistica Digitale*, nel senso che ne è l'unico autore o uno dei co-autori.
Utilizzando OpenRefine e il servizio di riconciliazione con Wikidata, è emerso che, in percentuale, il 43% ha un match sicuro, mentre per il restante 57% si tratta di riconciliare manualmente o di creare l'item in Wikidata (in questi casi il sistema di riconciliazione non ha considerato certo il possibile match, qualora presente, o è appunto necessario creare l'item).
Questo dato a nostro avviso è significativo perché implica la necessità di prevedere l'inserimento degli autori (e degli articoli) in Wikidata, al fine di favorire la creazione di LOD (Linked Open Data) a maggior vantaggio di tutta la comunità scientifica.
Rispetto alle parole chiave riscontrate, si è operata una minima uniformazione utilizzate in modo da poterle analizzare e rappresentare graficamente (figura 1).



Figura 1 – Word cloud delle principali parole chiave

I risultati dell'analisi hanno riscontrato l'assenza di keywords in 18 articoli, ma si tratta principalmente di introduzioni, recensioni, etc.
Nei 205 articoli della rivista si contano oltre 750 keyword diverse. Ciascun articolo presenta da un minimo di tre a un massimo di undici parole chiave, con una media di circa quattro keywords per articolo.
L'elemento di attenzione non è costituito tanto dal numero di keywords, fatto questo che potrebbe semplicemente dimostrare l'ampiezza del campo di studio e la grande varietà di tematiche affrontate dalla rivista, quanto la ripetitività, ovvero che molte di esse fanno riferimento sostanzialmente allo stesso ambito semantico, e allo stesso argomento, il quale viene però definito e declinato in modi diversi dai diversi ricercatori, rendendo di fatto impossibile una ricerca efficace per parole chiave, qualora si volesse utilizzare questa modalità per trovare articoli simili.
A titolo d'esempio, si trovano keyword quali: "Topic Model, Topic modeling, Topic modelling", "archivi digitali, archivi nativi digitali, Archivio digitale", codifica TEI, codifica XML-TEI, TEI encoding, XML/TEI, TEI", "Digital Scholarly Edition, digital scholarly editing, Digital Scholarly Editions".[4]

[4] La presenza di keyword in italiano o in inglese è dovuta alle scelte degli autori.

Con pochissime eccezioni non sono presenti keywords riconducibili all'ambito DEI.
I dati analizzati con OpenRefine sostanzialmente coincidono con quanto si può rilevare in Scopus e SciVal, perlomeno dal punto di vista quantitativo. Ciò che differisce è l'organizzazione delle parole chiave che, nei due prodotti commerciali, sono standardizzate e fortemente strutturate. Per quanto riguarda *Umanistica Digitale,* la banca dati SciVal presenta 98 topics e 65 topic clusters, in aree disciplinari concentrate ovviamente in prevalenza tra le scienze umanistiche e le scienze sociali.[5]
SciVal utilizza il suo Fingerprint Engine per attribuire delle keyphrases tramite text mining e varie tecniche di Natural Language Processing applicate ad articoli, abstract e dati bibliografici e confrontate con alcuni thesauri di riferimento, e per l'indicizzazione in Topics, Topic Clusters e Research Areas.
Per fare un esempio, usiamo il Topic *Digital; Critical Edition; Semantic Web* che è il più rilevante in termini di preminenza globale (su base citazionale naturalmente) rispetto alla rivista *Umanistica Digitale* (include dodici articoli) e viene definito nella banca dati come relativo a "the development and impact of digital scholarly editions in humanities research. It encompasses the use of digital methods to enhance access, preservation, and analysis of historical texts and manuscripts". Il Topic presenta le seguenti keyphrases: Digital Edition; Digital; Archival Document; Scholarly edition; Online; Critical edition; Semistructured Data; Review; Scholars; Collaborative; Software; Development Tool; Online Edition; Development Version; Textual Culture; Search Engine.
In sintesi, in aggiunta alle keywords presenti sulla rivista che vengono attribuite dall'autore o dagli autori dell'articolo, e che sono sempre presenti su Scopus, SciVal presenta diversi metodi di indicizzazione propri, arricchendo significativamente il campo semantico di ciascun articolo tramite molteplici keywords o keyphrases.
L'attenzione con cui l'editore della banca dati lavora sulle keywords ne dimostra l'importanza per la possibilità di scoperta e condivisione dei risultati della ricerca scientifica pubblicati. L'uso di diversi metodi massimizza *discoverability* e *findability* dei contenuti e consente innumerevoli efficaci rielaborazioni dei dati, soprattutto alla luce delle tecniche di text and data mining e delle applicazioni crescenti dell'intelligenza artificiale.
Per tornare alla rivista di nostro interesse, l'analisi sui dati di *Umanistica Digitale* in Scopus e SciVal conferma la quasi nulla presenza di tematiche legate a diversità, equità e inclusione. Questo dato viene confermato anche dalla mappatura degli SDG (Sustainable Development Goals) applicata agli articoli della rivista sempre in SciVal. Solo ventidue articoli sono riconducibili agli Obiettivi di sviluppo sostenibile dell'Agenda 2030 (Nazioni Unite 2015). Di questi uno è ricondotto all'SDG 10 – Ridurre le diseguaglianze, quindi direttamente ed esplicitamente legato ai temi DEI, e quattro articoli sono ricondotti all'SDG 16 – Pace, giustizia e istituzioni sane, un goal non direttamente e necessariamente rilevante per DEI ma in realtà i quattro articoli lo sono, se li si va a consultare.

4. CONCLUSIONI

Gli editori sono attenti e quindi in molti casi offrono indicazioni precise agli autori e di norma utilizzano un soggettario o un thesaurus per l'indicizzazione dei contenuti bibliografici digitali. Il modo in cui i thesauri coprono le tematiche DEI sarebbe da indagare ma esula dagli obiettivi di questo contributo. Senz'altro vi sono ormai diversi studi che si soffermano sui temi DEI nella comunicazione scientifica, alcuni dei quali sono stati riportati nell'introduzione. Anche l'Agenda 2030 e l'attribuzione degli SDG ai contenuti scientifici aiuta a individuarli.
L'assenza di parole chiave relative a DEI non implica necessariamente la scarsa attenzione a queste tematiche dagli articoli nelle riviste. Nonostante la loro importanza di "connettori tra l'articolo e il lettore" spesso gli autori sottovalutano l'importanza delle parole chiave, come indicano diversi degli studi citati, e allo stesso tempo in molti casi non vi sono linee guida e indicazioni specifiche da parte degli editori e dei manager delle riviste. Questo uso poco organizzato può avere ricadute sull'attribuzione di significato e sulla ricerca di informazione in generale, e in ambito DEI in particolare. La poca attenzione attribuita dagli autori alle parole chiave non permette di evidenziare tali specificità. L'assenza di keywords legate a DEI potrebbe indicare l'assenza di tali argomenti oppure la non precisa attribuzione di parole chiave. L'uso di

[5] I *Topics,* come spiegato nella banca dati, sono raccolte di pubblicazioni con un interesse intellettuale comune, basato su citazioni dirette, che rappresentano campi di ricerca a livello globale, costantemente in evoluzione in base ai cambiamenti nelle prospettive di ricerca. Una pubblicazione può appartenere a un solo Topic, e un Topic può appartenere a un solo Topic Cluster. La copertura temporale della rivista in Scopus (e quindi in SciVal) va dal 2019 al 2023.

keywords non corrette può essere fuorviante oppure non permettere il recupero dell'articolo; l'uso di keywords che non evidenzino i temi DEI oscura questi aspetti anche se trattati all'interno di un articolo. Tra i punti di attenzione va di certo ricompreso il modo in cui viene scritta la keyword fa la differenza. Per esempio, "person with disability" ha una valenza diversa di "disabled person" in quanto nel primo caso si accentua l'enfasi sulla persona mentre nel secondo caso sulla disabilità, così come "persona con disabilità" ha un'implicazione diversa da "disabile", soprattutto se estrapolata dal contesto (Thompson, 2020).
Un tema cruciale è poi quello linguistico, al quale bisognerebbe senz'altro prestare maggiore attenzione in un contesto scientifico colonizzato dalla lingua inglese (Horvath, A., Wagner, C. Wrisley, D.J., et al. 2024), in quanto la lingua è uno strumento fortemente identitario per una comunità e la convergenza verso un'unica lingua nella comunicazione scientifica non favorisce la diversità.
Tutti questi aspetti sono essenziali nella gestione automatizzata tramite text mining dei contenuti di una rivista o di una banca dati. In generale queste sono questioni alle quali gli editori stanno attribuendo considerazione crescente (APA 2021, Cuellar 2021, Dewidar 2022, Hinton & Lambert 2022; Marabelli & Chan 2024) poiché la centralità di questi temi nella società è di fatto crescente, anche alla luce dell'Agenda 2030, degli effetti della pandemia e di episodi come la morte di George Floyd o il movimento #meetoo (Marabelli & Chan 2024). Tali eventi hanno accresciuto la sensibilità verso queste tematiche, e aumentato perfino gli studi che hanno dimostrato i benefici anche in termini economici di un approccio rivolto alla diversità, all'equità, all'inclusione (McKinsey & Co. 2022). L'evoluzione degli ultimi anni ha messo in luce il lavoro ancora da compiere ma allo stesso tempo ha rappresentato un punto di svolta nelle tematiche DEI, dal quale auspicabilmente non si può tornare indietro.
La proposta è quindi, per tutte le riviste che ancora non lo fanno, di fornire e delle indicazioni precise, adottare linee guida, offrire suggerimenti e uno schema di tematiche DEI in cui far rientrare le keywords, oltre ad incrementare la trattazione di questi argomenti nella rivista, magari attraverso specifiche call for papers così come in questo convegno.

BIBLIOGRAFIA

APA. (2021). Equity, Diversity, and Inclusion Toolkit for Journal Editors, https://www.apa.org/pubs/authors/equity-diversity-inclusion-toolkit.

Cuellar, NG. (2021). Diversity, Equity, and Inclusivity in Peer-Reviewed Journals. *Journal of Transcultural Nursing*, 32(5), 449-450. doi:10.1177/10436596211034902

Dewidar, O., Elmestekawy, N. & Welch, V. (2022). Improving equity, diversity, and inclusion in academia. *Research Integrity and Peer Review*, 7. https://doi.org/10.1186/s41073-022-00123-z

Ferreira Garcia, D.C., Chaves Gattaz, C., & Chaves Gattaz, N. (2019). The Relevance of Title, Abstract and Keywords for Scientific Paper Writing, *Revista de Administração Contemporânea*, 3. http://doi.org/10.1590/1982-7849rac2019190178.

Hartley, J., & Kostoff, Ronald D. (2003). How useful are 'key words' in scientific journals? *Journal of Information Science*, 29, 433-438.

Hinton, A. Jr., & Lambert, M.W. (2022). Moving diversity, equity, and inclusion from opinion to evidence. *Cell Reports Medicine*, 3(4), 100619. https://doi.org/10.1016/j.xcrm.2022.100619

Horvath, A., Wagner, C. Wrisley, D.J., et al. (2024). Multilinguality in Action: Towards Linguistic Diversity and Inclusion in Digital Humanities. *magazén*, 5(2). DOI: 10.30687/mag/2724-3923/2024/02/001/

Marabelli, M., Chan, Y.E. (2024). The strategic value of DEI in the information systems discipline. *The Journal of Strategic Information Systems*, 33(1) https://www.sciencedirect.com/science/article/pii/S0963868724000052

McKinsey & Company (2022). What is diversity, equity, and inclusion? https://www.mckinsey.com/featured-insights/mckinsey-explainers/what-is-diversity-equity-and-inclusion

Morriello, R., & Sardo, L. (2023). Diversity management e prospettive DEI in biblioteca. *Biblioteche oggi Trend*s, 9(2), 33-42. http://doi.org/10.3302/2421-3810-202302-033-1

Nazioni Unite (2015). Obiettivi per lo Sviluppo Sostenibile, https://unric.org/it/agenda-2030/

Raju, J. (2020). Diversity, Inclusion, and Social Justice in the Information Context: Global South Perspectives, *The International Journal of Information, Diversity, & Inclusion*, 4(3/4). DOI:10.33137/ijidi.v4i3/4.34974

Thompson, K.M. (2020). Socializing Engagement: From Words to Action, *The International Journal of Information, Diversity, & Inclusion*, 4(1). DOI: 10.33137/ijidi.v4i1.33251

Carpe bias, quam minimum credula queries

Sabato Danzilli
Università di Catania/Universität Heidelberg, Italia/Germania, sabato.danzilli@phd.unict.it

ABSTRACT (ITALIANO)

L'intervento analizza i rischi culturali derivanti dall'uso acritico dei Large Language Models (LLMs) nelle scienze umane, focalizzandosi sul nesso tra formulazione delle query e amplificazione di bias. Attraverso il concetto di "costituzione della domanda euristica" di Johann Gustav Droysen – atto metodologico che trasforma la curiosità storica in indagine critica attraverso la selezione intenzionale di fonti e categorie interpretative – si dimostra come le query negli LLMs non siano mai neutre, ma incorporino premesse che orientano l'output verso narrazioni dominanti. Esempi concreti (dall'analisi di testi oraziani alla ricerca su corpora parlamentari) rivelano meccanismi di conferma bias che minacciano la diversità interpretativa. Si argomenta la necessità di una formazione umanistica "aumentata", capace di coniugare rigore ermeneutico e consapevolezza algoritmica, trasformando l'IA da amplificatore di stereotipi a strumento di esplorazione critica.
Parole chiave: bias; domanda euristica; Verstehen; epistemologia digitale; diversità interpretativa.

ABSTRACT (ENGLISH)

Carpe bias, quam minimum credula queries

The contribution examines the cultural risks arising from the uncritical use of Large Language Models (LLMs) in the humanities, focusing on the link between query formulation and bias amplification. Through Johann Gustav Droysen's concept of the "heuristic question constitution" – a methodological act that transforms historical curiosity into critical inquiry through the intentional selection of sources and interpretative categories – it demonstrates how LLM queries are never neutral but embed premises that steer outputs toward dominant narratives. Concrete case studies (from Horatian textual analysis to parliamentary corpora research) reveal confirmation bias mechanisms threatening interpretive diversity. The paper advocates for an "augmented" humanistic education, merging hermeneutic rigor with algorithmic awareness to transform AI from a stereotype amplifier into a tool for critical exploration.
Keywords: bias; heuristic question; Verstehen; digital epistemology; interpretive diversity.

1. INTRODUZIONE

Sebbene l'adozione di strumenti di LLM (Large Language Model) come ChatGPT sia ancora in una fase iniziale e non codificata nella prassi scientifica, il loro utilizzo informale sta rapidamente crescendo tra membri della comunità accademica, della popolazione studentesca e tra professionisti/e del settore pubblico e privato. Questi modelli sono impiegati per attività quali il *data mining* o la scrematura preliminare di vasti *dataset*, o anche per superare *gaps* linguistici, facilitando processi altrimenti complessi e dispendiosi in termini di tempo. L'intervento vuole offrire uno spunto di discussione sui mutamenti nell'approccio alla ricerca introdotti dall'interazione tra chi fa ricerca (in senso largo) nell'ambito culturale e i modelli di intelligenza artificiale (IA). Si vuole mostrare come un uso ingenuo di questi modelli rischi di amplificare narrazioni dominanti e stereotipate, emarginando prospettive minoritarie e ipotesi interpretative alternative. Questo fenomeno non solo compromette l'equità nell'accesso alla conoscenza, ma minaccia anche la diversità interpretativa, pilastro essenziale delle scienze umane.

2. METODOLOGIA

L'analisi presentata si basa su test condotti con ChatGPT (versione GPT-4o), utilizzando query in italiano. Il modello è stato interrogato tramite interfacce standard, senza personalizzazione dei parametri di generazione (es. temperatura, top-p), per simulare l'uso comune da parte di utenti non esperti. I dati di addestramento dei sistemi LLM utilizzati in misura maggiore nei paesi occidentali, come noto, derivano da corpora eterogenei e multilingue, con una sovrarappresentazione di fonti anglofone e occidentali che influisce sulle risposte e da cui deriva una maggiore "qualità" osservata delle risposte nella lingua di maggior addestramento, ossia l'inglese. La scelta di concentrarsi su casi di studio letterari (es. Orazio) e storici (es. verbali parlamentari) mira a evidenziare dinamiche trasversali alle diverse discipline umanistiche.

3. PER UN RECUPERO DEL CONCETTO DI COMPRENSIONE

L'IA eccelle nel generare spiegazioni causali basate su correlazioni statistiche, senza tuttavia accedere ai significati soggettivi, che restano centrali nel processo interpretativo umano. Tuttavia, essa può essere uno strumento ausiliare alle capacità umane di analizzare e interpretare dati complessi. In questa luce, il ruolo di ricercatrici e ricercatori non viene sminuito, ma enfatizzato: essi rimangono responsabili nell'attribuzione di significato.

Il concetto classico di *Verstehen* può essere utile per sottolineare la differenza tra un approccio ai dati (in senso ampio) puramente algoritmico e la comprensione umana che integra intenzionalità e contesto. Tuttavia, per evitare che il discorso diventi astratto o banalizzante, tale concetto dovrebbe essere utilizzato solo come punto di confronto metodologico: come monito per l'utente di essere consapevole dei limiti intrinseci dello strumento. Risulta esplicativo in questo senso richiamare la "costituzione della domanda euristica" come la intendeva Johann Gustav Droysen nella *Historik*. Per Droysen, formulare una domanda non è un atto neutro, ma un processo creativo e critico che delimita il campo d'indagine, seleziona le fonti e definisce i criteri interpretativi. Nell'uso dell'IA, questa consapevolezza metodologica diventa cruciale: una query mal strutturata non solo produce risposte distorte, ma cristallizza una "pre-comprensione acritica", trasformando l'AI da strumento euristico in una gabbia ermeneutica.

Il riferimento a Droysen rimarca in questa sede l'utilizzo non di un approccio mentalistico (es. "le macchine possono pensare?"), ma funzionalistico: si analizza l'atto della query come pratica ermeneutica situata, che richiede un controllo critico sulla coerenza tra premesse e risultati. Formulare una query, anche nell'uso dell'IA, significa già caricare l'indagine di intenzionalità, selezionando e strutturando i dati secondo una visione che non è mai neutra. In questo senso, la scelta delle premesse interpretative, siano esse linguistiche o tecniche, risulta determinante nella costruzione del sapere storico e filologico. L'uso dell'IA solleva dunque domande fondamentali sul rapporto tra spiegazione e comprensione nelle scienze umane. La dipendenza da modelli computazionali sposta l'attenzione dall'intenzionalità soggettiva a un approccio dominato da correlazioni. Tuttavia, se impiegata consapevolmente, l'IA può contribuire a una lettura stratificata dei risultati, integrando le intuizioni interpretative con nuovi livelli di profondità derivanti dall'analisi statistica.

L'intervento intende sottolineare di conseguenza i rischi insiti nell'utilizzo di strumenti di intelligenza artificiale da parte di operatrice e operatori privi di una formazione adeguata. L'assenza di competenze specifiche non solo espone a errori tecnici, ma amplifica le possibilità di incorrere in distorsioni interpretative. Per studiose e studiosi con una conoscenza preliminare limitata, l'IA tende a confermare le loro premesse implicite, rafforzando interpretazioni preesistenti e impedendo una riflessione critica sui dati. Questo effetto è amplificato dalla percezione errata che le risposte fornite dall'IA siano oggettive e neutrali. Le promotrici culturali e altri operatori del settore, che spesso utilizzano strumenti come i LLMs per elaborare rapidamente grandi quantità di dati, rischiano di consolidare prospettive stereotipate o parziali, riproducendo narrazioni dominanti a scapito di visioni alternative o minoritarie. La scarsa consapevolezza epistemologica può portare a decisioni e analisi meno rigorose. Gli accademici che si affidano all'IA per supportare aspetti secondari delle loro ricerche potrebbero inconsapevolmente rafforzare bias preesistenti, compromettendo la pluralità interpretativa e riproducendo schemi interpretativi riduttivi. Questi utilizzi errati trovano terreno fertile in forme di pensiero ancora maggioritarie tra i lavoratori e le lavoratrici della conoscenza. Da una parte, una presunzione di neutralità intrinseca negli strumenti digitali, associata a una scarsa conoscenza tecnica delle loro logiche operative. Si consideri inoltre che l'unicità dell'output fornito dagli LLMs in risposta alle query ammanta psicologicamente l'output stesso di un'oggettività che esso non possiede. Dall'altra, un insieme di pregiudizi radicati in molta ricerca umanistica o filosofica, e non solo negli eredi delle tradizioni "romantiche" che percepiscono la meccanizzazione come minaccia alla creatività umana, o di posizioni che liquidano l'IA come intrinsecamente incapace di cogliere il "vero" significato dei dati. Tali atteggiamenti, seppur di natura diversa, convergono nell'ostacolare una comprensione adeguata delle trasformazioni epistemologiche e metodologiche introdotte dall'IA. Il rischio di *confirmation bias* si rivela particolarmente insidioso, in quanto capace di minare il valore epistemico della produzione scientifica, con un impatto potenzialmente grave sulle scienze ermeneutiche.

L'uso dell'IA solleva dunque domande cruciali sul rapporto tra spiegazione e comprensione nelle scienze umane. La dipendenza da modelli computazionali sposta l'attenzione dall'intenzionalità soggettiva a un approccio dominato da correlazioni. Tuttavia, se impiegata consapevolmente, l'IA può contribuire a una lettura stratificata anche dei testi oggetto di indagine ermeneutica, integrando le intuizioni interpretative con l'analisi statistica. Emergono qui questioni di bias strutturali, ben documentate in letteratura (cfr. studi

di Buolamwini e Gebru sul gender bias, e di Caliskan et al. sui pregiudizi culturali negli algoritmi), che richiedono un'attenta valutazione delle fonti e del contesto di addestramento dei modelli.
L'analisi di grandi moli di dati, caricati sotto forma di allegati – ad esempio tramite testi storici in PDF – può evidenziare due ordini di problemi. Da un lato, errori palesi, come l'attribuzione errata di citazioni; dall'altro, rischi più sottili, quali l'amplificazione di pregiudizi latenti derivanti dalla formulazione stessa della query. L'IA può assecondare implicitamente la preconoscenza dello studioso, falsando così l'analisi e contribuendo al *confirmation bias*. Non occorre essere tuttavia "apocalittici": come nel caso dell'IA anche la trasmissione manoscritta o le scelte editoriali non sono mai state completamente asettiche; appare evidente: sebbene le biblioteche e i database non "producessero" direttamente risposte, la loro organizzazione e selezione dei testi già rifletteva un orientamento culturale e metodologico. A maggior ragione, tuttavia l'uso dell'IA solleva questioni analoghe di controllo e responsabilità, poiché l'algoritmo stesso codifica convenzioni e preferenze determinate da chi lo progetta o addestra.
Di seguito illustreremo attraverso alcuni esempi i rischi connessi a un uso ingenuo e poco critico delle query nei modelli di intelligenza artificiale. Sottolineiamo che il soggetto ideale di questo "esperimento" non è il tecnico esperto o già formato, ma figure come studiose, lavoratori del settore culturale o membri dell'accademia che necessitano di estrarre rapidamente informazioni da una vasta mole di dati per aspetti secondari o periferici delle loro ricerche.
Immaginiamo uno di questi operatori. Analizzando come egli formuli la sua query, possiamo osservare che l'output generato dall'IA non solo varierà significativamente, ma potrebbe risultare contraddittorio o addirittura fuorviante: un ricercatore che utilizzi un corpus storico digitalizzato di documenti parlamentari, come i verbali delle sedute del Parlamento italiano tra il 1861 e il 1946, per analizzare l'evoluzione del linguaggio politico legato alle riforme sociali. Se la query prevede termini moderni come "welfare" o "diritti civili", è possibile che il modello restituirà risultati poco rappresentativi o falsati, poiché tali termini potrebbero non essere presenti o avere significati completamente diversi nei contesti storici analizzati. Una formulazione superficiale della query potrebbe quindi omettere dibattiti centrali che utilizzavano un linguaggio diverso ma pertinente, una riformulata con linguaggio storico (es. "assistenza pubblica") consentirà esiti più robusti. Questo semplice aggiustamento, però, non basta: la selezione dei documenti e la pesatura statistica delle risposte dipendono da dinamiche opache (es. rappresentatività del corpus digitalizzato), che richiedono un'ulteriore contestualizzazione.
Ma un'insidia nell'uso ingenuo degli strumenti si percepisce a un livello ancora più basilare.
Abbiamo caricato su ChatGPT-4o il 10 gennaio 2025 una versione online facilmente reperibile delle *Opere complete* di Quinto Orazio Flacco.
Immaginiamo che una studentessa, o anche una promotrice culturale, cerchi per questa via di reperire facilmente informazioni riguardo la notissima *Ode I, 11*.
Immaginiamo che interroghi la macchina con questa domanda:
In che ode Orazio afferma che bisogna cogliere l'attimo?
Abbiamo ottenuto questa risposta:
Il famoso invito di Orazio a "cogliere l'attimo" si trova nell'Ode I, 11, rivolta a Leuconoe. In questa ode, Orazio esprime il concetto di carpe diem: "carpe diem, quam minimum credula postero", che si traduce come "cogli il giorno, confidando il meno possibile nel domani". Questo invito rappresenta il fulcro della filosofia esistenziale oraziana, che spinge a vivere il presente senza preoccuparsi eccessivamente del futuro incerto.
Si tratta di una risposta molto banale e che andrebbe approfondita, ma passabile per fini "enciclopedici".
Immaginiamo tuttavia che lo stesso operatore un po' più dubbioso, abbia invece domandato:
Orazio afferma che bisogna cogliere l'attimo?
In maniera interessante, a una domanda più neutrale, la risposta è stata più informativa:
Sì, Orazio afferma di cogliere l'attimo, ma con una sfumatura e una prospettiva specifica. Nel famoso verso "carpe diem, quam minimum credula postero" tratto dall'ode 1.11, Orazio esorta a "cogliere il giorno, confidando il meno possibile nel domani". Non si tratta di un invito all'edonismo sfrenato, ma piuttosto a vivere consapevolmente il presente, godendo di ciò che offre senza eccessi, in modo da evitare l'angoscia per il futuro incerto. Questo tema riflette la sua adesione a un'etica di moderazione e misura, ispirata anche alla filosofia epicurea e stoica, che cerca di bilanciare il piacere e la ragione.
Vediamo però cosa succedere se l'operatore pone una query già carica di un'interpretazione più forte:
Dove Orazio afferma che bisogna godere ogni giorno?
E la risposta della macchina accompagna l'interpretazione a monte dell'utente:

Orazio celebra l'importanza di godere ogni giorno nel carpe diem (afferrare il giorno) in Odi I, 11, rivolta a Leucònoe. In questo componimento, il poeta invita a non interrogarsi sul futuro o sulle sorti dettate dagli astri, ma a vivere intensamente il presente:
"Sappi cogliere allora il giorno come viene, senza illusioni: meno che puoi fidando nel domani."
Cosa insegnano questi esempi? Le domande formulate agli algoritmi IA non sono mai neutrali. Esse incorporano premesse semantiche e strutturali che influenzano profondamente le risposte. Nel contesto delle scienze umane, dove il significato è spesso polisemico, una query ambigua o implicante relazioni causali non dimostrate può condurre a risposte distorte, rafforzando narrative preconcette.
Il cuore della questione risiede nella tendenza dei modelli di intelligenza artificiale a confermare premesse o interpretazioni già presenti nelle query poste dagli utenti. Quando un utente pone una query carica di premesse implicite (ad esempio, chiedendo "Dove Orazio afferma che bisogna godere ogni giorno?"), il modello tende a generare una risposta che asseconda tali premesse, invece di metterle in discussione o evidenziare altre possibili letture. Questo effetto amplifica il rischio che l'utente confermi una narrazione preconcetta, compromettendo l'apertura ermeneutica che dovrebbe essere alla base della ricerca nelle *humanities*.
Il rischio epistemico evidente è che l'IA, confermando le premesse implicite dell'utenza, riduca la diversità interpretativa e rafforzi narrative dominanti o stereotipate. Ad esempio, nell'analisi di corpora storici, una query mal formulata potrebbe selezionare soltanto i dati che confermano una specifica ipotesi, lasciando in ombra contronarrazioni potenzialmente significative.
La progettazione delle query non dovrebbe essere vista solo come una tecnica ingegneristica, pertanto, ma come un atto ermeneutico. La precisione semantica nella formulazione non serve solo a ottenere risposte accurate, ma evita che il sistema IA diventi uno specchio dei pregiudizi dell'interlocutore. Una riflessione filosofica sulle premesse interpretative diventa centrale per affrontare i rischi epistemici. Il problema è pressante perché coinvolge il modo stesso in cui sono strutturati questi modelli, nella fase storica in cui essi diventano una tecnologia di massa.
Il motivo per cui una query più neutrale produce una risposta più informativa è che, in assenza di premesse forti o di un'interpretazione implicita nella domanda, il modello IA è "costretto" a generare una risposta che copra un ventaglio di significati più ampio, senza focalizzarsi su una lettura preconfezionata. Quando la query è carica di premesse, come nel caso di "Dove Orazio afferma che bisogna godere ogni giorno?", il modello utilizza quelle premesse come vincoli per generare una risposta. Di conseguenza, tenderà a produrre un'interpretazione che conferma l'input, privilegiando i dati che meglio si adattano a quella cornice semantica, viceversa, una query più neutrale, come "Orazio afferma che bisogna cogliere l'attimo?", lascia invece il modello libero di attingere a una gamma più ampia di contesti, permettendo una risposta più equilibrata che tiene conto di diverse sfumature interpretative.
I modelli linguistici avanzati, come ChatGPT, si basano su correlazioni statistiche apprese dai dati. Quando una query presenta un focus specifico (ad esempio, "godere ogni giorno"), il modello cerca corrispondenze strette, che limitano implicitamente la varietà delle informazioni richiamate. Le premesse cariche di significato nelle query agiscono come "guide" per il modello, ma spesso portano a confermare interpretazioni riduttive o parziali. Nel caso dell'esempio, chiedere esplicitamente "dove Orazio afferma che bisogna godere ogni giorno" trasforma il concetto di carpe diem in una lettura edonistica, portando il modello a costruire una risposta che rafforza questa interpretazione. Una query neutrale, invece, consente al modello di esplorare una maggiore varietà di significati, riflettendo, ad esempio, anche l'etica della moderazione epicurea o la prospettiva stoica, come nel secondo esempio fornito.
Le implicazioni per la pratica quotidiana nell'ambito delle *humanities* sono evidenti e determinano la necessità di educare l'utenza di questi strumenti a come mitigare i rischi di vedersi IA che si limitano ad accondiscendere ai loro pregiudizi pensando di fare *data mining*.
Una query apparentemente innocente, come quelle che abbiamo visto su Orazio, ha in realtà un numero molteplice di fallacie. Ne richiamiamo velocemente alcune:
Bias di conferma: Domande come "Dove Orazio afferma che bisogna godere ogni giorno?" guidano il modello a confermare un'interpretazione edonistica implicita, ignorando altre prospettive come l'etica della moderazione epicurea o stoica.
Bias di selezione: Le premesse nelle query spingono il modello a privilegiare dati che confermano interpretazioni dominanti nei testi di addestramento, escludendo contronarrazioni o letture filologicamente più accurate.

Bias culturale e linguistico: L'associazione di carpe diem con il godimento immediato è spesso rafforzata da interpretazioni moderne e da pregiudizi culturali nei dati di addestramento, ignorando la complessità del contesto storico e culturale.
Bias di disponibilità: Il modello attinge alle interpretazioni più comuni presenti nei dati di addestramento, trascurando quelle meno frequenti ma più rilevanti per una lettura critica del testo.
Bias algoritmico: Gli algoritmi ottimizzano risposte basate su correlazioni statistiche, non considerando il contesto critico della query. Questo limita la capacità del modello di generare risposte che sfidino le premesse implicite.
L'amplificazione dei bias attraverso query cariche di premesse ha anche un impatto diretto sull'eliminazione o sottorappresentazione delle minoranze e delle ipotesi interpretative alternative.
L'esperimento con le *Odi* di Orazio illustra il contrasto tra una domanda carica di premesse ("godere ogni giorno") e una aperta ("Orazio afferma di cogliere l'attimo?"). Come nella "costituzione della domanda euristica" di Droysen, la differenza non sta nella "verità" dell'IA, ma nell'interazione tra struttura della query e contesto culturale. Una query del tipo "Quali diritti civili sono stati discussi nel Parlamento italiano nel XIX secolo?" potrebbe portare il modello a enfatizzare le discussioni più comuni (es. riforme elettorali) e ignorare temi meno rappresentati, come il dibattito sull'emancipazione femminile o i diritti delle minoranze religiose. Una query che riflette una visione maggioritaria (es. "Come il Rinascimento ha promosso la libertà intellettuale?") amplifica la narrazione dominante, ignorando eventuali critiche, come quelle legate all'esclusione delle donne, e al filone femminista, o delle minoranze religiose ed etniche durante il periodo.
La scena è comune, quotidiana: uno studente universitario, alle prese con la ricerca di informazioni rapide per un esame, o una professionista che deve esplorare un argomento secondario fuori dal proprio ambito di *expertise*, si rivolge a un modello di intelligenza artificiale, usando magari lo strumento di caricamento di file per "sentirsi più sicuro". La query posta è diretta, apparentemente innocente, formulata per ottenere risposte sintetiche e immediate. Tuttavia, proprio in questo gesto, che può sembrare una prassi per ottimizzare il tempo, talvolta necessaria, si nasconde un rischio insidioso e profondo.
Dalla stesura della prima versione di questo contributo (gennaio 2025), l'ecosistema degli LLM ha subito un'accelerazione: modelli come DeepSeek e strumenti integrati in app di messaggistica, lettori PDF (es. Adobe AI Assistant) o ulteriore diffusione di piattaforme accademiche (es. Scite, Elicit) hanno reso l'IA ubiqua. Non si tratta più solo di sottoporre un prompt a un modello, ma di processi in cui l'IA "sgrossa" o addirittura ricerca dati preliminari, sintetizza letteratura o addirittura suggerisce ipotesi di lettura. Questa automazione parziale crea un salto qualitativo: l'utente medio interagisce e sottopone ad analisi output che sono essi stessi già mediati da algoritmi, spesso inoltre senza accesso al ragionamento sottostante. Modelli di deepthinking (es. Chain-of-Thought prompting) che esplicitano il processo logico della macchina offrono un contraltare promettente, ma rimangono pienamente comprensibili solo a utenti "avvertiti" dei rischi.
Gli esempi discussi (dall'analisi di Orazio alle query sui verbali parlamentari) mostrano che il rischio principale non è la "malafede" dell'IA, ma l'assenza di consapevolezza metodologica nell'utente. Come sottolinea Droysen, la storia (e per estensione le scienze umane) non è mera accumulazione di dati, ma costruzione guidata da domande critiche. Gli strumenti di IA, se usati passivamente, rischiano di ridurre la ricerca a un enciclopedismo frammentario. In Italia, dove l'educazione digitale fatica a illustrare le criticità ermeneutiche, ciò richiede interventi sistemici: non basta insegnare a "verificare le fonti", come suggeriscono gli stessi modelli di IA in fondo alle loro interfacce, ma serve formare a decostruire le premesse implicite delle query, riconoscere i bias cognitivi e valutare l'adeguatezza contestuale delle risposte. Questo approccio richiama in parte l'avvertimento di Droysen contro il positivismo: la ricerca non può ridursi a leggi universali o correlazioni statistiche, ma deve integrare categorie interpretative (politiche, culturali, morali). A tal proposito, strumenti come i framework di "AI literacy" proposti da UNESCO o il progetto europeo AI4T offrono spunti operativi, ma vanno adattati alle specificità delle discipline umanistiche.
L'IA non è un osservatore neutrale; le risposte che produce non sono mere fotografie oggettive della realtà. Ogni query agisce come una lente che amplifica le prospettive e le narrazioni presenti nei suoi dati di addestramento, oltre a indurre un falso senso di sicurezza nello scienziato sociale, che si sente spesso gratificato da informazioni che vanno nella direzione che lui auspicava, ma solo perché il prompt è stato scritto in un modo poco avvertito. Il risultato è un ciclo che rafforza schemi già consolidati, relegando al margine le voci minoritarie, le interpretazioni alternative e le ipotesi innovative. Invece di arricchire il panorama epistemico, rischia di impoverirlo. Infine, resta aperto il problema della possibile "responsabilizzazione" degli algoritmi, ovvero di come la futura implementazione tecnica possa mitigare

bias e distorsioni. Sul punto, possono risultare utili riflessioni che prospettano una maggiore trasparenza e verificabilità nell'intero processo di sviluppo dei modelli, permettendo così di arginare parzialmente i rischi epistemici connessi al loro impiego nelle scienze umane.
Questo, va da sé, non è solo un limite tecnico, ma un rischio etico e scientifico. Se lasciato incontrollato, l'uso ingenuo dell'IA potrebbe minare la diversità interpretativa, uno dei pilastri fondamentali delle scienze umane. Diffondere prospettive stereotipate non è solo un errore di metodo; è un insuccesso del compito della ricerca accademica. La sfida, quindi, non è se utilizzare l'IA, ma come promuovere un utilizzo consapevole, evitando che un atto quotidiano si trasformi in uno strumento di consolidamento di stereotipi e di riduzione della pluralità interpretativa.

4. CONCLUSIONI

Il problema, in conclusione, non è nella "verità" dell'IA, ma nell'interazione tra struttura della query, bias del modello e contesto culturale dell'utente. Non si tratta pertanto di essere "apocalittici o integrati", ma riconoscere che l'IA necessita un cambiamento cognitivo nella definizione delle pratiche di ricerca. La posta in gioco non è etica in astratto, ma metodologica: come evitare che strumenti progettati per ottimizzare efficienza riducano la complessità interpretativa. La soluzione non sta, o non sta solo, in controlli tecnici ex post (es. filtri anti-bias), ma nella formazione di una professionalità umanistica che coniughi filosofia del linguaggio e filosofia della scienza, teoria delle fonti e critica degli algoritmi. Solo così si potrà scongiurare un'IA quale amplificatore di bias e rendere, al contrario, tale strumento utile per una ricerca più rigorosa e inclusiva.

BIBLIOGRAFIA

Bender, E. M. et al. (2021) On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency, DOI:10.1145/3442188.3445922

Buolamwini, Joy (2024) Unmasking AI: My Mission to Protect What Is Human in a World of Machines.

Caliskan, A. et al. (2017) Semantics derived automatically from language corpora contain human-like biases, Science, Vol 356, Issue 6334, 183-186, DOI: 10.1126/science.aal4230;

Crawford, K. (2021) The Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence. New Haven: Yale University Press;

Droysen, Johann Gustav, Sommario di Istorica. Pisa: Edizioni della Normale, 2014; New York: Random House USA;

EU AI4T Project (2024) AI Pedagogy in Humanities: A Framework for Critical Engagement. Brussels: European Commission;

Gebru, T. et al. (2021) Datasheets for Datasets, v8 https://doi.org/10.48550/arXiv.1803.09010;

Mercier, H., Sperber, D. (2017) The Enigma of Reason. Cambridge (Massachusetts): Harvard University Press, 2017;

O'Neil, C. (2017) Armi di distruzione matematica. Milano: Bompiani, 2017;

Orazio. (2013) Tutte le opere. Roma: Salerno, 2013;

Ricoeur, P. (2000) La mémoire, l'histoire, l'oubli. Paris: Seuil;

Ruha, B. (2019) Race After Technology: Abolitionist Tools for the New Jim Code. Medford: Polity, 2019

Umoya Noble, S. (2018) Algorithms of Oppression. How Search Engines Reinforce Racism. New York: NYU, 2018;

UNESCO (2023) Guidance for Generative AI in Education and Research. Paris: UNESCO.

Evaluating bias within an epistemological framework for AI-based research in the humanities

Sarah Oberbichler*❖, Cindarella Petz*☼

DH Lab, Leibniz Institute of European History,
Alte Universitätstr. 19, 55116 Mainz, Germany
❖ oberbichler@ieg-mainz.de
☼petz@ieg-mainz.de
*These authors contributed equally to this work.

ABSTRACT (ENGLISH)

In this paper, we propose an epistemological framework for AI-based digital research in the humanities. Based on two of our current research projects using Large Language Models (LLMs) for various analysis tasks, we demonstrate the critical assessment of bias examination as one vital aspect of our proposed epistemological framework. There, we are focusing on biases inherent to the LLMs, and biases, that emerge due to the application of the LLM, which significantly impact research outcomes. This includes how language-specific prompts may favor certain languages, resulting in differing outcomes, as well as how LLMs can amplify biases present in source materials, such as reproducing problematic labels or using discriminatory language. Source critical approaches and bias output evaluation helps identify these patterns and allows researchers to make informed decisions when using LLMs for analysis tasks such as information extraction.

Keywords: LLMs; Output Evaluation; Bias Examination; FAIR AI; Ethical AI

TITLE AND ABSTRACT (ITALIANO)

La valutazione dei bias in un framework epistemologico per la ricerca basata sull'IA nelle scienze umane

In questo lavoro proponiamo un framework epistemologico per la ricerca digitale basata sull'intelligenza artificiale nelle scienze umane. Basandoci su due dei nostri attuali progetti di ricerca che utilizzano Large Language Models (LLM) per vari compiti di analisi, dimostriamo che la valutazione critica dei *bias* è un aspetto vitale del nostro framework epistemologico proposto. In questo caso, ci concentriamo sui *bias* intrinseci agli LLM e su quelli che emergono dalla loro applicazione, che hanno un impatto significativo sui risultati della ricerca. Ciò include il modo in cui i prompt testuali possono favorire alcune lingue, dando luogo a risultati diversi, nonché il modo in cui gli LLM possono amplificare i pregiudizi presenti nei materiali di partenza, come la riproduzione di denominazioni problematiche o l'uso di un linguaggio discriminatorio. Gli approcci critici alle fonti e la valutazione dei *bias* nei risultati aiutano a identificare questi schemi e consentono ai ricercatori di prendere decisioni informate quando utilizzano gli LLM per compiti di analisi come l'estrazione di informazioni.

Parole chiave: LLM; Valutazione degli Output; Analisi dei *bias*; IA FAIR; IA Etica

1. An epistemological framework for AI-based research in the humanities

Digital research is changing rapidly through lowered barriers to digital and computational tools — in particular based on Large Language Models (LLMs) — and intensified efforts to link FAIR cultural heritage data with them.[1] This presents us with both opportunities to democratize knowledge through broader engagement with cultural heritage, but also increases the challenge on how to provide guidance for the responsible use of those resources and the appropriate application of these new tools.

1 An example for such an effort is the European ECHOES project, which is designed to establish the European Collaborative Cloud for Cultural Heritage (ECCCH), a shared platform to link heritage professional and researchers to open data and scientific resources from GLAM institutions fostering collaboration across previously fragmented communities. There, advanced digital tools and specifically AI technology plays a crucial within the supported projects, such as AUTOMATA and TEXTaiLES (https://www.echoes-eccch.eu/projects/).

To address this, we argue that established digital epistemological frameworks within the Digital Humanities disciplines need to be extended and adapted to include the deployment of AI models: to combine traditional practices such as verifying information with new approaches to critically assess complex AI systems, their prompts, and outputs.

We therefore introduce an epistemological framework for AI-based research in the humanities, providing guidance for the critical reflection and systematic examination of how knowledge is created, validated, and transmitted through digital tools and data — specifically when working with generative AI. This includes critical evaluation of sources, methodologies, and tools used in knowledge production, as well as understanding how these elements shape our interpretation and understanding of cultural heritage material.

Drawing from the authors' research projects on "Transnational Flows of News in Historical Newspapers" and on the "Construction of Political Criminality during the Dollfuß-/Schuschnigg-Regime", this paper demonstrates how the application of such a framework can enhance methodological rigor and critical reflection in the AI-assisted analysis of cultural heritage resources based on the example of understanding and evaluating biases within sources and algorithms.

This framework goes beyond interpretability and explainability approaches – such as technical tools like Neuronpedia[2], which visualizes internal model activations, or surface-level transparency like DeepSeek's[3] natural language explanations of its reasoning steps — and black-box vs. white-box models. Both black-box and white-box AI models face fundamental challenges with explicability and explainability. While black-box models inherently struggle with explainability due to their opaque nature, white-box models, despite being interpretable in theory, can become equally challenging to explain when their complexity increases. Both struggle with explicability, the broad challenge of enabling accountable usage and appropriate understanding across different stakeholders. The complexity of AI systems means that even when we can trace how a decision is made (as in white-box models), we may still fail to make this understanding accessible and meaningful to the stakeholders who need to work with these systems (Herzog, 2022).[4]
In order to mitigate this and to understand and use AI responsibly, this requires examining:

- The quality and reliability of knowledge sources
- The appropriateness and trustworthiness of the methods
- The validation, evaluation, and ethical implications of how we use this knowledge

The proposed framework as visualized in Figure 1 can provide such a comprehensive approach to assessing AI models from an epistemological standpoint, connecting source, data, and tool criticism with prompt and output evaluation (compare to Oberbichler & Petz 2025). This framework addresses three interconnected dimensions:

The first dimension focuses on source and data criticism, where traditional research practices must evolve to meet new challenges. Epistemologically, the reliability of knowledge is directly tied to the quality of its sources. While it has a long tradition in humanities research to evaluate representativeness, appropriateness, quality, authenticity, and bias of used sources, these need to be extended also onto LLMs, which are both data and tool / method. Additionally, researchers must carefully consider the legal dimensions of their data: copyright protections, licensing restrictions, and the handling of personal data under EU AI Act and or national laws requirements which need to influence the choice of which models and tools to work with, and how the corpora needs to be created and analyzed. Compliance with these regulations ensures that the knowledge generation process adheres to legal and ethical standards and enhances trustworthiness.

2 Neuronpedia, https://docs.neuronpedia.org/.

3 DeepSeek, https://chat.deepseek.com/ .

4 Similarly, the European Data Protection Supervisor (2023) stressed the need for transparent documentation in order for a model to be understood beyond mere technical transparency.

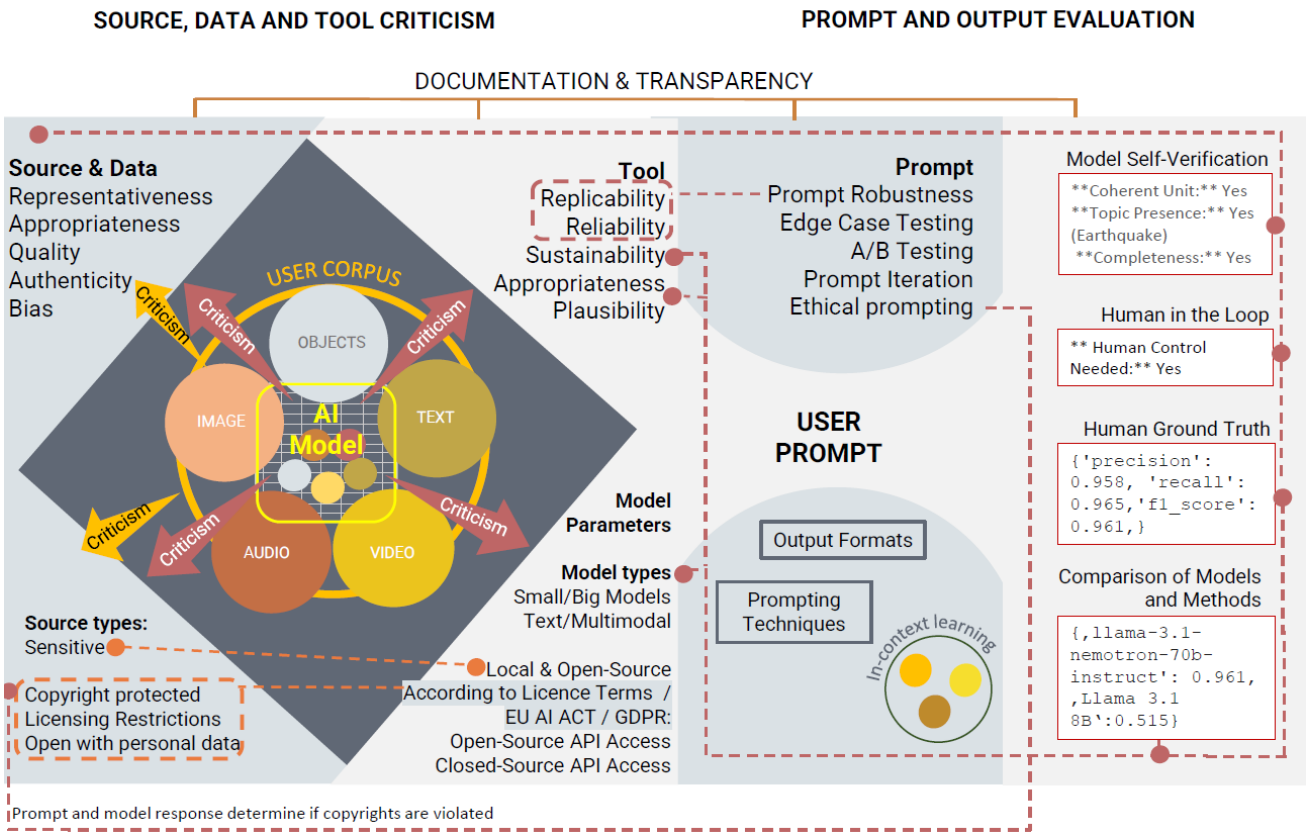


Figure 1. Epistemological framework for AI-based research in the humanities (Image: S. Oberbichler & C. Petz 2024).

The second dimension centers on tool and model criticism, emphasizing the replicability, reliability, sustainability, appropriateness, and plausibility of chosen LLM tools. These factors are crucial for justifying the methods used in knowledge generation. The tools and methods must be reliable and appropriate to produce trustworthy knowledge. This includes careful consideration of model parameters and types, whether working with small or large models, text-only or multi-modal systems. For researcher or any person using LLMs as an analysis tool, it is suggested to evaluate different access options, weighing the implications of open-source / open-weight versus closed-source model options, and assessing the advantages and limitations of local versus cloud-based deployment options. The recent FAIRification of LLMs — i.e., the extension and re-interpretation of the FAIR data principles to include AI models, algorithms, software and tools which produce data — answered to the need for thorough technical documentation of model characteristics[5] and persistent identification (compare e.g., Ravi et al., 2022; Huerta et al, 2023; Hupont et al., 2023; Alkemade et al., 2023; Golpayegani et al., 2024). This transparency furthers the scientific re-usability of trained models and AI tools, and increases both the reliability and interpretability of models, allowing for cross-comparisons and the reproducibility of results. If copyright or licensing restrictions are present in the data base at question, using local models should be considered the most responsible option. While small local models have the advantage of not needing additional computing power, comparative evaluations can determine whether larger LLMs achieve significantly better results that justify their usage.

The final dimension involves prompt and output evaluation, requiring the implementation of robust prompting techniques and evaluation guidelines. This includes systematic edge case testing, A/B testing, prompt iteration, ethical prompting, and the integration of model self-verification protocols. These elements ensure that the interaction with LLMs is both consistent and ethical. Critical to this dimension is

5 This includes model training data, input and output data types, configurations, dependencies, benchmark metrics, and uncertainty quantification for performance evaluation, as well as usage (and storage) guidelines in model cards.

the establishment of the human-in-the-loop validation processes and systematic comparison of different models and methods to ensure optimal results.
Throughout all these dimensions, maintaining comprehensive documentation and transparency of decisions remains crucial. An example template of such a documentation is shown in the *AI model Research Documentation Sheet* that is part of this framework (Oberbichler, 2024).

2. Practical application of the epistemological framework on the example bias examination

Based on two of our current research projects using AI models for various analysis tasks, we will demonstrate the critical assessment of bias examination as one vital aspect of our proposed epistemological framework. Biases manifest in AI models on various levels: in the model design, in the training data and it's selection, which then may translate into biased model outputs, and in its (mis-)application on context the model wasn't trained for (compare e.g., Ferrer et al., 2021). Source and data selection biases can result in misrepresentation of the context at study due to unbalanced domain distributions, cultural, temporal and language constraints, or demographic skews. This can lead to demonstrable social biases in model outputs, including prejudices and even discrimination based on gender, age, ethnicity, religion, disability status, and socioeconomic class (compare e.g., Dai et al., 2024; Gallegos et al., 2024; Navigli et al., 2023; European Union Agency for Fundamental Rights, 2022). Source critical approaches and bias output evaluation helps identify these patterns and allows researchers to make informed decisions when using LLMs for analysis tasks such as information extraction.

Amplification of bias and Ethical AI in historical court records

Based on the research project on "The Construction of Political Criminality during the Dollfuß-/ Schuschnigg-Regime (1933—1938)," we reflect on the amplification of biases through LLMs, both due to the models used and due to inherent biases within the source basis at study. There, we are interested in the appropriateness of the AI-assisted analysis:

1. Do different models have different capabilities for text extraction of key characteristics of the trial based on what they have been trained on? How does prompt engineering facilitates this?
2. How can we combat the amplification of biases inherent to the source basis at study? And does Ethical AI[6] obscure the analysis of discriminatory language present in historical sources?

In this study, we work with historical court records, which offer an ultimately non-objective worldview imposed by the court personnel, the judiciary, and the police onto the defendants and the case at trial. Schwerhoff (2011, p. 40) characterized the nature of court records as 'involuntary ego-documents' of the accused, where the self-declarations of the accused become altered through the court's perspectives and their recording decisions, obscuring factual representation. In our case study on the semantic construction of political criminality in the prosecution of political-religious offenses during the autocratic Dollfuß-/Schuschnigg-regime in Austria (1933—1938), this becomes evident, when specific labels are imposed on the accused within the trials, influencing their case outcomes. In the context of political-religious offenses, these are often moralizing, evoke ethnic differences, and/or dependent on the political orientation of the accused (e.g., framing someone as morally corrupt, Socialist, Jewish, or in health-related terms), revealing hidden patterns of the justice systems' labeling processes of morals, deviant behaviors, as well as political judiciary. When found guilty of charges, these labels and the punishment of conviction too have repercussions for the social standing of the accused, leading to stigmatization and ostracization in Austrian society.
The identification of these inherent perspectives of court records tie into the first dimension of our proposed framework. Another aspect of this are legal considerations: court records as sources of sensitive data require the adherence to their legal protections not only when publishing the study's results and accompanying datasets with person-related details, but already when analyzing these records with generative AI. This makes it necessary to opt for localized data and prompt processing with our chosen

6 Ethical AI in the context of non-maleficent and just AI is trained to avoid unfair biases (compare to overview on Ethical AI frameworks in Prem, 2023, p. 702).

model (or on German Data Privacy Regulations-bound servers when using high performance computing services).
The second and third dimension of our proposed framework are closely connected when assessing model characteristics, optimizing prompts, and evaluating outputs: what data have the chosen models been trained on, and how does this influence the results? Which prompts result in appropriate results? This makes it necessary to opt for open-access models in order to be able assess the appropriateness and plausibility of chosen LLMs when systematically comparing results from multiple models and different prompting strategies with a human-created ground truth.
While the source and data biases described above necessarily reappear in model output when utilizing them as Retrieval Augmented Generators, we do need to make sure not to amplify discriminatory historical terms. This study demonstrates the need for context-awareness of the human-in-the-loop for output evaluation in order to frame and possibly even offer a counter-narrative for the automatic findings of the model used, and suggest further editing to combat problematic or discriminatory language when presenting findings and when publishing the underlying dataset.

Bias evaluation when analyzing multilingual historical newspaper corpora

Using the research project "Transnational Flows of News: Analysis and Visualization of Historical News across Languages and Countries, 1850–1950" as an example, we demonstrate how systematic evaluation of AI model outputs contributes to understanding the model's appropriateness. This is crucial when working with multilingual historical sources where cultural and linguistic contexts vary greatly. We specifically want to determine whether model's training data sufficiently represents historical and linguistic diversity for accurate analysis, and whether its design effectively accommodates human-in-the-loop processes: and

1. Do different models have varying capabilities across languages? Does the language of the prompt matter?
2. How good is the historical context understanding? Are the models able to find articles related to a specific event without providing further historical context?
3. Are LLMs biased towards specific output formats?
4. How well are the models aligned for historical analysis?

The underlining project for this evaluation examines transnational news flows in Germany, France, Great Britain, Italy and Switzerland from 1850-1950, focusing on news about (re)migration and environmental / natural disasters. The goal of this study is to provide new insights into the creation and manipulation of narratives across languages and time.

Crucial for this project is the creation of a topic-specific corpus containing well separated articles. Previous layout based article separation methods currently don’t provide good enough (>70% accuracy) results (Sun et al., 2024) or are not flexible enough to be adapted for this research project. Therefore, we created a novel approach using OCR'ed text and context-understanding for article extraction via LLMs. For this purpose, we created an evaluation framework for three LLMs tasks:

- Classification: Testing models' ability to classify articles containing OCR mistakes as relevant or not relevant to the specific topic of the 1908 Messina earthquake.
- Extraction: Evaluating accuracy in extracting complete relevant articles from newspaper in different languages
- Boundary Detection: Assessing ability to correctly mark beginning and end of articles using different output formats, especially when several articles were published in the same newspaper issue.

The proposed bias evaluation within this case study occurs through multiple steps within the epistemological framework. Analyzing both language and historical domain bias involves addressing representation issues within model training data, which falls under data and source criticism (first and second dimension of our proposed framework. Language evaluation also constitutes part of tool / model criticism, as selecting a monolingual model would be inappropriate for multilingual datasets. Although

critical assessment of model training sources is often limited, both language and domain biases can be quantified by comparing model outputs with human ground truth (third dimension). Such comparisons can reveal whether a model performs significantly worse when extracting information from non-English texts compared to English texts. Historical domain bias can be evaluated by including both topic-relevant and non-relevant texts in a test corpus. Comparing the model's classification results against human-created ground truth will demonstrate whether the model possesses sufficient contextual understanding to complete this task. Additionally, closely examining a model's reasoning and explanations for classifying sources as relevant or non-relevant can provide valuable insights into its historical knowledge and its ability to integrate geographical, cultural, and historical information (Mauermann and Oberbichler,2025). Model design biases, such as distortions introduced through a model's alignment or instruction-following approach, are part of tool/model criticism and fall under the aspect of model appropriateness. Model alignment and instruction-following issues may manifest as failures to follow specific instructions—such as the directive not to truncate when extracting long articles, to preserve OCR mistakes, or to adhere to a specified output format. These issues can be quantified by comparing model outputs to provided ground truth. Another critical consideration is a model's confidence alignment. A well calibrated model should be able to highlight cases of ambiguity or uncertainty to facilitate human verification. Models that exhibit overconfidence in their answers fail to create space for human-in-the-loop approaches. This type of bias can be evaluated by comparing models' self-estimated confidence levels with output evaluation results, evaluating instructions that encourage human interaction, or analyzing the logits of a model (Mauermann and Oberbichler, 2025).

In both case studies, these approaches to identify biases demonstrate the interconnectedness of our proposed framework's dimensions, when critically assessing our sources, models, operationalisations, and outputs. This multilayered approach equips researchers with the necessary critical framework for understanding both the strengths and limitations of a model.

3. Outlook: Bias and Ethical AI

The intersection of AI technology and humanities research presents both opportunities and challenges for knowledge democratization. Our proposed epistemological framework offers a systematic approach to understanding and addressing these challenges, particularly regarding the complex relationship between ethical AI principles and historical research integrity.
Through our case studies - examining political criminality in Austrian court records and analyzing multilingual newspaper coverage - we demonstrate how ethical AI considerations must be carefully balanced with historical authenticity. While ethical AI aims to minimize discriminatory language and biased representations, humanities researchers must sometimes preserve these elements to accurately document historical discrimination and prejudice. While this preservation is integral for historical inquiry, when presenting and publishing these results and underlying datasets, every effort needs to be made to mitigate discriminatory language and labels. Our proposed framework helps researchers navigate this tension by:

1. Recognize and document biases inherent in both historical sources and AI models
2. Develop systematic evaluation processes that account for linguistic and cultural diversity
3. Maintain human oversight while leveraging AI capabilities
4. Balance ethical AI principles with accurate historical representation

This approach to understanding AI tools helps counter biases and distortions, which is central to democratizing access to knowledge and cultural heritage. By implementing robust evaluation frameworks, we can better harness AI's potential while maintaining the rigorous standards of humanities research.
The future of AI technology in the humanities will require continuous refinement of these approaches. Our work provides a foundation for this ongoing development, while acknowledging that ethical AI evaluation is not a fixed destination but rather a continuous process of improvement and adaptation.

ACKNOWLEDGEMENTS
No AI-based tools were used for creating the content of this text.

REFERENCES

Alkemade, H., Claeyssens, S., Colavizza, G., Freire, N., Irollo, A., Lehmann, J., Neudecker, C., Osti, G., & van Strien, D. (2023). Datasheets for Digital Cultural Heritage Datasets. https://doi.org/10.5281/zenodo.8375034.

European Data Protection Supervisor. (2023). TechDispatch: explainable artificial intelligence. #2/2023. Publications Office. https://data.europa.eu/doi/10.2804/802043.

European Union Agency for Fundamental Rights. (2022) Bias in Algorithms – Artificial Intelligence and Discrimination. Vienna, 2022. https://fra.europa.eu/de/publication/2022/bias-algorithm#publication-tab-1.

Ferrer, X., van Nuenen, T., Such, J. M., Coté, M., & Criado, N. (2021). Bias and discrimination in AI: A cross-disciplinary perspective. IEEE Technology and Society Magazine, 40(2), 72-80. https://doi.org/10.1109/MTS.2021.3056293.

Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Dernoncourt, F., Yu, T., Zhang, R., & Ahmed, N. K. (2024). Bias and Fairness in Large Language Models: A Survey (No. arXiv:2309.00770). arXiv. https://doi.org/10.48550/arXiv.2309.00770.

Golpayegani, D., Hupont, I., Panigutti, C., Pandit, H. J., Schade, S., O'Sullivan, D., & Lewis, D. (2024). AI Cards: Towards an Applied Framework for Machine-Readable AI and Risk Documentation Inspired by the EU AI Act. https://doi.org/10.48550/ARXIV.2406.18211.

Herzog, C. (2022). On the risk of confusing interpretability with explicability. AI and Ethics, 2(1), 219–225. https://doi.org/10.1007/s43681-021-00121-9.

Huerta, E. A., Blaiszik, B., Brinson, L. C., Bouchard, K. E., Diaz, D., Doglioni, C., Duarte, J. M., Emani, M., Foster, I., Fox, G., Harris, P., Heinrich, L., Jha, S., Katz, D. S., Kindratenko, V., Kirkpatrick, C. R., Lassila-Perini, K., Madduri, R. K., Neubauer, M. S., Psomopoulos, F. E., Roy, A., Rübel, O., Zhao, Z. & Zhu, R. (2023). FAIR for AI: An interdisciplinary and international community building perspective. Scientific Data, 10(1), 487. https://doi.org/10.1038/s41597-023-02298-6.

Hupont, I., Fernández-Llorca, D., Baldassarri, S., & Gómez, E. (2023). Use case cards: a use case reporting framework inspired by the European AI Act. https://doi.org/10.48550/ARXIV.2306.13701
LIAS: Layout Information-Based Article Separation in Historical Newspapers | SpringerLink. (n.d.). Retrieved January 2, 2025, from https://link.springer.com/chapter/10.1007/978-3-031-72437-4_15.

Mauermann, J., & Oberbichler, S (2025). LLM Biases: Expected and Unexpected Model Design Effects in Historical Newspaper Article Extraction on the Messina Earthquake. DH Lab. https://doi.org/10.58079/137qr.

Navigli, R., Conia, S., & Ross, B. (2023). Biases in Large Language Models: Origins, Inventory, and Discussion. J. Data and Information Quality, 15(2), 10:1–10:21. https://doi.org/10.1145/3597307

Oberbichler, S. (2024). AI Model Research Documentation Sheet (AIRDocS). https://doi.org/10.5281/zenodo.14550113.

Oberbichler, S., & Petz, C. (2025). Working Paper: Implementing Generative AI in the Historical Studies (1.0). Zenodo. https://doi.org/10.5281/zenodo.14924737.

Prem, E. From ethical AI frameworks to tools: a review of approaches. AI Ethics 3, 699–716 (2023). https://doi.org/10.1007/s43681-023-00258-9.

Ravi, N., Chaturvedi, P., Huerta, E. A., Liu, Z., Chard, R., Scourtas, A., Schmidt, K. J., Chard, K., Blaiszik, B., & Foster, I. (2022). FAIR principles for AI models with a practical application for accelerated high energy diffraction microscopy. Scientific Data, 9(1), 657. https://doi.org/10.1038/s41597-022-01712-9.

Schwerhoff, Gerd. 2011. Historische Kriminalitätsforschung. Frankfurt/M.: Campus Verlag.

Sun, W., Tran, H. T. H., González-Gallardo, C.-E., Coustaty, M., & Doucet, A. (2024). LIAS: Layout Information-Based Article Separation in Historical Newspapers. In A. Antonacopoulos, A. Hinze, B. Piwowarski, M. Coustaty, G. M. Di Nunzio, F. Gelati, & N. Vanderschantz (Eds.), Linking Theory and Practice of Digital Libraries (pp. 256–272). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-72437-4_15.

Dai, S., Xu, C., Xu, S., Pang, L., Dong, Z., & Xu, J. 2024. Bias and Unfairness in Information Retrieval Systems: New Challenges in the LLM Era. In: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24). Association for Computing Machinery, New York, NY, USA, 6437–6447. https://doi.org/10.1145/3637528.3671458.

Semplificare la lettura dei manoscritti utilizzando tecnologie WEB interattive e interazioni "hover"

Giacomo Marchioro[1], Andrea Brugnoli[2], Francesca Carnazzi[3], Paolo Pellegrini[4], Edoardo Ferrarini[5]
[1] Università degli Studi di Verona, Italia – giacomo.marchioro@univr.it
[2] Università degli Studi di Verona, Italia – andrea.brugnoli@univr.it
[3] Università degli Studi di Firenze - Università degli Studi di Verona, Italia – francesca.carnazzi@univr.it
[4] Università degli Studi di Verona, Italia – paolo.pellegrini@univr.it
[5] Università degli Studi di Verona, Italia – edoardo.ferrarini@univr.it

ABSTRACT

Il patrimonio manoscritto rappresenta una parte significativa dell'eredità culturale. Tuttavia, rispetto, ad esempio, alle opere di arti figurative, esso può essere meno accessibile al pubblico. Leggere e interpretare il testo di un manoscritto è spesso impegnativo, anche per i professionisti del settore. Di conseguenza, i visitatori delle mostre che presentano documenti manoscritti di solito si limitano a un'osservazione superficiale o alla lettura di una traduzione o di un riassunto forniti in una didascalia separata, riducendo la loro interazione con il documento. Anche gli studenti di discipline letterarie o storiche possono incontrare difficoltà quando si confrontano per la prima volta con la documentazione originale. È quindi essenziale fornire strumenti che consentano un'esperienza di lettura più semplice e interattiva di tali documenti. In questo contributo, presentiamo un potenziale approccio per facilitare la lettura di un documento manoscritto digitalizzato tramite un'applicazione che consente un'interattività migliorata tra il lettore e il testo. L'applicazione consente agli utenti di visualizzare, in tempo reale, tramite il movimento di un cursore, una trascrizione diplomatica del testo, espandendo le abbreviazioni ed evidenziando la traduzione della porzione di testo selezionata in una finestra separata. Inoltre, è possibile associare un audio alle singole parole per ascoltarne la pronuncia corretta.
Parole chiave: Patrimonio manoscritto; Filologia digitale; Lettura interattiva; Accessibilità; Inclusione

ABSTRACT

Simplifying the reading of manuscripts using interactive WEB technologies and "hover" interactions.
The manuscript heritage represents a significant part of the cultural heritage. Still, at the same time, compared to works of visual arts, it can be the least accessible to the public. Reading and interpreting the text of a manuscript is often challenging, even for professionals in the field. As a result, visitors to exhibitions featuring manuscript documents usually limit themselves to superficial observation or reading a translation or summary provided in a separate caption, reducing their interaction with the document. Even students in literary or historical disciplines may face difficulties when first engaging with original documentation. It is, therefore, essential to provide tools that enable an easier and more interactive reading experience of such documents. In this contribution, we present a potential approach to facilitating the reading of a digitized manuscript document through an application that enables an enhanced interactivity between the reader and the text. The application allows users to view, in real-time, through the movement of a cursor, a diplomatic transcription of the text, expand abbreviations, and highlight the translation of the selected text portion in a separate window. Additionally, external audio can be associated with individual words to hear the correct pronunciation.
Keywords: Manuscript heritage; Digital philology; Interactive reading; Accessibility; Inclusion

1. INTRODUZIONE

Ampliare l'accessibilità del patrimonio culturale è uno degli obiettivi della Convenzione di Faro[1], nonché del MiC, anche in relazione al patrimonio manoscritto, librario e documentario[2]. Per quanto concerne i manoscritti, si pongono però delle sfide complesse. In particolare, la fruizione del bene non può limitarsi a

[1] https://www.coe.int/en/web/conventions/full-list/-/conventions/treaty/199; ratifica per l'Italia con L. 133/2020.
[2] Ministero della Cultura, Piano della performance 2021-2023. https://performance.gov.it/performance/piani-performance/allegato/3778: 4.1.1.

considerare il documento come oggetto, ma deve necessariamente comprendere la lettura e la comprensione del suo contenuto, campo in cui anche gli specialisti possono incontrare difficoltà. Da tempo si sottolinea come strumenti informatici possano costituire un'utile intermediazione nell'approccio ai manoscritti (Basili et al., 2017), favorendo l'accessibilità del patrimonio a tutti gli stakeholders. Comunità locali (Salvatori et al. 2017), semplici appassionati o studenti (Mugelli et al. 2020; Crucitti et al. 2021), anche con Disturbi Specifici dell'Apprendimento (DSA), pur non possedendo specifiche competenze, potrebbero interagire con i testi nella loro veste originale.
Questo contributo intende riflettere su quali siano le difficoltà di comprensione di un testo manoscritto e su come permettere a utenti con formazioni diverse di approcciarsi alla lettura in contesti espositivi o didattici in modo interattivo e stimolante, attraverso la progettazione di un applicativo con un'interfaccia utente estremamente intuitiva. L'edizione digitale di testi con l'integrazione della digitalizzazione è stata già oggetto di diversi studi. Per esempio,Theleme[3] offre esercizi interattivi, EVT (Edition Visualization Technology) (Rosselli Del Turco et al. 2014; Cacioli et al. 2022) permette di visualizzare su finestre separate la digitalizzazione del testo e la trascrizione diplomatica, quella interpretativa o la traduzione. Non sembrano invece ancora disponibili applicativi pensati per un più ampio pubblico, finalizzati a facilitare la lettura dei manoscritti negli spazi espositivi o in contesti didattici. Alla base della progettazione di questi applicativi, si devono considerare le conclusioni di Rowe et al. (2017), che ha identificato alcuni aspetti comuni nelle applicazioni ritenute efficaci ai fini della didattica e della divulgazione, tra cui la facilità di utilizzo e l'incentivazione della curiosità e di un approccio esplorativo.

L'applicativo illustrato in questo contributo permette una lettura accompagnata del testo manoscritto, risolvendo passo dopo passo le sfide interpretative poste dalle grafie antiche, sfruttando ciò che Sutcliffe (2016) ha efficacemente definito "interattività migliorata". Il lettore non si trova davanti direttamente alla traduzione del testo, ma giunge attivamente a questa attraverso le tappe intermedie della trascrizione diplomatica, dello scioglimento delle abbreviazioni e della trasposizione in caratteri moderni.

2. LE BARRIERE ALL'ACCESSIBILITÀ DEL TESTO ANTICO

Riguardo all'accessibilità di un testo trasmesso da un manoscritto, si possono identificare in primo luogo le seguenti barriere:

- Stato di conservazione del supporto
- Tipologia di scrittura e varianti grafiche
- Utilizzo di sistemi abbreviativi
- Pronuncia delle parole
- Lingua del testo.

Al di là dello stato di conservazione del supporto, la lettura di un testo tràdito da un manoscritto, specialmente antico, è resa difficoltosa anzitutto dalla grafia, che risulta diversificata per peculiarità riconducibili a determinati contesti storici e geografici e quindi molto distante dalle odierne convenzioni. Questo aspetto, nella maggior parte dei casi, rappresenta per il lettore non specialista uno scoglio insormontabile che, unitamente alla presenza di eventuali varianti grafiche dovute all'*usus scribendi*, respinge già al primo approccio tutti quei potenziali fruitori che non abbiano avuto modo di acquisire alcuna competenza paleografica. Lo stesso vale per studenti che si avvicinino alle prime letture e alle trascrizioni dei manoscritti, tanto più quando manifestino Disturbi Specifici dell'Apprendimento (dislessia, disgrafia, disortografia e discalculia). Infatti, la distanza tra le grafie antiche e le caratteristiche compensative che i testi devono avere per essere fruibili da chi presenti DSA[4] può rivelarsi davvero incolmabile.
Quanto appena detto vale a maggior ragione se si considera che, anticamente, era frequente impiegare sistemi abbreviativi e segni tachigrafici (Cencetti 1954), sia generali sia propri di specifici ambiti, per esempio quello liturgico o giuridico (Giovè Marchioli 1991), che attualmente è possibile sciogliere soltanto a seguito di specifiche e iterate esperienze di lettura, supportate da strumenti specialistici quale il classico *Dizionario di abbreviature* del Cappelli (1912).

[3] http://theleme.enc.sorbonne.fr/dossiers/(Ultimo accesso 9/4/2025)
[4] British Dislexia Association, *Creating a dyslexia friendly workplace*
https://www.bdadyslexia.org.uk/advice/employers/creating-a-dyslexia-friendly-workplace/dyslexia-friendly-style-guide

Un ulteriore problema, connesso alla decodificazione del testo manoscritto e che può incidere anche sull'edizione dei testi, soprattutto in volgare, riguarda il rapporto tra grafia e pronuncia. In presenza di grafie dotte o di forme marcate diatopicamente, così come diastraticamente, il lettore non specialista potrebbe trovarsi in difficoltà non soltanto a comprendere forme linguistiche distanti dalla sua lingua standard, ma anche a replicarne i suoni nel momento di una lettura a voce alta. Più in generale, dal punto di vista della comprensione del testo, la stessa lingua utilizzata dallo scrivente potrebbe risultare difficilmente accessibile dal lettore nella sua morfologia, sintassi e lessico.

3. L'APPROCCIO SVILUPPATO

Abbiamo impostato l'applicativo in modo che eliminasse gradualmente le barriere di accessibilità alla lettura del manoscritto attraverso diversi gradi di esplicitazione del testo, dal riconoscimento delle singole lettere allo scioglimento delle abbreviazioni, per giungere al testo integrale e alla sua lettura (possibilmente secondo la corretta pronuncia del tempo) e a una sua traduzione i per facilitare il collegamento tra singoli termini ed espressioni (vd. Fig.1)

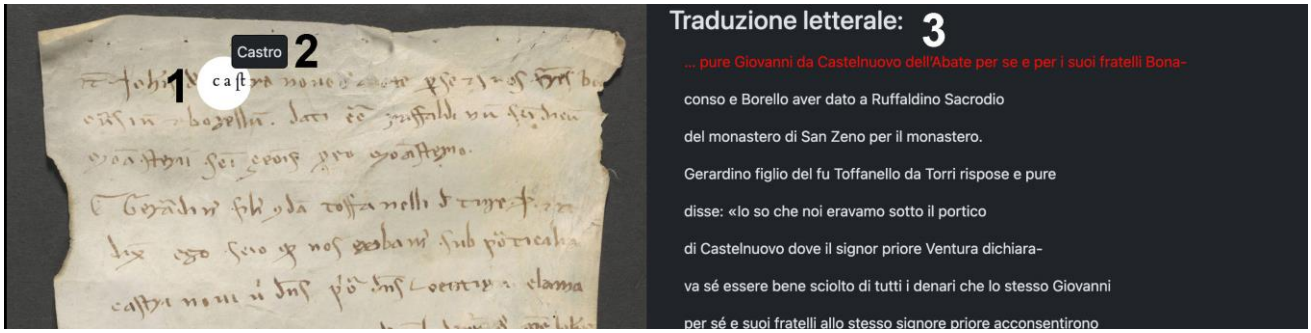


Figura 1. Screenshot di una porzione dell'interfaccia grafica, 1 – spotlight mobile che mostra la trascrizione usando i caratteri MUFI, 2 – tooltip con lo scioglimento della parola, 3 – pannello laterale con la traduzione letterale delle righe (la riga su cui è posizionato il cursore viene evidenziata in rosso).

L'esperienza utente si basa su un'interazione di tipo hover o on-mouse-over, in cui la semplificazione del testo è attivata dal focus dell'utente su una particolare parola. Nella pratica questo si traduce nel posizionare il cursore del mouse o del trackpad sopra la parola (nel caso di un sistema con queste periferiche) oppure nell'area in cui avviene l'interazione touch, se il dispositivo su cui viene installata l'applicazione supporta un touch-screen.
La visualizzazione base mostra la digitalizzazione dell'oggetto ad alta risoluzione. Nel nostro workflow di digitalizzazione seguiamo le linee guida internazionali ISO/TR 19263-1:2017 e FADGI per acquisire una riproduzione accurata del bene culturale. Un'immagine di cattiva qualità è, infatti, il primo impedimento all'accessibilità della digitalizzazione. Una riproduzione accurata, invece, permette all'utente di avere un'esperienza più vicina a quella di lettura del documento. Nel caso di mostre, il documento può essere messo in prossimità del totem multimediale dove è installata l'applicazione in modo da favorire il confronto e permettere poi la lettura direttamente sul documento originale.

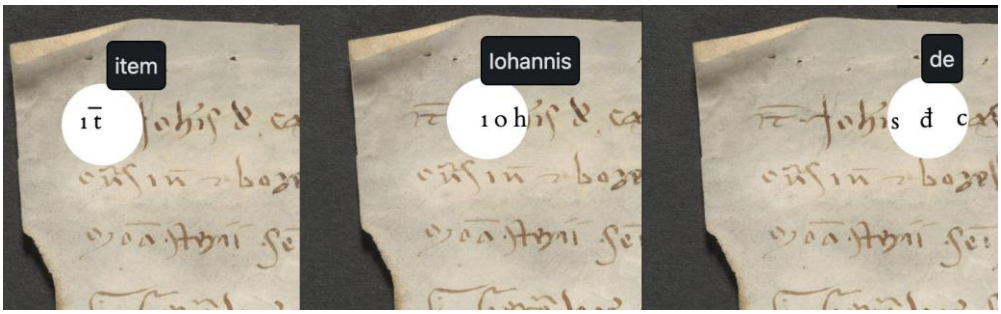


Figura 2. L'elemento spotlight al muoversi del cursore evidenza la trascrizione diplomatica e lo scioglimento delle abbreviazioni in tooltip.

La prima barriera di leggibilità del documento è rappresentata dalla tipologia di scrittura, su cui si impostano le oscillazioni grafiche dello scrivente. Per abbattere questa prima barriera usiamo quello che definiamo come 'spotlight', cioè un'area circoscritta nei pressi del punto di interazione dell'utente che permette la visualizzazione dell'area in esame semplificata (vd. Fig.2). Abbiamo utilizzato un font che supporti il Medieval Unicode Font Initiative (MUFI), in modo da poter trascrivere anche caratteri non più in uso come quelli utilizzati per particolari abbreviazioni, come i compendi e le tachigrafie (Ferrarini 2007: 108-111). L'utilizzo di un font analogo alla scrittura del documento permette di eliminare solo le variazioni grafiche della mano, ma spesso è più pratico utilizzare un font moderno per permettere un accesso diretto alla parola traslitterata. Per avere un'esperienza utente efficace, l'allineamento dei caratteri con il contenuto dell'immagine è fondamentale. I documenti antichi spesso non presentano una rigatura costante; bisogna quindi suddividere il testo in modo da posizionare le parole nell'area corrispondente della digitalizzazione. Operare a livello della singola parola si è rivelato un buon compromesso tra accuratezza del posizionamento e facilità della marcatura, anche se ci sono casi in cui un singolo glifo può non essere allineato con il resto della parola.
Per risolvere il terzo livello di accessibilità della scrittura, ovvero lo scioglimento delle abbreviazioni, abbiamo invece testato diversi approcci. Il primo approccio è stato quello di sciogliere l'abbreviazione direttamente nella spotlight: questo approccio, tuttavia, portava spesso a sovrapposizioni di testo che risultavano fastidiose per l'utente. Abbiamo quindi optato per l'utilizzo di tooltips, aree di testo che appaiono nei pressi dello spotlight, contenenti lo scioglimento dell'abbreviazione. Nelle tooltips viene utilizzato un font moderno che l'utente può leggere senza difficoltà, con lo scioglimento e la sostituzione dei caratteri speciali del MUFI nel loro corrispettivo moderno.
L'ultimo livello di difficoltà nella lettura del testo è la lingua del testo stesso. Gran parte del patrimonio librario medioevale europeo è in latino o in lingue che la maggior parte degli utenti non riescono a comprendere. L'applicativo che abbiamo sviluppato affianca alla digitalizzazione una traduzione letterale del testo. Il passaggio a quest'ultimo livello di accessibilità è sicuramente il più complicato, e il rischio che si perda la progressione nella lettura del testo è molto alto. Spesso collegare univocamente la parola nella lingua originale alla sua traduzione nella finestra affiancata non è possibile, anche in ragione di diversi costrutti sintattici; abbiamo quindi scelto di evidenziare semplicemente la riga di riferimento e proporre traduzioni il più possibile letterali, in modo che l'utente riesca comunque a ricollegare la parola al testo tradotto. L'approccio è comunque personalizzabile a seconda del documento che si vuole rendere accessibile. Per esempio, nel caso di un documento senza abbreviazioni e il cui testo può essere letto facilmente attraverso l'uso dello spotlight, si potrebbe sfruttare il tooltip per mostrare la parola tradotta e quindi mostrare una traduzione meno letterale nella finestra affiancata.
Infine, è possibile associare un audio alle singole parole, in modo da permettere all'utente di ascoltare come queste venivano presumibilmente pronunciate, facendo in modo che la pronuncia corrisponda a specifici usi storicamente e geograficamente definiti, evitando, per esempio, che vengano mal interpretate alcune particolarità grafiche, come la resa della laterale palatale [ʎ] o della nasale palatale [ɲ] indifferentemente con *gl* o *lg* e *gn* o *ng*, o la restituzione delle occlusive velari sorde e sonore e delle affricate alveolari.

4. LE SCELTE PER L'IMPLEMENTAZIONE DELL'APPLICATIVO

Lo sviluppo di un prototipo di visualizzatore è stato fatto utilizzando tecnologie standard: HTML, CSS, SVG e Javascript. L'intento è quello di facilitare la creazione di applicativi personalizzati a seconda delle esigenze didattiche o della tipologia di testo che si vuole esporre.
La digitalizzazione e la trascrizione diplomatica viene strutturata in un file SVG che viene poi incorporato nel file HTML. Il file SVG contiene il link alla riproduzione fotografica del documento e la trascrizione con la traduzione del testo, gestite a diversi livelli di granularità. Ad ogni parola viene assegnato un id come attributo dell'elemento <text> che contiene un sotto elemento <tspan> con le coordinate dell'area di testo e il cui testo contiene la trascrizione diplomatica della parola in esame. Gli attributi globali[5] vengono usati per aggiungere la trascrizione con le abbreviature sciolte (data-trasc) ed eventualmente la traduzione della singola parola (data-trad). Le parole appartenenti alla stessa riga vengono raggruppate in un elemento group <g> che contiene anch'esso l'attributo data-trad con la traduzione dell'intera frase. Il seguente estratto mostra come viene codificata la prima parola della riga con id "g44":

[5] https://developer.mozilla.org/en-US/docs/Web/HTML/Global_attributes/data-*

```
<g data-trad="disse: «Io so che noi eravamo sotto il portico" id="g44">
 <text id="text37">
  <tspan x="10" y="195" id="parola0036" data-trasc="dixit:" data-trad="Egli ha detto:">dıx</tspan>
  </text>
   ....
</g>
```

Degli script Javascript consentono l'estrazione delle traduzioni, la possibilità di vedere nella tooltip la forma sciolta delle parole (o la loro traduzione), implementano lo spotlight e infine permettono l'eventuale connessione dei file audio con la pronuncia della parola.
Oltre al fatto che i file SVG possono essere visualizzati in tutti i browser, un altro importante vantaggio di questo formato è la possibilità di utilizzare direttamente dei software per la grafica vettoriale al fine di realizzare la trascrizione del testo. Nel nostro caso, abbiamo usato il software open-source Inkscape. Abbiamo poi impiegato un editor di testo per aggiungere gli elementi data-trasc e data-trad. Questo processo può essere ulteriormente automatizzato sviluppando degli script che aggiungano la traduzione da un file esterno o che utilizzino un dizionario per sciogliere le abbreviazioni. Creando uno script Python è stato possibile, per esempio, tradurre automaticamente le singole parole e aggiungere la traduzione all'attributo data-trasc della parola corrispondente. Allo stesso modo è possibile creare uno script per esportare il dato in formato TEI. La TEI, infatti, definisce un modo standardizzato per associare a digitalizzazioni la loro trascrizione e associare alla digitalizzazione un sistema di riferimento che può essere usato per abbinare il testo trascritto. Questo esempio dalla documentazione ufficiale[6] mostra come collegare un'immagine attraverso l'elemento <graphic> è inoltre possibile creare delle sotto aree usando l'elemento <zone> mentre le sotto aree permettono di gestire layout più complessi[7].

5. DISCUSSIONE

La creazione di strumenti per aumentare l'accessibilità di un documento manoscritto e la valutazione in ordine alla loro efficacia (Mancinelli & Pierazzo, 2020: 105-107) possono portare notevoli vantaggi alla fruizione del bene. In questo contributo abbiamo definito quali sono le barriere di accessibilità di un testo trasmesso da un manoscritto anche per gli utenti che vi si avvicinano senza specifiche competenze. Sebbene esistano già degli strumenti che facilitano la visualizzazione e l'interpretazione del testo con un confronto diretto con la digitalizzazione, questo contributo ha elencato i vantaggi di un approccio basato su interazioni hover e quello che abbiamo definito spotlight, cioè un'area di indagine mobile che mostra la trascrizione della porzione di testo su cui l'utente sta dirigendo la sua attenzione. Rispetto a un approccio più tradizionale, che utilizzi diverse finestre affiancate, abbiamo riscontrato i seguenti vantaggi:

- c'è un confronto diretto con la fonte primaria e la trascrizione: l'utente non deve spostare l'attenzione tra le due finestre ma può controllare direttamente l'area in esame notando differenze tra la trascrizione e la digitalizzazione[8].
- l'interazione è immediata e più immersiva: non è richiesta una particolare formazione dell'utente per utilizzare l'applicativo, l'interazione si sviluppa attraverso lo spostamento del cursore sull'area di interesse.
- l'utente può decidere quando vuole essere aiutato nella lettura del testo: egli modula lo sforzo cognitivo a seconda del suo interesse, valutando quando ricevere l'aiuto e dove focalizzare la propria attenzione.

RINGRAZIAMENTI

Gli autori ringraziano Andrea Giacchetti per i preziosi consigli in fase di stesura dell'articolo.

BIBLIOGRAFIA

Basili, C., Biorci, G., & Emina, A. (2017). Digital Humanities and Society: An impact requiring 'intermediation'. Umanistica Digitale, 1, Articolo 1. https://doi.org/10.6092/issn.2532-8816/7196

Cacioli, G., Cerretini, G., Di Pietro, C., Maenza, S., Rosselli Del Turco, R., & Zenzaro, S. 2022. "There and back again: what to expect in the next EVT version". In Ciracì, F., Miglietta, G., & Gatto, C. (Eds.),

[6] https://tei-c.org/release/doc/tei-p5-doc/it/html/PH.html#PHFAX
[7] https://tei-c.org/release/doc/tei-p5-doc/it/html/PH.html#PH-surfzone
[8] Questo oltre ad essere potenzialmente uno strumento didattico può essere anche uno strumento di revisione molto efficace.

AIUCD 2022 – Digital cultures. Intersections: philosophy, arts, media. Proceedings of the 11th national conference, Lecce, 1-3 June 2022: 212-217. http://amsacta.unibo.it/6848

Cappelli, A (1912). Lexicon abbreviaturarum. Dizionario di abbreviature latine ed italiane, Hoepli.

Cencetti, G. (1954). Lineamenti di storia della scrittura latina, Bologna: Patron.

Crucitti, M., Benedetti, M., Mirandola, R., Maneschi, G., Soldani, A., Amato, L., Lepori, F., Taddei, A., & Boschetti, F. (2021). La collaborazione inclusiva: Un'esperienza didattica di annotazione tramite Euporia. Umanistica Digitale, 11, Articolo 11. https://doi.org/10.6092/issn.2532-8816/13680

Ferrarini, E. (2007). "La trascrizione dei testimoni manoscritti: metodi di filologia computazionale". In Ciula, A., & Stella, F. (Eds.), Digital Philology and Medieval Texts. Proceedings of the Arezzo Seminar 2006, 19-21 January. Pisa: Pacini, 103-20. https://iris.univr.it/handle/11562/309385

Giovè Marchioli, N. (1991). Per una storia delle abbreviature latine. Gazette du livre médiéval, 18(1), 17–24. https://doi.org/10.3406/galim.1991.1159

Mancinelli T., & Pierazzo E. (2020). Che cos'è un'edizione scientifica digitale. Roma: Carocci.

Mugelli, G., Re, G., & Taddei, A. (2020). Annotazione digitale di testi antichi. Lingue antiche e Digital Humanities, tra ricerca e didattica. Umanistica Digitale, 9, Articolo 9. https://doi.org/10.6092/issn.2532-8816/9962

Rosselli Del Turco, R., Buomprisco, G., Pietro, C. D., Kenny, J., Masotti, R., & Pugliese, J. (2014). Edition Visualization Technology: A Simple Tool to Visualize TEI-based Digital Editions. Journal of the Text Encoding Initiative, Issue 8, Articolo Issue 8. https://doi.org/10.4000/jtei.1077

Rowe, J. P., Lobene, E. V., Mott, B. W., & Lester, J. C. (2017). Play in the museum: Design and development of a game-based learning exhibit for informal science education. International Journal of Gaming and Computer-Mediated Simulations (IJGCMS), 9(3), 96–113.

Salvatori, E., Turco, R. R. D., Alzetta, C., Pietro, C. D., Mannari, C., & Miaschi, A. (2017). Il Codice Pelavicino tra edizione digitale e Public History. Umanistica Digitale, 1, Articolo 1. https://doi.org/10.6092/issn.2532-8816/7232

Sutcliffe, A. (2016). Designing for User Experience and Engagement. In H. O'Brien & P. Cairns (Eds.), Why Engagement Matters (pp. 105–126). Springer International Publishing. https://doi.org/10.1007/978-3-319-27446-1_5

Accessibilità e inclusione per la documentazione del restauro: gli archivi del Centro Conservazione e Restauro "La Venaria Reale"

Stefania De Blasi, Responsabile Area Comunicazione e Documentazione, Fondazione Centro Conservazione e Restauro "La Venaria Reale", Italia, stefania.deblasi@ccrvenaria.it
Edi Guerzoni, PhD student, Università di Torino - Fondazione Centro Conservazione e Restauro "La Venaria Reale", Italia, eguerzoni@ccrvenaria.it
Chiara Pipino, archivista, Area Comunicazione e Documentazione, Fondazione Centro Conservazione e Restauro "La Venaria Reale", Italia, chiara.pipino@ccrvenaria.it

ABSTRACT (ITALIANO)
Questo contributo vuole presentare il progetto di accessibilità degli archivi del Centro Conservazione e Restauro "La Venaria Reale" (CCR). La digitalizzazione e la pubblicazione dei dati riguardanti le opere restaurate dal CCR hanno infatti previsto un'ultima fase di facilitazione della consultazione attraverso una metadatazione con testi facilitati realizzati con gli strumenti Easy to Read. Sono state realizzate circa 1200 schede per la fruizione del sito in linguaggio accessibile, divise tra le schede dedicate all'archivio dei restauri del CCR (realizzati tra 2005 e 2024) e altre dedicate al Fondo Pinin Brambilla Barcilon. Il fine è quello di una sensibilizzazione, non solo alla protezione del patrimonio, ma anche nei confronti di discipline scientifiche e tecniche che attorniano la disciplina del restauro, per sua natura ibrida tra gli approcci umanistici e quelli del settore scientifico.
Parole chiave: conservazione; restauro; accessibilità; digitalizzazione; archivio

ABSTRACT (ENGLISH)
This paper presents the accessibility project developed for the archives of the Centro Conservazione e Restauro "La Venaria Reale" (CCR). As part of the digitisation and publication of records related to artworks restored by the CCR, a final phase was introduced to enhance consultation through metadata and Easy to Read texts. Around 1,200 accessible records were created—some concerning the CCR's restoration archive (2005–2024), others related to the Pinin Brambilla Barcilon collection. The project aims to raise awareness not only about the importance of heritage protection, but also about the scientific and technical disciplines that support conservation—a field inherently situated between the humanities and science.
Keywords: conservation; restoration; accessibility; digitisation; archive

1. INTRODUZIONE
Le attività di digitalizzazione e metadatazione dei fondi archivistici conservati presso la Fondazione Centro per la Conservazione ed il Restauro dei Beni Culturali "La Venaria Reale" (d'ora in poi: CCR) sono state parte del più ampio progetto di accessibilità fisica e cognitiva messo in atto dal CCR a partire dal 2023. Per la digitalizzazione e valorizzazione di un archivio è oggi fondamentale partire da fonti che prevedano già in origine l'orientamento all'accessibilità (Marin, A., 2017). La sfida di sviluppare le attività d'archivio e valorizzare le *digital humanities* per un istituto che si occupa di formazione, conservazione e ricerca nell'ambito dei beni culturali (Lewi, H., (2020) ha trovato una possibilità nel progetto dedicato all'abbattimento delle barriere fisiche e cognitive degli archivi, biblioteca e percorsi di visita, grazie a fondi NextGenerationEU erogati dal Ministero della Cultura attraverso i bandi per il PNRR[1]. Il CCR ha previsto un importante cambiamento delle strutture digitali del dipartimento di documentazione con la priorità dell'accessibilità indicata nel bando. Oltre al patrimonio librario di 30.000 volumi, presso la biblioteca ha sede l'archivio dei restauri del CCR, con oltre 3200 interventi di restauro su circa 8000 opere e oltre 135.000 file di immagini ad alta risoluzione, elaborati nel corso dei vent'anni di vita del CCR. La biblioteca ospita inoltre donazioni di fondi bibliografici, tra cui il Fondo e l'Archivio di Pinin Brambilla Barcilon (1925-2020), restauratrice italiana che restaurò il *Cenacolo* di Leonardo da Vinci e che ebbe un ruolo cardine anche per la creazione del CCR stesso. Uno degli obiettivi previsti dal progetto è stato quello di diffondere i contenuti della biblioteca e dell'archivio attraverso strumenti di divulgazione universalmente accessibili con

[1] Bando del Ministero della Cultura per la presentazione di proposte progettuali di intervento finalizzate alla rimozione delle barriere fisiche, cognitive e sensoriali, per consentire un più ampio accesso e partecipazione alla cultura, in questo caso, del restauro. Il progetto dal titolo "Humanities in Conservation" è stato finanziato nell'ambito del PNRR, Intervento 1.2, MC1C3.3, sostenuto dall'Unione europea – NextGeneration EU.V

modalità inedite per il settore della conservazione e degli archivi. Il primo obiettivo del progetto è stato caratterizzato dalla digitalizzazione e la sistematizzazione della documentazione fotografica e degli interventi di restauro all'interno di un *database web based* adottato e personalizzato sulla base delle esigenze del CCR.

2. IL PROGETTO

A partire dalla piattaforma di archiviazione sono stati scelti 1200 restauri emblematici per lessico, tipologia di intervento e complessità, da rendere liberamente fruibili ad un pubblico allargato direttamente sul sito del CCR[2]. Le descrizioni sono state elaborate in linguaggio accessibile, includendo sia attività eseguite dai 9 laboratori del CCR tra 2005 e il 2024, sia interventi storici provenienti dal Fondo Brambilla Barcilon. Una strategia di più ampia visibilità dell'archivio digitale, che riporta contenuti sullo studio delle opere e sugli interventi di conservazione, è stata perseguita attraverso la creazione di codici QR posizionati in prossimità delle opere restaurate e ritornate presso i luoghi di provenienza, con un rimando diretto allo spazio digitale dell'archivio[3]. Mentre il materiale dell'archivio CCR si presentava come nativo digitale, sia per le immagini che per la documentazione tecnica, ma non ancora sistematizzato in un database, la documentazione del Fondo Brambilla Barcilon era interamente cartacea e analogica. L'archivio Pinin Brambilla Barcilon è stato conferito nel 2017 al CCR ed è composto da materiale prodotto dalla restauratrice durante gli oltre 50 anni di attività[4]. Il progetto ha previsto una prima digitalizzazione a tappeto di fotografie, rilievi, mappature e relazioni tecniche, a cui è seguita una informatizzazione e catalogazione all'interno della piattaforma adottata dal CCR. Il lavoro di traduzione linguistica in italiano facilitato si è svolto, invece, in corrispondenza della definizione di un layout per il sito web che, attraverso i codici API (Application Programming Interface), acquisisce i contenuti resi visibili nel back end della piattaforma.

3. DALLA TEORIA ALLA PRATICA

Questo progetto fonda le sue basi deontologiche sull'importanza della circolazione della conoscenza nel campo della tutela del patrimonio storico-artistico attraverso progetti di digital humanities (Murphy B. L., 2016) nell'ambito di istituzioni culturali e archivistiche legate al patrimonio (Drucker, J., 2021). Se in ambito storico-artistico e della cultura visuale tout court queste prospettive sono sempre più attuali, resta da approfondire come applicarle anche all'ambito della storia del restauro e della conservazione del patrimonio, oltre che in ambito archivistico (Carmeliti, M., et al., 2023). La comunicazione delle attività conservative è, di fatto, parte essenziale della tutela[5] e ha un duplice impatto: da una parte, per l'educazione e sensibilizzazione ai temi della tutela; dall'altra, in ambito scientifico e specialistico, per la condivisione di metodi e materiali d'intervento. Isabelle Dahy (Dahy, I., 2003) pone la questione di quali modalità di comunicazione seguire da parte degli operatori della conservazione per aprirsi al pubblico. Il fine è quello di una sensibilizzazione, non solo alla protezione del patrimonio, ma anche nei confronti di discipline scientifiche e tecniche che attorniano il settore del restauro, per sua natura ibrida tra gli approcci umanistici e quelli del settore scientifico. Il campo del restauro, dalla data di pubblicazione del contributo di Isabelle Dahy, ha avuto un'importante evoluzione dettata da una riconosciuta autorevolezza in ambito scientifico e accademico, soprattutto in Italia con l'introduzione della formazione del restauratore nel

[2] https://www.centrorestaurovenaria.it/archivio-digitale-ccr/

[3] Tale attività è stata realizzata in accordo con enti proprietari delle opere restaurate dal CCR nel corso degli anni. Una prima presentazione del progetto ha visto l'allestimento dei QR code con rimando alle schede accessibili dell'archivio digitale direttamente nel percorso di visita della Reggia di Venaria e della Palazzina di Caccia di Stupinigi.

[4] La documentazione è stata conservata nel Laboratorio di Restauro Pinin Brambilla a Milano fino al giugno 2017, quando, grazie al sostegno di Ra.Mo S.p.A, è stata trasferita interamente al CCR. Nel 2019 è stato portato a termine l'inventario dell'archivio a cura di Stefania Circosta. Nel 2023 è stata attivata una borsa di dottorato nell'ambito del corso di Dottorato Patrimonio Culturale e Produzione Storico-Artistica, Audiovisiva e Multimediale dell'Università degli Studi di Torino cofinanziata dal CCR e attribuita a Edi Guerzoni con una ricerca dal titolo *La documentazione e la divulgazione del restauro. La storia del restauro e la sua narrazione attraverso il fondo archivistico Pinin Brambilla Barcilon.* In questo modo è stato possibile incrociare le competenze e avviare il progetto di digitalizzazione e studio dell'archivio.

[5] "A museum is a not-for-profit, permanent institution in the service of society that researches, collects, conserves, interprets and exhibits tangible and intangible heritage. Open to the public, accessible and inclusive, museums foster diversity and sustainability. They operate and communicate ethically, professionally and with the participation of communities, offering varied experiences for education, enjoyment, reflection and knowledge sharing.", Definizione di Museo, ICOM 2022.

Codice dei Beni Culturali, ma anche dalla sempre maggiore incidenza dei media e della divulgazione online. Tuttavia, rimane da colmare un vuoto di conoscenza tecnica di chi si occupa di mediazione culturale e linguistica sul tema specifico e complesso della conservazione. In generale, nel 2003 come oggi, si tratta ancora di creare una relazione più forte tra la conservazione dei beni culturali e il pubblico, effettivo possessore di tali beni. Nel caso degli archivi del CCR si è optato per due diversi livelli di target: uno per il pubblico specializzato, che avrà la possibilità di consultare online in libero accesso le fotografie degli interventi di restauro direttamente sul sito del Centro e, previa richiesta, approfondire consultando le relazioni tecniche e scientifiche dei restauri, sia per il fondo del CCR che per quello di Pinin Brambilla Barcilon. Un altro livello, immediatamente visibile e disponibile sempre sul nuovo sito internet del CCR, rilasciato a novembre 2024, è destinato ad una concezione allargata di pubblico universale, con una particolare attenzione all'accessibilità cognitiva e culturale. Un significativo numero di schede è stato pubblicato con fotografie e metadati presentati in lingua italiana facilitata per pubblici diversi, comprensiva di audiodescrizioni e traduzioni automatiche del testo in diverse lingue[6]. I metadati relativi al materiale digitalizzato sono stati redatti in base alle informazioni archivistiche e all'interpretazione critica, prevedendo un livello di restituzione altamente accessibile, in linguaggio Easy to Read, adatto alla comprensione da parte di persone con deficit cognitivi, neurodivergenti, persone sorde e appartenenti a diversi background culturali (Puthiya Purayil S., 2022). Il progetto permette di valutare quali siano le difficoltà e le soluzioni della digitalizzazione di fonti d'archivio (Jaillant, 2022), come l'impiego di Large Language Model (LLM), non ancora adeguati per la creazione di testi accessibili nell'ambito del restauro, e un'analisi critica da parte di personale specializzato sui temi dello universal design for all, per il passaggio dalla digitalizzazione alla restituzione in termini di comunicazione. Nel caso del fondo storico Pinin Brambilla Barcilon, la digitalizzazione si è concentrata principalmente sulla sezione più significativa e con maggiore consistenza, relativa a "Cantieri e Restauri". Questa sezione, i cui estremi cronologici sono 1964-2017, è quella contenente il più alto numero di fototipi e di relazioni, divise in ordine topografico a seconda del cantiere di restauro di afferenza. Tale struttura è stata mantenuta e conservata in qualità di fonte storica essa stessa. Le relazioni e le immagini sono state digitalizzate e riversate in una piattaforma open source customizzata.

4. GLI STRUMENTI INFORMATICI

L'infrastruttura della piattaforma è costituita dall'utilizzo di due software open source, "Collective Access" e "Pimcore"[7], e da due layer principali: un PIM[8] modificato e ottimizzato per creare e gestire un archivio storico e un'interfaccia di consultazione a uso esclusivo aziendale, protetta da una profilazione utente. Il livello di back end è dedicato all'organizzazione, alla gestione e all'arricchimento dei dati esclusivamente in chiave storica e funziona sulla base di un PIM (Pimcore) e un database di archiviazione (Collective Access), popolato da schede metadatate in maniera differente a seconda della tipologia (es. prodotti, oggetti, opere, luoghi, persone, eventi, ecc.), a cui sono associati media e relazioni tipologiche (per il collegamento ad altre schede, es. autore, luogo, ...). Il PIM in questione è stato ed è regolarmente arricchito con strumenti sviluppati per rendere possibile la creazione e gestione ordinaria e straordinaria di un archivio

[6] Un caso esemplare per applicare questi presupposti teorici è stato proprio il progetto sul fondo archivistico Pinin Brambilla Barcilon, con la digitalizzazione del fondo e alla conseguente restituzione dello stesso attraverso una metadatazione specifica volta ad ottenere una comunicazione accessibile dei contenuti d'archivio, per lo più report di restauri effettuati tra anni Sessanta del Novecento e anni Duemila. Dal Fondo Brambilla, costituito da 455 faldoni contenenti circa 50000 immagini di cui sono stati digitalizzati i fototipi positivi e la completezza dei report di restauro.

[7] Lo strumento si chiama "Archivio Platform" (già "Caveau digitale") ed è stato ideato e sviluppato da Promemoria Group con lo scopo di valorizzare il patrimonio storico di aziende ed enti culturali.

[8] Product Information. Si riferisce al processo di gestione di tutti i dati, contenuti e altro materiale necessario per vendere e promuovere prodotti. Il PIM, inoltre, garantisce che vengano creati dati di qualità per uso interno e distribuzione multicanale.

storico, garantendo il rispetto degli standard scientifici attuali[9]. Le principali caratteristiche dell'interfaccia front end, invece, permettono la gestione di diversi gruppi di utenti con altrettanti differenti privilegi di accesso; cercare, consultare e filtrare le schede per data (timeline), tipologia, categorie, tag gerarchici, presenza/assenza di media e numerose altre attività di ordinamento del materiale. L'interfaccia front end attraverso un sistema di API permette l'organizzazione dei dati che si sono voluti rendere pubblici e senza necessità di login, direttamente sulla landing page del sito del CCR. Un limite che dovrà essere superato implementando le funzioni della piattaforma "Archivio Platform" di Promemoria Group consiste nel rendere accessibile con alternative testuali la descrizione delle immagini. Una volta implementata tale funzione sulla piattaforma e previste apposite API, anche sulla landing page del sito dell'archivio digitale sarà possibile rendere visibili i testi alternativi. Al momento per ogni scheda facilitata dell'Archivio restauri CCR è stata predisposta una descrizione testuale della raffigurazione dell'opera. Il lavoro di perfezionamento sarà l'obiettivo dello staff dell'archivio che è in corso di stabilizzazione grazie a investimenti sul piano strategico scientifico del CCR, che ha tra le finalità lo sviluppo dell'area dedicata alle digital humanities e valorizzazione dell'archivio (Videtta C., 2020).

5. LESSICI TECNICI, ACCESSIBILITÀ E INTELLIGENZA ARTIFICIALE

Il lavoro di metadatazione dell'archivio e il passaggio critico tra il registro linguistico della documentazione tecnica e la sua traduzione ad un registro di comunicazione facilitata ha fatto emergere la necessità di realizzare un glossario come elemento complementare dello sviluppo del progetto. Negli ultimi quarant'anni nel settore della conservazione diversi gruppi di lavoro e letteratura tecnico scientifica sono stati dedicati alla normalizzazione dei glossari tecnici relativi alla denominazione degli oggetti, delle tecniche esecutive, dei materiali e dei fenomeni di degrado[10]. A valle di questa disamina critica relativa al tema dei lessici è risultato inesistente ma imprescindibile un glossario che rendesse la terminologia tecnica accessibile secondo le vigenti norme europee. Realizzare un tale glossario è stato proposto come obiettivo di un laboratorio dal titolo Cultural Accessibility: Principles and Methodologies for Cultural Heritage, organizzato durante la IV edizione dello "Young Professional Forum 2024", rassegna annuale del CCR. L'esperienza laboratoriale ha offerto una lezione partecipativa sulla scrittura accessibile, applicando le tecniche di scrittura accessibile al campo della conservazione. L'obiettivo è stato quello di aggiungere l'ausilio del glossario alla digitalizzazione. Il glossario, ora in corso di pubblicazione con traduzione in inglese, arabo e lingua dei segni, sarà funzionale a rendere le informazioni e le conoscenze nel campo della conservazione del patrimonio culturale più accessibili a un pubblico ampio. Il volume sarà fruibile liberamente sul sito, periodicamente incrementato e potrà costituire per altri enti e professionisti un punto di partenza per nuovi progetti di accessibilità della conservazione e del restauro. L'esperienza ha portato alla creazione del glossario per i primi cinquanta termini maggiormente ricorrenti nelle relazioni di restauro conservate nell'Archivio restauri CCR e nell'Archivio Pinin Brambilla Barcilon. La scelta ha riguardato, quindi, la terminologia attualmente utilizzata e normalizzata e quella relativa a interventi storicizzati. Il

[9] Di seguito alcune delle funzionalità di Pimcore che si sono rivelate utili ai fini della digitalizzazione: utilizzare delle API di autenticazione (o sistema SSO); la creazione dei record di oggetti a cui associare i media digitali; l'uso di diverse classi di oggetti con tracciati descrittivi individuali per ogni classe; la possibilità di creare dei tracciati descrittivi da interfaccia (metadati, relazioni) senza richiedere assistenza tecnica; creare metadati di diversa tipologia: testo libero, rich text, select, multiselect, relazioni, data, campi numerici; creare relazioni tra record: one to one, one to many, many to many, e visualizzare relazioni bi-univoche;, visualizzazione a elenco di una selezione di record o risultati di ricerca con la possibilità di vedere e modificare rapidamente il contenuto di tutti i metadati, tag, relazioni, anteprime dei media collegati; ricerca semplice, gerarchica, avanzata (su tutti i metadati) e per tag; la possibilità di modificare massivamente dei record in base alla collocazione gerarchica (seleziona nodo padre e compila tutti i figli o una selezione di figli) o in base a un criterio di ricerca avanzata (modifica tutti i risultati di ricerca o una selezione di risultati).

[10] Per una disamina diacronica del tema che riprende i principali progetti si veda *La questione dei lessici, l'Associazione Giovanni Secco Suardo e il progetto europeo LMCR - Lessico tecnico Multilingue di Conservazione in* De Blasi S., (2016) Il restauro silenzioso. La conservazione preventiva: un sistema sostenibile di gestione e controllo, Cronache 6, Firenze, Nardini Editore, 2016, pp. 105 – 117, si veda in particolare i volumi sui lessici: Nimmo M., (a cura di), (2001), *La pittura murale, proposta per un glossario, Regione Lombardia – Associazione Giovanni Secco Suardo Lurano*, 2001; Giannini C. (a cura di), (2010), Dizionario del restauro. Tecniche, diagnostica, conservazione, Firenze, Nardini Editore, 2010; Weyer Angela, Roig Picazo Pilar, Pop Daniel, Cassar Joann, Özköse Aysun, Vallet Jean-marc, Srša Ivan, Petersberg, Imhof Verlag Michael, (a cura di), *EwaGlos, European illustrated glossary of conservation terms for wall paintings and architectural surfaces. English definitions with translations into Bulgarian, Croatian, French, German, Hungarian, Italian, Polish, Romanian, Spanish and Turkish,* 2015.

progetto ha consentito inoltre la sperimentazione di sistemi LLM che si sono rivelati non ancora sufficientemente strutturati per recepire tali argomenti. Una traccia a cui il CCR sta lavorando è la specializzazione degli strumenti su basi Easy to Read e attraverso le implementazioni di definizioni facilitate del restauro idonee a una prima stesura di metadatazione a partire dalle fonti, con l'obiettivo di semplificare e velocizzare il processo di creazione dei testi, seguito dalla imprescindibile validazione da parte di specialisti del settore. A partire da questi presupposti il tema dell'implementazione di sistemi di Intelligenza Artificiale, che sono alla base di progetti di ricerca di cui il CCR è partner tecnico, sono l'esito atteso sia per la generazione di contenuti scientifici che per la normalizzazione di contenuti facilitati[11].

6. DIRITTI DI RIPRODUZIONE, CREATIVE COMMONS E AUTORIZZAZIONI

Un ulteriore piano di aggiornamento e sviluppo dei progetti relativi agli archivi digitali del CCR consiste nella definizione dei diritti associati alle immagini. Il CCR rappresenta un caso di studio attualmente al vaglio di pareri giuridici nazionali e comunitari. Infatti, la totalità delle immagini custodite e digitalizzate è riferita ad opere che non sono di proprietà della Fondazione CCR, ma di enti differenti con requisiti di accesso specifici e sovente non paragonabili, poiché normati da discipline e processi autorizzativi diversi (es. opere pubbliche ministeriali, opere pubbliche territoriali, opere di istituti internazionali, opere private, opere di enti ecclesiastici). È attualmente in corso una campagna di acquisizione delle autorizzazioni ai singoli enti che sarà la base per definire i livelli di riproducibilità delle immagini e di riutilizzo. Per il momento il temporaneo livello che dovrà essere a breve pubblicato sul sito del CCR indicherà la licenza di massima restrizione CC BY-NC-ND 4.0 in attesa di sciogliere le riserve rispetto a un utilizzo più libero delle immagini (Orlandi S.D., et al., 2021). Tutte le immagini sono scaricabili a bassa definizione e con richiesta di citazione come riportato nella pagina principale della maschera di ricerca dell'archivio digitale[12].

[11] In particolare, tasks relativi alla generazione attraverso AI di contenuti corretti e normati sono parte del progetto "AI4Re-Story: storie di restauro raccontate dall'intelligenza artificiale" condotto dal CCR insieme al Dipartimento di Studi Storici dell'Università di Torino, HighEst Lab, MIT Technology Review e finanziato da Fondazione CRT.

[12] https://www.centrorestaurovenaria.it/archivio-digitale-ccr/

BIBLIOGRAFIA

Carmeliti, Maria, Marziali, Stefano, Catalano Chiara Eva, (2023), Catalogazione e gestione del restauro in archivi digitali, in Lo Stato dell'arte 21, XXI Congresso Nazionale IGIIC, Firenze, Nardini, 2023

Dahy, Isabelle (2003), La conservation-restauration du patrimoine et la médiation culturelle, Lettre de l'OCIM, n°86, 2003

De Blasi S., (2016) Il restauro silenzioso. La conservazione preventiva: un sistema sostenibile di gestione e controllo, Cronache 6, Firenze, Nardini Editore, 2016

Drucker, Johanna, (2021), The Digital Humanities Coursebook. An introduction to Digital Methods for Research and Scholarship, Londra – New York, Routledge, 2021

Giannini C. (a cura di), (2010), Dizionario del restauro. Tecniche, diagnostica, conservazione, Firenze, Nardini Editore, 2010

Jaillant, Lise, (a cura di), (2022), Archives, access, and Artificial Intelligence. Working with born digital and Digitized Archival Collections, Bielefeld, Bielefeld University Press, 2022

Lewi Hannah (a cura di) (2020), The Routledge International Handbook of New Digital Practices in Galleries, Libraries, Archives, Museums and Heritage Sites, Londra – New York, Routledge, 2020

Marin, Anne-Catherine, (2017). Archivistes, tous médiateurs ? Petites réflexions sur les pratiques professionnelles, in "Gazette des archives", Anno 2017, n. 247

Murphy Bernice L., (a cura di), (2016), Museums, Ethics and Cultural Heritage, ICOM Series, Londra – New York, Routledge, 2016

Nimmo M., (a cura di), (2001), La pittura murale, proposta per un glossario, Regione Lombardia – Associazione Giovanni Secco Suardo Lurano, 2001

Puthiya Purayil Sneha, Alternative Histories of Digital Humanities: Tracing the Archival Turn, in Domenico Fiormonte, Sukanta Chaudhuri, Paola Ricaurte (a cura di), Global Debates in the Digital Humanities, Minneapolis, University of Minnesota Press, 2022

Sarah Dominique Orlandi, Anna Maria Marras, Deborah De Angelis, Pierfrancesco Fasano, Cristina Manasse, & Mirco Modolo. (2021). Open Acces - Diritto D'autore, Copyright E Licenze Aperte Per La Cultura Nel Web (Versione 2021) [Computer software]. Zenodo. https://doi.org/10.5281/zenodo.4593914

Weyer Angela, Roig Picazo Pilar, Pop Daniel, Cassar Joann, Özköse Aysun, Vallet Jean-marc, Srša Ivan, Petersberg, Imhof Verlag Michael, (a cura di), EwaGlos, European illustrated glossary of conservation terms for wall paintings and architectural surfaces. English definitions with translations into Bulgarian, Croatian, French, German, Hungarian, Italian, Polish, Romanian, Spanish and Turkish, 2015.

Videtta Cristina, (2020), Sviluppo sostenibile: un ruolo per la tutela del patrimonio culturale, Giornata di studi Gli archivi leader della sostenibilità, 16 settembre 2020, Milano

Il patrimonio culturale digitale delle minoranze etniche: il progetto DIGICHer tra le comunità Sámi, ladine ed ebraiche in Europa.

Matteo Cova[1], Eleonora De Longis[2]
[1] Istituto Italiano di Studi Germanici, Italia – cova@studigermanici.it
[2] Istituto Italiano di Studi Germanici, Italia – delongis@studigermanici.it

ABSTRACT (ITALIANO)
Il contributo presenta obiettivi e metodologia di sviluppo del progetto europeo DIGICHer, dedicato alla revisione delle linee guida per la digitalizzazione del patrimonio culturale delle minoranze etniche in Europa. In collaborazione con la comunità Sámi, ladina ed ebraica, grazie ad un approccio altamente multidisciplinare, il progetto mira a costruire una metodologia per affrontare un processo di transizione digitale dei beni culturali che si confronti con il tema dell'etica e della sostenibilità.
Parole chiave: Etica; digitalizzazione; patrimonio culturale; minoranze etniche; design

ABSTRACT (ENGLISH)
The digital cultural heritage of ethnic minorities: the DIGICHer project among the Sámi, Ladin and Jewish communities in Europe.
This contribution presents the objectives and methodological development of the European project DIGICHer, dedicated to revising the guidelines for the digitalization of the cultural heritage of ethnic minorities in Europe. In collaboration with the Sámi, Ladin, and Jewish communities, through a highly multidisciplinary approach, the project aims to build a methodology to address a process of digital transition of cultural heritage that confronts the issues of ethics and sustainability.*Paper Title for AIUCD2025.* Translation of the abstract in English language.
Keywords: Ethics; Digitalization; Cultural Heritage; Ethnic minorities; Design

1. IL PROGETTO DIGICHER
Il contributo intende presentare obiettivi, metodologia di lavoro e primi risultati della ricerca condotta dall'Istituto Italiano di Studi Germanici (IISG) all'interno del progetto europeo denominato "DIGICHer: digitisation of cultural heritage of minority communities for equity and renewed engagement", in corso di svolgimento. Il progetto, avviato nel febbraio del 2024, è risultato tra i vincitori del bando Horizon EU CL2 2023 - Heritage 01-03 (Re-visiting the digitisation of cultural heritage: What, how and why?), destinato a proposte che avessero come obiettivo la revisione delle attuali pratiche nella digitalizzazione del patrimonio culturale promosse dall'Unione Europea, alla luce dei considerevoli investimenti che quest'ultima sta impiegando per la valorizzazione in digitale di collezioni museali, biblioteche, archivi, monumenti, edifici, siti archeologici. I progetti selezionati affrontano in primo luogo le carenze in questo ambito specifico, partendo da una analisi critica degli errori e dei rischi – ma anche delle potenzialità – nella digitalizzazione del patrimonio culturale, allo scopo di riformulare tali processi di transizione digitale per garantirne accessibilità e sostenibilità. L'output effettivo sarà l'elaborazione di raccomandazioni e linee guida metodologiche applicabili nel futuro prossimo, indirizzate alle istituzioni europee attive nella conservazione e valorizzazione dei beni culturali.
Un consorzio internazionale di undici enti di ricerca e istituzioni culturali, di cui fa parte IISG, ha saputo intercettare la richiesta europea declinandola in un progetto, DIGICHer, che si propone di affrontare e superare l'insorgere di possibili questioni di natura etica nel caso in cui i beni interessati dalla digitalizzazione fossero espressione e testimonianza della cultura e della storia di un gruppo di minoranza etnico esistente all'interno dello spazio europeo. In tal senso, al fine di svolgere una ricerca che comprendesse una fase di sperimentazione e applicazione di un nuovo 'metodo', sono state coinvolte tre comunità di minoranza presenti in Europa, tra quelle particolarmente attive e sensibili al tema dei beni culturali digitali: le comunità Sámi stanziate nella Lapponia finlandese (Sápmi), la comunità ladina della Val di Fassa in Trentino e la comunità ebraica della città di Vilnius, in Lituania.
La ricerca prende avvio da una considerazione generale sugli attuali cambiamenti che sta apportando la transizione digitale nel settore dei beni culturali: a fronte degli innegabili vantaggi per tutela, conservazione e valorizzazione offerti dalle attuali risorse tecnologiche, sono evidenti alcuni ritardi con cui altri rami del sapere e della conoscenza si adeguano all'esponenziale evolversi delle soluzioni per la digitalizzazione (Zhelev 2021). La repentina corsa alla realizzazione di 'gemelli digitali' di ciò che

appartiene al patrimonio culturale, materiale ed immateriale, si accompagna quindi ad una serie di rischi talvolta imprevedibili o dai contorni non meglio definiti, che non hanno trovato ancora una strategia di contrasto poiché intersecano su vari aspetti della società. D'altro canto la diffusione e l'uso del bene culturale digitale intercetta molti interessi e settori: non solo quello della ricerca accademica ma anche quello educativo e della didattica ad esempio; quello economico, se si pensa alle varie possibilità di sfruttamento e di generare indotto; quello tecnologico, là dove si sperimentano nuove forme di fruizione sempre più avanzate e creative; quello giuridico, nel necessario aggiornamento delle norme sulla tutela dei dati e del copyright.

Rischi ed ostacoli risultano ancora più evidenti quando la digitalizzazione va a trattare un patrimonio culturale che ha una forte valenza identitaria: soprattutto, si pongono quesiti di natura etica sulle modalità del suo sfruttamento (Fouseki & Vacharopoulou 2013; Manžuch 2017). La mancanza di metodologia nel condurre la transizione digitale, sommata ad una regolamentazione insufficiente in materia, rischia di tradursi in una fruizione del bene, da parte dell'utente finale, totalmente avulsa dal contesto culturale di appartenenza. Ovvero, in una distorsione di ciò che quell'opera, contenuto, azione o soggetto riprodotto in digitale rappresenta nella propria realtà d'origine (Leopold 2013). Un uso ai fini commerciali ad esempio, senza un accordo con la minoranza, ha già condotto in passato a riproposizioni fuorvianti di alcuni tratti culturali rappresentati nelle opere e nelle immagini. A ciò si aggiunge l'eventualità che la comunità di riferimento non abbia pieno accesso ai beni digitali o che la tecnologia su cui si basa la diffusione non sia da essa fruibile, per cui la comunità stessa potrebbe perdere traccia dello sfruttamento di quelle risorse e non essere nella condizione di esercitare una qualche forma di controllo.

In seguito ad una preliminare valutazione dell'attuale panorama europeo relativo a progetti, attivi o conclusi, in fatto di beni culturali digitali legati alle minoranze, il programma DIGICHer mira quindi a riconsiderare, nell'ottica di questi molteplici aspetti, l'intera metodologia di sviluppo di iniziative future, dall'individuazione e studio dei beni culturali fino al prodotto digitale finale.

La prima fase di ricerca si è concretizzata in un approccio diretto con le tre comunità di minoranza coinvolte, al fine di raccogliere le criticità sollevate nei confronti dei progetti di digitalizzazione svolti in passato. Dagli incontri tenuti a Rovaniemi, Inari, Vilnius e Vigo di Fassa sono emerse differenti problematiche riguardanti l'efficacia o la percezione delle iniziative di digitalizzazione svolte negli ultimi anni. Nonostante la diversità e la distanza geografica che separa le tre minoranze, le criticità evidenziate sono state sorprendentemente trasversali a tutte e possono riassumersi in tre aree macroscopiche.

La prima riguarda la scarsità di risorse umane adatte a condurre la digitalizzazione, che è risultata in diversi casi sommaria o incompleta. Soprattutto, è stata evidenziata la mancanza di descrizioni scientifiche adeguate a corredo di beni e documenti, e la carenza di competenze tecniche specifiche nella metadatazione. Inoltre, di frequente i sistemi di metadati non supportano l'uso delle lingue parlate dalle minoranze (si pensi alle lingue ladine o alle nove varianti delle lingue Sámi), o comunque non sono stati integrati con le giuste traduzioni. Si tratta di un limite che impedisce la corretta identificazione di determinati contenuti, tra cui oggetti, fenomeni, rituali, secondo il termine in uso nella minoranza linguistica, a volte intraducibile poiché inesistente in altre lingue (Conway 2009; Conway 2010).

La seconda fa riferimento all'ambito giuridico, e quindi alle legislazioni vigenti in materia di protezione e diritti di utilizzo degli *asset* digitali. Il quadro è assai vario e complesso, poiché dipende da differenti fonti del diritto, a più livelli e con specificità nazionali, che interessano la tutela del patrimonio culturale, la proprietà intellettuale, il copyright, la protezione dei dati, la riproduzione di opere di pubblico dominio, le linee guida etiche nella ricerca. I rappresentanti culturali delle comunità hanno riportato due ordini di difficoltà. Da un lato, sono stati segnalati problemi di violazione della privacy nella pubblicazione e diffusione di documenti e immagini digitali, o altresì nel loro sfruttamento senza logica di contesto, fuori dal consenso della minoranza. Un primo significativo passo per contrastare questo rischio è stato ottenuto con il progetto Nuohtti ('rete da pesca' in Sámi), che unifica molti archivi digitali legati ai Sámi in un portale web ad accesso libero, tuttavia regolato da dettagliate linee guida etiche per l'utilizzo dei contenuti.[1] In prospettiva opposta, sono emersi problemi di accessibilità da parte delle comunità stesse al proprio patrimonio culturale digitale, a causa di restrizioni sul copyright e per la mancanza di diritti sulla proprietà intellettuale. Non raramente ciò è generato dal fatto che i progetti di digitalizzazione sono attivati in una visione *top-down*, cioè ideati e promossi da enti esterni e sovraordinati alla comunità (ad esempio su iniziativa di università piuttosto che dal governo nazionale o regionale) che ne limitano poi l'accesso. E' questo il caso, ad esempio, di documentazioni audio e video di interesse etnografico raccolte negli anni da

[1] https://nuohtti.com

vari enti, sia a scopo di ricerca sia per produzioni cinematografiche, i quali se ne riservano i diritti di utilizzo, limitando la minoranza nell'usufrutto della loro stessa rappresentazione.
La terza area in cui sono state sollevate criticità è la sostenibilità dei progetti nel tempo. La mancanza di risorse finanziare e quindi di un adeguato e costante aggiornamento delle tecnologie e delle ricerche, hanno trasformato alcuni database online in luoghi obsoleti nei contenuti e nelle interfacce per l'utente. Nuovamente, si tratta talvolta di progetti promossi da enti esterni che, terminato l'interesse di ricerca, non sono stati in grado di assicurare il mantenimento delle piattaforme online. La digitalizzazione dei beni culturali risente anche in questi casi di una discrepanza tra l'emissione di sempre più numerosi progetti e bandi di ricerca che richiedono una qualche forma di digitalizzazione e le corrispettive risorse messe in campo per assicurarne la vita futura una volta cessato l'interesse di ricerca. Anche la sostenibilità di progetti simili pare dipendere in questi casi da una mancata partecipazione della comunità ai processi decisionali.
In sintesi, la soluzione alle criticità elencate è sostanzialmente quella di trovare un modo per 'riportare' queste collezioni digitali alla comunità di appartenenza, affinché possa partecipare alla loro realizzazione, averne la gestione e il controllo sullo sfruttamento, regolarne eventualmente la diffusione, garantirne l'accessibilità evitando barriere tecnologiche, riutilizzarne i contenuti per generare nuove forme di valorizzazione. Così, la comunità dovrebbe anche essere messa in condizione di implementarle e partecipare al loro sostentamento nel futuro. In ultima analisi, ogni problema pare riconducibile essenzialmente ad un'unica questione di fondo: il mancato coinvolgimento della comunità di riferimento, dal momento dell'ideazione alla conclusione dei progetti. Ciò non va confuso con una mera assenza di comunicazione: non sono mancate in passato campagne di informazione capillare delle iniziative in atto presso le minoranze che tuttavia, per quanto efficaci, si inseriscono ancora nell'ottica dell'approccio *top-down* e non sono finalizzate a una partecipazione attiva della popolazione. A sua volta, quindi, possiamo invece interpretare questo mancato coinvolgimento attivo come una più estesa carenza nella progettazione.

2. SÁMI, LADINI, COMUNITA' EBRAICHE: DIGITALIZZAZIONE DEL PATRIMONIO CULTURALE

Risulta evidente che le problematiche da affrontare incrociano differenti aspetti e dinamiche che vertono e insistono sulla transizione digitale dei beni culturali, con impatti diversi sui risultati finali. Per fronteggiare un panorama di sfide di tale complessità, ciascun indirizzo di ricerca all'interno di DIGICHer è stato assegnato ad un ente leader, a seconda della competenza specifica dei singoli istituti facenti parte del consorzio (Università di Jena per la mappatura dell'attuale situazione europea sulla digitalizzazione e per la ricerca sulle potenzialità e le barriere tecnologiche; Università di Vilnius per la ricerca nel settore socio-economico; Università della Lapponia per la ricerca sugli aspetti legislativi e giuridici; Europeana per la diffusione e comunicazione).
Il gruppo di ricerca dell'Istituto Italiano di Studi Germanici è a capo della fase di sintesi e sperimentazione, riassumibile in due task: in primo luogo la progettazione di un quadro di lavoro (*framework*), formato da una serie ordinata di metodi e buone pratiche per affrontare la digitalizzazione dei beni culturali delle minoranze. Tali metodi sono formulati basandosi su tutte le considerazioni emerse nella valutazione dei limiti e dei rischi dell'attuale situazione e sono quindi rivolti al superamento e risoluzione degli stessi. Il *framework* è qui inteso come un modello di progettazione capace di unificare, in senso olistico, le soluzioni proposte dagli altri partner per affrontare i problemi emersi nei rispettivi ambiti di loro competenza, all'esperienza maturata nel confronto diretto con le minoranze. Esso dovrà essere replicabile su comunità diverse per dimensione e cultura e quindi scalabile, per esser proposto all'interno delle nuove linee guida europee che andranno a formarsi.
Il secondo task è costituito dal lavoro 'sul campo' e consiste nella realizzazione, presso le già menzionate comunità, di almeno un progetto di digitalizzazione per ciascuna, seguendo i principi del nuovo *framework*, in stretta collaborazione con i rispettivi organi di rappresentanza culturale (Parlamento Sámi finlandese e Archivio Nazionale Finlandese per i Sámi, il Jewish Heritage Network per le comunità ebraiche, l'Istitut Ladin Majon de Fascen per la comunità ladina). A tal proposito, in cooperazione con le rappresentanze culturali, sono state avanzate diverse proposte su materiali che saranno oggetto della fase di digitalizzazione. Presso la comunità Sámi del villaggio di Inari si è rivelato di particolare interesse l'archivio storico della corporazione di gestione comune delle acque e dei pescatori (*vesialueiden osakaskunta*), ricco di informazioni e testimonianze sulle tecniche tradizionali di pesca dei Sámi. Dato che la regione geografica della Lapponia – terra oltre il circolo polare artico abitata dai Sámi sin dall'età del bronzo finale – si

estende su un territorio che comprende Norvegia, Svezia, Finlandia e Russia (penisola di Kola), gli archivi e le collezioni contenenti materiali legati ai Sámi (*Sámi archives*) sono dislocati in più sedi e sparsi su tre nazioni diverse, gestiti rispettivamente dagli Archivi Nazionali finlandesi, svedesi e norvegesi. Parte dei *Sámi archives* sono virtualmente riunificati in rete grazie al già menzionato progetto sovranazionale Nuohtti. Altri fondi considerati significativi sono quelli delle associazioni Sámi per la promozione della cultura indigena (conservati a Oulu, a Rovaniemi e in altre località) e l'archivio personale dell'antropologo Tim Ingold, studioso dei Sámi Skólt e del loro rapporto con l'ambiente naturale, che ha lasciato un prezioso fondo di scritti e appunti personali redatti nei primi anni '70 durante la ricerca sul campo presso il lago Sevettijärvi.
Per quanto riguarda la comunità ladina della Val di Fassa, altrettanto varia è stata la selezione di proposte per la digitalizzazione, a cominciare dal fondo fotografico di Franz Dantone "Pascalin", costituito da più di mille lastre al collodio di grande valore etnografico, raffiguranti il paesaggio e gli abitanti ladini della valle alla fine del sec. XIX. Un secondo progetto di importanza etnografica riguarda il restauro conservativo e la riproduzione in modello tridimensionale di un apiario tradizionale, antica struttura di dislocazione delle arnie per la produzione del miele. L'azione intende documentare in digitale le diverse fasi del restauro, completando la documentazione con la descrizione in lingua ladina degli elementi costitutivi e della terminologia legata all'apicoltura, oltre che avviare una mappatura di altre strutture similari dislocate nella valle, al fine di mantenere e promuovere un'attività tradizionale per il territorio (Acke 2021; Malik 2021). Altre iniziative potrebbero riguardare l'aggiornamento e implementazione di realtà in digitale già esistenti, come la Mediateca Ladina o la piattaforma lessicografica-linguistica Tone, strumento basato su intelligenza artificiale per lo studio e la traduzione del ladino fassano.
Diverso e ancora più articolato è infine il panorama delle possibilità di digitalizzazione di beni culturali delle comunità ebraiche in Europa. Non configurandosi come una minoranza con dimensione territoriale definita, la presenza storica di comunità ebraiche è testimoniata, come noto, in molte città europee. Il Jewish Heritage Network ha collaborato alla nascita e alla gestione di numerosi programmi per la valorizzazione in digitale della cultura ebraica. In questo orizzonte, l'intervento di DIGICHer consiste nella riapertura di *repository* online ormai obsoleti o piuttosto terminati in seguito alla mancanza di adeguato aggiornamento e risorse per mantenerli in funzione. Tra questi, si segnala soprattutto il progetto Synagogen Internet Archive, corposa raccolta di documenti, informazioni e fotografie riguardanti più di 2200 sinagoghe esistenti in Germania e Austria dal medioevo al 1933. L'archivio digitale, ideato dall'Università di Darmstadt, è attualmente offline e i dati sono stati temporaneamente trasferiti allo Jüdisches Museum di Berlino.

3. COMMUNITY DESIGN, CO-CREATION E CITIZEN SCIENCE

Considerata la diversità delle minoranze e dei progetti proposti, quale strategia adottare per costruire un *framework* di lavoro, e quindi una metodologia, che sia scalabile, replicabile e possa adattarsi alle differenti realtà? Che sia in grado di considerare simultaneamente tutti gli aspetti della società su cui la digitalizzazione ha un impatto garantendo il rispetto di parametri etici? Che possa coinvolgere attivamente le comunità di minoranza? Se, come anticipato, gli elementi critici derivano perlopiù da una carenza nella capacità di progettare, per trovare una soluzione metodologica è stato dunque esaminato il problema dal punto di vista della progettazione, ovvero di *design* del sistema (Besana 2019). Dalla collaborazione tra l'Istituto Italiano di Studi Germanici e il Design Research Lab dell'Università di Trento è nata, di conseguenza, l'idea di una soluzione basata sul *design thinking*, sul *service design* e sul metodologie da essi derivate (Lupo 2023). Gli strumenti del design (*design tools*) consentono infatti di trovare soluzioni a problemi progettuali complessi che richiedono approcci multidisciplinari come, in questo caso, il far convogliare in un solo pacchetto di linee guida tutte le buone pratiche per guidare un processo di digitalizzazione, che contempli ogni ricaduta sociale ed economica, gli aspetti giuridici, i *driver* tecnologici e il rispetto dei fattori etici (Whaanga 2015). Non solo: una strategia di design viene applicata anche nel secondo task, per il test sul campo e il coinvolgimento attivo delle minoranze, e quindi per verificare e validare le buone pratiche adottate.
L'idea alla base è tanto lineare nella teoria quanto articolata nella realizzazione: si tratta di coinvolgere i membri della comunità instaurando un continuo dialogo, e al tempo stesso di attivare la comunità in senso pratico, tramite l'applicazione di design tools pensati allo scopo, affinché partecipi direttamente alla realizzazione dei progetti. Soprattutto, alcuni specifici metodi del design (ad esempio *co-creation*, *citizen science* e *community design*) sono utilizzati per raggiungere un efficace coinvolgimento delle comunità in tutto il percorso di progettazione e gestione della transizione digitale: dal confronto sulle questioni etiche

alla scelta dei fondi d'archivio o delle opere da digitalizzare, dal reperimento delle risorse economiche per sostenere i progetti in futuro (*crowdfunding* ed altre forme) all'effettiva partecipazione e condivisione nelle modalità di sfruttamento e di riutilizzo dei beni digitali nonché nelle ricadute economiche. In tal senso già si sono visti in tempi recenti alcuni esempi virtuosi di progetti che sono riusciti ad unire digitalizzazione e co-creazione con la comunità. E' questo il caso del progetto europeo ICH North, dedicato alla trasmissione del patrimonio immateriale della musica tradizionale nell'area chiamata 'Aurora' (regioni nord della penisola scandinava compresa la Lapponia), cui partecipano anche le comunità Sámi, particolarmente attive su vari fronti.[2] Sempre con la supervisione scientifica dei ricercatori per garantire la qualità del risultato, i membri della comunità sono stati coinvolti nell'integrazione dei metadati, completando le descrizioni con l'utilizzo delle categorie culturali loro proprie, e interpellati nella nomenclatura degli oggetti, rinominandoli secondo terminologia specifica della lingua Sámi, o ancora nelle traduzioni. Tali azioni ricadono nel più ampio spettro della cosiddetta *community-based research*. Alcuni portali online connessi al progetto, inoltre, si configurano come collezioni digitali aperte: i membri possono integrare nuove scoperte e nuove forme d'arte sviluppate all'interno della tradizione culturale o ancora rielaborare le tradizioni per reinterpretarle e attualizzarle. I database online diventano così non solo luoghi di conservazione e promozione, ma dei luoghi vivi che godono di un continuo rinnovarsi (Waterton 2010).
Fondamentale si dimostra quindi la capacità di coinvolgere diversi settori della società nella realizzazione del progetto e in particolare quelli più legati alla cultura e alla creatività, chiamando a raccolta esponenti ed esperti in varie discipline, mestieri e arti provenienti dalla comunità. DIGICHer si muove in definitiva su questo binario, considerando un programma di ricerca per la digitalizzazione come risultato, a tutti gli effetti, di un ecosistema di innovazione cui partecipano diversi attori. Il modello di innovazione cui si fa riferimento è definito 'quadrupla elica' (Carayannis & Campbell 2009). Se graficamente visualizzato, si tratta di un sistema in cui ogni pala dell'elica corrisponde ad un soggetto che partecipa al processo innovativo, in una costante interazione con gli altri. Quattro sono i soggetti coinvolti nella quadrupla elica: la sfera dell'università e dell'accademia, il governo, l'industria e la componente civile. Per lungo tempo la ricerca e l'innovazione sono state appiattite su due poli: la ricerca avviata dal mondo accademico o dalle istituzioni, mentre l'innovazione generalmente promossa dal settore delle imprese. Il modello di innovazione della quadrupla elica prevede invece una virtuosa e costante collaborazione multidisciplinare tra i quattro attori per il conseguimento di un obiettivo comune (*co-design*), soprattutto con l'inserimento della società civile (in questo caso attraverso il coinvolgimento di associazioni culturali). Altresì, il ruolo dell'industria nel settore dei beni culturali è, nella visione del progetto DIGICHer, svolto dalle imprese attive nel campo culturale e creativo. Queste ultime in particolare posso essere una risorsa decisiva per l'innovazione, considerando il loro impatto nell'ambito culturale: si tratta di soggetti che lavorano *nella* cultura, *con* la cultura e *per* la cultura. Rispettivamente, sono coloro che gestiscono direttamente collezioni o svolgono ricerca sul campo o ancora artisti e artigiani (lavorano *nella* cultura); coloro che da beni o collezioni preesistenti promuovono eventi e opere di valorizzazione e diffusione (lavorano *per* la cultura); coloro che svolgono la loro attività in ambiti tangenti (ad esempio l'innovazione tecnologica) ma che possono offrire ottime soluzioni per la digitalizzazione del patrimonio culturale (lavorano *con* la cultura). Un ulteriore risultato atteso dall'azione congiunta tra i diversi attori è l'individuazione di diverse fonti di risorse (sia umane che economiche, provenienti dalla rete che si viene a creare) per garantire la sostenibilità negli anni dei progetti di archivi e beni digitali avviati con DIGICHer, riportando così anche la gestione futura nelle mani delle minoranze coinvolte. In tal modo, si vogliono individuare non solamente nuove forme di interrelazione virtuosa con le comunità di minoranza di cui i beni culturali sono rappresentazione identitaria, ma si intende anche dare valore alle esperienze, alle competenze e alle risorse insite nella comunità stessa. Si trasformano così, in un risultato fruttuoso, il senso di responsabilità verso il proprio patrimonio culturale e il desiderio di partecipazione diretta della popolazione ad un più esteso progetto vòlto al rafforzamento, al riconoscimento e al rispetto della propria storia e tradizione.

[2] https://net.centria.fi/en/project/ich-north/

BIBLIOGRAFIA

Acke, L., De Vis, K., Verwulgen, S., & Verlinden, J. (2021). Survey and literature study to provide insights on the application of 3D technologies in objects conservation and restoration. Journal of Cultural Heritage, 49, 272–288.

Besana, D. (2019). Cultural Heritage Design: theories and methods for the project complexity management. EGE - Revista de Expresión Gráfica en la Edificación, 11, 31-43. https://doi.org/10.4995/ege.2019.12864.

Carayannis, E. & Campbell, D. (2009). 'Mode 3' and 'Quadruple Helix': Toward a 21st century fractal innovation ecosystem. International Journal of Technology Management, 46, 201-234.

Conway, P. (2009). Building Meaning in Digitized Photographs. Journal of the Chicago Colloquium on Digital Humanities and Computer Science, 1, 1–18.

Conway, P. (2010). Modes of Seeing: Digitized Photographic Archives and the Experienced User. American Archivist, 73, pp. 425-462.

Fouseki, K., & Vacharopoulou, K. (2013). Digital Museum Collections and Social Media: Ethical Considerations of Ownership and Use. Journal of Conservation and Museum Studies, 11, 1–10.

Leopold, R. (2013). Articulating Culturally Sensitive Knowledge Online: A Cherokee Case Study. Museum Anthropology Review, 7, 85–104.

Lupo, E., Camosino, G., Gobbo, B., Motta, M., Mauri, M., Parente, M., Spadoni, E., Spallazzo, D., & Rubino, F. (2023). Digital for Heritage and Museums: Design-Driven Changes and Challenges. Proceedings of the IASDR 2023: Life-Changing Design. Milan, 9-13 October, 2023. De Sainz Molestina, D., Galluzzo, L., Rizzo, F., Spallazzo, D. (eds.). https://doi.org/10.21606/iasdr.2023.397

Malik, U. S., Tissen, L. N., & Vermeeren, A. P., (2021). 3D Reproductions of Cultural Heritage Artefacts: Evaluation of signific ance and experience. Studies in Digital Heritage, 5, 1–29.

Manžuch, Z. (2017). Ethical issues in digitization of cultural heritage. Journal of Contemporary Archival Studies, 4, 1-17.

Mason, M., & Vavoula, G. (2021). Digital Cultural Heritage Design Practice: A Conceptual Framework. The Design Journal: An International Journal for All Aspects of Design, 24, 405-424. https://doi.org/10.1080/14606925.2021.1889738

Waterton, E. (2010). The advent of digital technologies and the idea of community. Museum Management and Curatorship, 25, 5–11.

Whaanga, H., Bainbridge, D., Anderson, M., Scrivener, K., Cader, P., Roa, T., & Keegan, T. T. (2015). He Matapihi Mā Mua, Mō Muri: The ethics, processes, and procedures associated with the digitization of indigenous knowledge - The Pei Jones collection. Cataloging & Classification Quarterly, 53, 520–547.

Zhelev, Y., Monova-Zheleva, M., & Stewart, R. (2021). The Digital Future of Intangible Cultural Heritage – Challenges and Initiatives. Digital Presentation and Preservation of Cultural and Scientific Heritage, 11, 41-49.

Fulfilling GEN-der AIms: do image-generating tools discriminate? An on-field study

Francesco Meledandri
University of Bari "Aldo Moro", Italy – francesco.meledandri@uniba.it

ABSTRACT (ENGLISH)

Recently, the digital turn entered a new era as some technological tools proved to represent a breakthrough in virtual and real-life practices. Though the idea of Artificial Intelligence (AI) has been circulating since the late 1950s (McCarthy et al., 1955) it has only been part of everyone's lives for a few years, since users had the possibility to create language and multimedia contents from scratch. Nowadays the most promising application is probably Generative AI (GenAI), which generates "synthetic data that closely resemble real-world data" (Bandi, 2023: 1). Trained upon LLMs (Large Language Models) developed by huge stakeholders (Dao, 2023), AI is now closer to the idea of "human reasoning" as it generates contents based on human-induced prompts. Ethical issues are paramount to the discussion of AI outputs (Dubber et al., 2020; Boddington, 2023). As a matter of fact, GenAI creates plausible outputs, though they are not always *true* since they rely upon the *synthetic* gathering of corpus-based data, thus leading to hallucinations (Ji et al., 2023). Another critical issue involves gender-related depictions, showing a disparity that leads to a proper machine-induced bias (Leavy, 2018; Foka, 2024). Against this background, this paper aims at providing evidence from specific AI-powered generative tools that create realistic images from textual prompts, thus carrying out an intersemiotic translation process (Dusi, 2015). In particular, providing different easy-to-use GenAI platforms (including the newly-introduced Grok featured in *Twitter/X*) with 'neutral' prompts, the analysis would assess the level of possible equality (or inequality) in terms of gender, thus assessing the level of potential bias that may influence the perception of users in terms of visual narratives (Chen et al., 2024).

Keywords: Artificial Intelligence; Equality in language; Gender issues; Intersemiotic translation.

ABSTRACT (ITALIANO)

Stabilire obiettivi di GENer-AI: gli strumenti AI di generazione immagini discriminano? Uno studio empirico

In tempi recenti la 'svolta digitale' ha abbracciato una nuova era grazie ad alcuni strumenti che rappresentano una rivoluzione nelle pratiche virtuali e reali. Sebbene l'idea di Intelligenza Artificiale (IA) sia presente sin dalla fine degli anni '50 (McCarthy et al., 1955), essa è diventata parte delle nostre vite solo da alcuni anni, soprattutto da quando è stato possibile creare contenuti linguistici e multimediali. Ad oggi l'applicativo più promettente in questo senso è la c.d. IA Generativa (GenAI), che "genera dati sintetici che assomigliano a dati reali" (Bandi, 2023: 1). L'IA è addestrata secondo gli LLM (Large Language Models) sviluppati da grandi attori del settore (Dao, 2023) ed è sempre più vicina al concetto di "ragionamento umano" poiché produce contenuti basati su indicazioni (prompt) generate da umani. Le questioni etiche sono al centro della discussione in tema di IA (Dubber et al., 2020; Boddington, 2023); infatti, la GenAI crea risultati plausibili sebbene non sempre *veri* poiché dipendono dal processo di associazione *sintetica* dei dati, e questo porta ai fenomeni di c.d. allucinazioni (Ji et al., 2023). Un altro problema riguarda le rappresentazioni basate sul genere, dato che vi è una disparità che porta ai cosiddetti *bias* della macchina (Leavy; 2018; Foka 2024). In questo contesto, questa presentazione mira a fornire alcuni dati empirici derivanti da strumenti specifici di IA generativa che creano immagini a partire da prompt testuali, in chiave di traduzione intersemiotica (Dusi, 2015). In particolare, attraverso l'immissione di prompt 'neutrali' su piattaforme di facile utilizzo (tra cui Grok, il nuovo strumento di IA su Twitter/X), il presente studio analizza il livello di possibile uguaglianza di genere, valutando il livello di *bias* che può influenzare la percezione degli utenti in termini di narrative visive (Chen et al., 2024).

Parole chiave: Intelligenza Artificiale; Uguaglianza linguistica; Questioni di genere; Traduzione intersemiotica.

1. INTRODUCTION

Over the last few years, the digital turn entered a new era. After achieving the goal to create an interconnected world via the popularisation of online, digitally-connected environments, thus overcoming time and space constraints for a globalised audience, other technological tools proved to represent a real

breakthrough in virtual and real-life practices. The contemporary 'elephant in the digital room' is surely represented by the advent of Artificial Intelligence (AI), a paradigm that has lately proved to be ubiquitous and has been shaping our activities in different environments and also in daily routines. Mehan (2024: 61) claims that "we take AI in its many forms for granted. It is so embedded in our daily lives that we see it practically everywhere" since it has been implemented in a growing number of devices and services.
The paradigm of AI, though, became a widespread reality only when the global community came across easy-to-use, usable tools available within everyone's reach. This allowed for a new, massive area of research that analysed the impact of AI in terms of new opportunities (Tomašev et al., 2020) but also potential shortcomings and dangers (Tredinnick & Laybats, 2023). The latter perspective seems to have particular implications, since it intertwines with the negative consequences of Human-Computer Interaction (HCI), a field of study that needs to re-instate new principles in the light of the AI era (Harper, 2019). The 'dark' side of AI is voiced as fear-mongering campaigns raised concerns about the possibility that this new technology could be 'too' smart and could be harmful to mankind, since "AI can be applied in ways that are detrimental" (Federspiel et al., 2023: 1), advocating for a wise understanding of the technology behind it (Carpio, 2023). Yet, there are still problems affecting AI and its universal reception due to issues that need to be analysed so that possible remedial actions could be taken. In the light of the above-mentioned considerations, this presentation aims at analysing a specific theme involving equality in AI-based contents, in particular those newly generated upon a human-induced input. In particular, the case study focuses on the analysis of data generated by AI systems with a definite output, in this case images. One of the opportunities provided by AI systems is to generate original, unpublished contents that result from pre-constructed, trained data. It is in the nature of such data that inequality may arise as users provide inputs (prompts) to generate this multimedia content, as this study tries to assess. Properly trained data should result in balanced outputs considering all variables involved in the input (prompt) entered for image generation-related purposes; yet, the generation of images may be affected by bias-based frameworks, thus providing inaccurate results. Starting from a supposedly neutral prompt, what is the degree of unbalanced representation in gender-related terms? Do these images reinforce certain pre-constructed images when it comes to the generation of a visual output from textual prompts? These research questions help in guiding a case study involving GenAI images with possible gender-dependent implications. A certain degree of unbalance in gender terms is somehow expected, even in cases in which GenAI systems could provide a number of alternatives for a given query.

2. SHIFTING PARADIGM(S): FROM AI TO GEN-AI FOR A GLOBALISED USE

The idea of non-human forms of intelligence dates back long before the current era dominated by AI. Actually, the prelude and pioneering conception has been circulating since the late 1950s (McCarthy et al., 1955). The phrase Artificial Intelligence has to be retraced in an era that tried to make the most of the potential offered by computers, since they represented the idea of "otherness" compared with the potential of human brains. Advances in AI always progressed, and AI-based systems permeated into our daily lives softly in terms of automation or predictive behaviours via electronic devices. At the same time, the 'new' era of AI has been part of everyone's lives for a few years now. Indeed, machines were able to go beyond 'pure' computing, giving users the possibility to create language and multimedia contents from scratch, in a sort of 'new linguistic turn' that accounts for our digital and always-on needs, and considers machines as linguistically intelligent entities (Roncaglia, 2023). The latest paradigm in AI, then, became part of our daily lives also in linguistic terms, since the creation of new language content (in different sign systems) is envisaged. This is why the term Generative AI (GenAI) barged through our customary activities almost relentlessly as a promising way to provide innovative forms of HCI. In IT terms, GenAI "focuses on developing algorithms and models capable of generating synthetic data that closely resemble real-world data" (Bandi, 2023, p.1). In order to create contents from scratch, data is trained via Large Language Models (LLMs), which are developed by major stakeholder in the field of *big data* (e.g. OpenAI, Microsoft, and Google; Dao, 2023). These systems are able to provide contents on the basis of human-induced texts (prompts), thus 'answering' users' requests and getting closer to the idea of human reasoning.
While most of the literature focussed on the earliest forms of GenAI solutions based on text-to-text generation (such as ChatGPT; see Mohsin & Masood, 2023; Watters & Lemanski, 2023), an underestimated though promising scenario is represented by image-based generative AI, also known as text-to-image (TTI) AI. Even in this case pictures are created from scratch and are the result of a process of synthesis of pre-existing data, though the final output only depends on the dataset and the training process, resulting in an image that reflects the prompt that has been provided. These images are

generated since large datasets of existing, semantically-based (i.e. described in order to associate a meaning to a given image) pictures are interconnected when elicited from a text prompt, using different synthesis techniques (Cerulli, 2023). Consequently, a textual prompt transformed into an image is a process of intersemiotic translation *à la Jakobson* (1959) since a change of sign system is envisaged (Dusi, 2015). These systems, then, rely on the importance of a well-defined prompt: the more detailed the input, the more realistic and close to the user's needs the pictorial result.

3. TOWARDS AI-NEQUALITY: ETHICS, BIAS, HALLUCINATIONS

When it comes to the analysis of AI-based outputs, the assessment should envisage a multi-layered interpretation. The most obvious classification depends on parameters such as consistency, accuracy and logical response (De Cesare, 2023; Taecharungroj, 2023; Mao et al., 2024). In this sense, then, this layer is more user-centred since the latter is interested in getting a good result from the machine, whether it is a text, or an image, or a video matching a prompt as much as possible. Other layers of assessment, instead, evaluate AI outputs on the basis of other parameters. Ethical issues are paramount to the discussion of AI outputs, and the debate is still open (Dubber et al., 2020; Boddington, 2023). The notion of truth (even in scientific domains; see Alkaissi & McFarlane, 2023) is questioned, since there is a subtle margin between plausible outputs and proper facts. Since they rely upon the *synthetic* gathering of data stored in huge corpora, the association of ideas may be inaccurate though perfectly credible and consistent with a desired prompt. This leads to the phenomenon known as *hallucinations*, or fabricated data that could seem truthful but it is not (Salvagno et al., 2023; Azamfirei et al., 2023), and this poses a threat in Natural Language Generation outputs (Ji et al., 2023). Another issue is represented by biased information, or "the intentional or unintentional introduction of systematic error into sampling or testing by, intentionally or not, selecting or encouraging one outcome or answer over others" (Gichoya et al., 2023: 1). Though the issue has been addressed to provide mitigation (Nazer et al., 2023), it has been a long-standing problem affecting machine learning and the relationship between humans and machines (Christian, 2020). This 'digital dilemma' highly depends on the kind of information machines need to process. In other words, bias is a(n unconscious) human condition (Emberton, 2021), therefore it is somehow transferred into the algorithms that make up AI systems. After all, "bias is as old as human civilization" (Ntoutsi et al., 2020: 2) and a possible solution needs understanding, mitigating, and accounting for bias. Standards aiming at its identification and management are necessary (Roselli et al., 2019; Schwartz et al., 2022); yet, as AI systems evolve and become popular tools, the risk to spread unequal results is high, thus triggering a vicious cycle that starts with human bias, feeds machines and generates results that are acknowledged non-objectively. In terms of biased and unbalanced representations, gender and race are two major 'battlegrounds'. Studies on the detection of racism in AI (Intachomphoo & Gundersen, 2020) have been flourishing since training artificial intelligences on larger data sets seems to strengthen their race-based biases (Hsu, 2023), questioning the idea that data to be trained is processed "with high precision, speed, and supposed lack of bias" (Adib-Moghaddam 2023: 4). Gender issues are another topic of interest in the field of AI, as the representation of uneven images in the description of groupings of men vs. women has been significantly noted (Leavy, 2018). A recent study by Chen et al. (2024) analysed a dataset of images processed by AI systems in news outlets and public resources, finding out that the resulting visual framing in terms of gender reinforces stereotypes and a consequent underrepresentation. Buolamwini and Gebru (2018) combine both gender and race to assess the level of disparity in AI-driven recognition systems; in particular, a well-defined classification of both categories is needed in order to assess the level of inequalities perpetrated by AI tools (in this sense, gender is labelled according to the binary classification corresponding to male vs. female). These analyses lay the foundations for a related analysis involving AI and the way through which gender is depicted, especially in the field of TTI technologies.

4. ASSESSING GENDER IN TTI GEN-AI: AN ON-FIELD STUDY

Since machines are built upon human schemes, AI embraced multimodality both in terms of learning and outputs; after all, "communication between humans is multimodal" (Xiao 2024) so does AI in replicating human tasks, fulfilling their needs according different media. For this reason, TTI technologies have been on the rise in terms of use, also due to its accessibility and usability. As stated above, a more widespread use leads to a possible diffusion of non-objective or unequal images, thus instilling a framing that does not account for a comprehensive view of diversities in a fair perspective. Gender mismatches are a consolidated evidence in TTI AI tools (Gorska & Jemielniak, 2023; Locke & Hodgdon, 2024; Foka, 2024)

and some projects aim at providing possible solutions to overcome bias-based image generation (Chinchure et al. 2025). This sample study tries to verify if such assumptions are true and can be verified via a hands-on approach by using prompts to generate TTI AI images. The methodology, then, relies on the use of some simple prompts that in a way could imitate the standard input of an average end-user who tries these easy-to-use AI tools. The platforms and apps used are consistent with the methodology described above, since some free tools are more likely to be used by a larger number of users, in a sort of digital paradigm of the so-called 'Zero Price Effect' that makes people "value free things too much" (Shampanier et al., 2007: 743). In this perspective, two platforms have been chosen. The first one is *Grok*[1], a newly-introduced built-in feature of *X* (formerly known as Twitter) which could represent a good benchmark to observe the level of synthetic generation of images it may achieve. Grok is available within the X mobile or web app, therefore users can choose to use this AI assistant to fulfil several tasks, including the generation of images from a textual prompt. The second app chosen is *GenCraft*[2], available as a web platform or mobile app, and on the digital market since 2023. Both interfaces (desktop and mobile) are usable and intuitive, and solely rely on the input of a prompt (also via a prompt enhancer) to create images with possible visual alternatives, thus being TTI tools only. This app envisages a *freemium* option, that is it allows for free-to-use (though limited) features and a subscription plan to unlock all available features.

As stated before, the prompts chosen are really simple and somehow neutral. Following Gorska & Jemielniak's (2023) work, the input on both platforms simply designate a professional figure without any other significant details (Query 1). Unlike other languages (e.g. Romance languages such as French or Italian), English has no gender-specific variation, therefore a simple indication of a job would generate some results based on the 'raw' idea synthetized by Grok and Gencraft. The professional figures to be included in the list, indicated as gender-neutral job titles[3], are: *a doctor, a lawyer, a teacher, an engineer, a scientist, a manager, a fashion model, a flight attendant, a coal miner,* and *a football athlete*.

Afterwards, these results will be compared to a slightly different prompt including a quality in the form of a pre-modifier that somehow matches the professional figured listed above, with no reference to gender whatsoever (Query 2). Thus, results can be compared to find out whether a slight variation in the prompt could provide alternative results as for gender equality. Results for the first, 'raw' query for the output on Grok are exemplified in Table 1:

Professional figure	Number of images	Male professional figures	Female professional figures
Doctor	4	4	0
Lawyer	4	4	0
Teacher	4	0	4
Engineer	4	4	0
Scientist	4	4	0
Manager	4	4	0
Fashion Model	4	0	4
Flight attendant	4	0	4
Coal miner	4	4	0
Football athlete	4	4	0
Total	40 (100%)	28 (70%)	12 (30%)

Table 1. Results of TTI query 1 in Grok

Results for the first, 'raw' query for the output on Gencraft are exemplified in Table 2:

Professional figure	Number of images	Male professional figures	Female professional figures
Doctor	4	4	0
Lawyer	4	4	0
Teacher	4	0	4
Engineer	4	4	0
Scientist	4	4	0
Manager	4	4	0
Fashion Model	4	0	4
Flight attendant	4	0	4
Coal miner	4	4	0
Football athlete	4	4	0
Total	40 (100%)	28 (70%)	12 (30%)

Table 2. Results of TTI query 1 in Gencraft

[1] Grok, available at: https://x.ai/ (cons. 24/01/2025)
[2] Gencraft, available at: https://gencraft.com/ (cons. 24/01/2025)
[3] Butler, 2024, *Top 25 Gender-Neutral Job Titles*, available at: https://universalenglish.org/gender-neutral-job-titles/ (cons. 24/01/2025)

As stated, a second query has been carried out using the same professional figures, though an attribute has been added. Frequent attributes to the category have been chosen from a list of frequent collocations[4], and are supposed to be gender neutral, thus enhancing the query. The queries for the second TTI generation are: *qualified doctor, successful lawyer, good teacher, chartered engineer, brilliant scientist, assistant manager, catwalk fashion model, skilled flight attendant, surface coal miner, amateur football athlete.* The results for the enhanced queries submitted to Grok and Gencraft are the following (Table 3 and Table 4):

Professional figure	Number of images	Male professional figures	Female professional figures
Qualified Doctor	4	2	2
Successful Lawyer	4	4	0
Good Teacher	4	3	1
Chartered Engineer	4	4	0
Brilliant Scientist	4	4	0
Assistant Manager	4	3	1
Catwalk Fashion Model	4	0	4
Skilled Flight attendant	4	0	4
Surface Coal miner	4	4	0
Amateur Football athlete	4	4	0
Total	40 (100%)	28 (70%)	12 (30%)

Table 3. Results of TTI query 2 in Grok

Professional figure	Number of images	Male professional figures	Female professional figures
Qualified Doctor	4	4	0
Successful Lawyer	4	4	0
Good Teacher	4	3	1
Chartered Engineer	4	4	0
Brilliant Scientist	4	4	0
Assistant Manager	4	0	4
Catwalk Fashion Model	4	4	0
Skilled Flight attendant	4	0	4
Surface Coal miner	4	4	0
Amateur Football athlete	4	4	0
Total	40 (100%)	31 (77.5%)	9 (22.5%)

Table 4. Results of TTI query 2 in Gencraft

Average results are showed in Table 5, grouped per platform and type of query.

Platform/App used	Overall number of TTI	Average number of Male professional figures	Overall number of Female professional figures
Grok	80	56 (70%)	24 (30%)
Gencraft	80	59 (73.75%)	21 (26.25%)
Grok+Gencraft, query 1	80	56 (70%)	24 (30%)
Grok+Gencraft, query 2	80	59 (73.75%)	21 (26.25%)

Table 5. Aggregated data per platform and type of query

Though the output generated cannot be discussed in terms of faithfulness of results, the images show a clear disproportion in gender-based terms, as more than 7 images out of 10 generate a male professional figure. Generally speaking, there is a clear tendency to get a sort of 'binary' result (for a given category, all professional figures are male OR female), and only in one case (Query: *Qualified doctor* in Grok) a true equality is showed. In other cases, an enhanced query generates different results in favour of the other category compared with the 'raw' query (e.g., *a teacher* has four examples of female figures in both Grok and Gencraft; *a good teacher* has three examples of male figures in both Grok and Gencraft; *a manager* has four examples of male figures in both Grok and Gencraft; *an assistant manager* has three results of male characters and one female figure in Grok, while four female figures in Gencraft). Though it is a sample study, results confirm the insight from Gorska & Jemielniak's study, as "gender bias is a widespread issue in AI-generated images of professionals, with men being overrepresented and certain

[4] Free Collocation, available at: https://www.freecollocation.com/ (cons. 24/01/2025)

professions and AI image generators being especially susceptible to this bias" (Gorska & Jemielniak 2023: 4373). Stereotypes are also present in this binary representation (e.g. fashion models as female-only figures), leading to the idea that only due to these pre-constructed framing schemes results are not completely unbalanced in favour of a male-based depiction. The data in this study confirms that training data for AI results is not simple and depends on the quality and quantity of the information submitted into machine learning engines. Further studies can rely upon other variables (other platforms, other professional figures involved, enhanced queries, a larger sample, etc.) to verify if such inequality is conveyed alike. For the time being, TTI AI provides unbalanced and somehow discriminatory results that may influence the perception of users in terms of visual narratives (Chen et al., 2024). Only fair practices of data training can enhance the current situation, aiming at the development of an unbiased paradigm.

REFERENCES

Adib-Moghaddam, A. (2023). *Is Artificial Intelligence Racist? The Ethics of AI and the Future of Humanity*. London: Bloomsbury.

Alkaissi, H., McFarlane, S. I. (2023). Artificial Hallucinations in ChatGPT: Implications in Scientific Writing. *Cureus 15*(2): e35179. Doi:10.7759/cureus.35179

Azamfirei, R., Kudchadkar, S. R., & Fackler, J. (2023). Large language models and the perils of their hallucinations. *Critical Care 27*(120). Doi: 10.1186/s13054-023-04393-x

Bandi, A., Adapa P. V. S. R., & Kuchi Y. E. V. P. K. (2023). The Power of Generative AI: A Review of Requirements, Models, Input–Output Formats, Evaluation Metrics, and Challenges. *Future Internet, 15*(8). Doi: 10.3390/fi15080260

Boddington, P. (2023). *AI Ethics*. Singapore: Springer Nature.

Boulamwini, J., Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Proceedings of Machine Learning Research, 81*,1-15.

Carpio, E. T. (2023). Overcoming Fear, Uncertainty, and Doubt: Artificial Intelligence (AI) and the Value of Trust. *Cureus 15*(9):e45576. Doi:10.7759/cureus.45576

Cerulli, G. (2023). *Fundamentals of Supervised Machine Learning*. Cham: Springer.

Chen, Y., Zhai, Y., & Sun, S. (2024). The gendered lens of AI: examining news imagery across digital spaces. *Journal of Computer-Mediated Communication, 29*(1), zmad047. Doi: 10.1093/jcmc/zmad047

Chinchure, A., Shukla, P., Bhatt, G., Salij, K., Hosanagar, K., Sigal, L., & Turk, M. (2025). TIBET: Identifying and Evaluating Biases in Text-to-Image Generative Models. In A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, & G. Varol (Eds.), *Computer Vision – ECCV 2024. 18th European Conference Milan, Italy, September 29–October 4, 2024*, 429-446. Cham: Springer.

Christian, B. (2020). *The Alignment Problem. Machine Learning and Human Values*. New York and London: W. W. Norton & Company.

Dao, X. Q. (2023). Performance Comparison of Large Language Models on VNHSGE English Dataset: OpenAI ChatGPT, Microsoft Bing Chat, and Google Bard. *arXiv*:2307.02288. Doi:10.48550/arXiv.2307.02288

De Cesare, A. M. (2023). Assessing the quality of ChatGPT's generated output in light of human-written texts. A corpus study based on textual parameters. *CHIMERA. Romance Corpora and Linguistic Studies, 10*, 179-210.

Dubber, M. D., Pasquale, F., & Das, S. (Eds.). 2023. *The Oxford Handbook of Ethics of AI*. New York: Oxford University Press.

Dusi, N. (2015). Intersemiotic translation: Theories, problems, analysis. *Semiotica, 206*, 181-205.

Emberton, M. (2021). Unconscious Bias Is a Human Condition. *The Permanente Journal, 25*(2). Doi:10.7812/TPP/20.199

Federspiel, F., Mitchell, R., Asokan, A., Umana, C., & McCoy D. (2023). Threats by artificial intelligence to human health and human existence. *BMJ Global Health 8*:e010435. Doi:10.1136/bmjgh-2022-010435

Foka, A. (2024). She Works, He Works: A Curious Exploration of Gender Bias in AI-Generated Imagery. *ArXiv*:2407.18524. Doi:10.48550/arXiv.2407.18524

Gichoya, J. W., Thomas, K., Celi, L. A., Safdar, N., Banerjee, S., Banja, J. D., Seyyed-Kalantari, L., Trivedi, H., & Purkayastha, S. (2023). AI pitfalls and what not to do: mitigating bias in AI. *British Journal of Radiology, 96*(1150). Doi:10.1259/bjr.20230023.

Gorska, A. M., & Jemielniak, D. (2023). The invisible women: uncovering gender bias in AI-generated images of professionals. *Feminist Media Studies, 23*(8), 4370–4375.

Harper, R. H. R. (2019). The Role of HCI in the Age of AI. *International Journal of Human-Computer Interaction, 35*(15) 1331-1444.

Hsu, J. (2023). Big data may make AI more racist. *New Scientist, 259*(3448), 12.

Intachomphoo, C., & Gundersen, O. E. (2020). Artificial Intelligence and Race: a Systematic Review. *Legal Information Management, 20*, 74–84.

Jakobson, R. (1959). On linguistic aspects of translation. In R. Brower (Ed.), *On Translation*, 232–239. Cambridge, MA: Harvard University Press.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Chen, D., Chan, H. S., Dai, W., Madotto, A., & Fung, P. (2023). Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys, 55*(12), 248, 1-38.

Leavy, S. (2018). Gender Bias in Artificial Intelligence: The Need for Diversity and Gender Theory in Machine Learning. *2018 ACM/IEEE 1st International Workshop on Gender Equality in Software Engineering*.

Locke, L. G., & Hodgdon, G. (2024). Gender bias in visual generative artificial intelligence systems and the socialization of AI. *AI & Society*. Doi:10.1007/s00146-024-02129-1

Mao, R., Chen, G., Zhang, X., Guerin, F., & Cambria, E. (2024). GPTEval: A Survey on Assessments of ChatGPT and GPT-4. *LREC-COLING 2024*, 7844–7866.

McCarthy, B., Minsky, M. L., Rochester, N., & Shannon, C. E. (1955). *A proposal for the Dartmouth summer research project on artificial intelligence*.

Mehan, J. E. (2024). Artificial Intelligence. *Ethical, social, and security impacts for the present and the future*. Ely: IT Governance Publishing.

Mohsin, H., & Masood, S. (2023). Unveiling the Potential of ChatGPT: Applications, Challenges, and Future Directions. *International Journal of Computer Applications, 185*(21), 37-49.

Nazer, L. H., Zatarah, R., Waldrip, S., Xue Chen Ke, J., Moukheiber, M., Khanna, A. K., Hicklen, R. S., Moukheiber, L., Moukheiber, D., Ma, H., & Mathur, P. (2023). Bias in artificial intelligence algorithms and recommendations for mitigation. *PLOS Digital Health, 2*(6): e0000278. Doi:10.1371/journal.pdig.0000278

Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdl, W., Vidal, M. E., Ruggieri, S., Turini, F., Papadopoulos, S., Krasanakis, E., Kompatsiaris, I. Kinder-Kurlanda, K., Wagner, C., Karimi, F., Fernandez, M., Alani, H., Berendt, B., Kruegel, T., & Heinze, K. (2020). Bias in data-driven artificial intelligence systems—An introductory survey. *WIREs Data Mining and Knowledge Discovery, 10*(3). 10:e1356. Doi: 10.1002/widm.1356

Roncaglia, G. (2023). *L'architetto e l'oracolo: Forme digitali del sapere da Wikipedia a ChatGPT*. Bari-Roma: Laterza Editori.

Roselli, D., Matthews, J., & Talagala, N. (2019). Managing Bias in AI. *Proceedings of the 2019 World Wide Web Conference*. 978-1-4503-6675-5/19/05. Doi: 10.1145/3308560.3317590

Salvagno, M., Taccone, F. S., & Gerli, A. G. (2023). Artificial Intelligence Hallucinations. *Critical Care 27*(180). Doi:10.1186/s13054-023-04473-y

Schwartz, R., Vassilev, A., Greene, K., Perine, L., Burt, A., & Hall, P. (2022). Towards a Standard for Identifying and Managing Bias in Artificial Intelligence. *NIST Special Pubblication 1270*. Doi: 10.6028/NIST.SP.1270

Shampanier, K., Mazar, N., & Ariely, D. (2007). Zero as a Special Price: The True Value of Free Products. *Marketing Science, 26*(6), 742-757.

Taecharungroj, V. (2023). "What Can ChatGPT Do?" Analyzing Early Reactions to the Innovative AI Chatbot on Twitter. *Big Data Cogn. Comput. 7*(35). Doi:10.3390/bdcc7010035

Tomašev, N., Cornebise, J., Hutter, F., Mohamed, S., Picciariello, A., Connelly, B., Belgrave, D. C. M., Ezer, D., Cachat van der Haert, F., Mugisha, F., Abila, G., Arai, H., Almiraat, H., Proskurnia, J., Snyder, K., Otake- Matsuura, M., Othman, M., Glasmachers, T., & de Wever, W. (2020). AI for social good:

unlocking the opportunity for positive impact. *Nat Commun, 11*(2468). Doi:/10.1038/s41467-020-15871-z

Tredinnick, L., & Laybats, C. (2023). The dangers of generative artificial intelligence. *Business Information Review, 40*(2), 46-48.

Watters, C., & Lemanski, M. K., (2023). Universal skepticism of ChatGPT: a review of early literature on chat generative pre-trained transformer. *Frontiers in Big Data 6*:1224976. doi: 10.3389/fdata.2023.1224976

Xiao, H. (May 8, 2024). Multimodal: AI's new frontier. *MIT Technological Review Insights*. https://www.technologyreview.com/2024/05/08/1092009/multimodal-ais-new-frontier/

From Bias Paralysis to Bias as a Category of Analysis

Introducing the Bias-Aware Framework

Mrinalini Luthra[1], Amber Zijlma[2]
[1] Huygens Institute, Royal Dutch Academy of Arts and Sciences, Amsterdam, mrinalini.luthra@huygens.knaw.nl
[2] Huygens Institute, Royal Dutch Academy of Arts and Sciences, Amsterdam, amber.zijlma@huygens.knaw.nl

ABSTRACT (ENGLISH)
As digital humanities (DH) increasingly rely on computational methods—from handwritten text recognition to automated metadata extraction—addressing bias becomes both more urgent and more complex. Datasets inherit and perpetuate biases through multiple channels: discriminatory language in archives, unequal representation in collection practices, and algorithmic biases in AI-assisted processing. These biases are compounded throughout the research process, yet the term "bias" itself lacks a clear definition, often causing "bias paralysis." This paper introduces the Bias-Aware Framework, a research framework that transforms bias from a paralyzing problem into an analytical category with clear action points.

The framework consists of three key components: (1) a *Bias Thesaurus,* which establishes a shared vocabulary of 'bias' across disciplines to address its conceptual instability; (2) a *Data Lifecycle Model* showing where biases enter the research process; and (3) *Guidelines* for documenting, describing, and mitigating bias. Together, these components can transform abstract concerns about bias into concrete opportunities for methodological improvement. We therefore approach bias not simply as an error, but as a revealing analytical lens that shapes knowledge production. By explicitly describing these conditions of production, researchers can improve transparency, improve dataset documentation, and enable more informed reuse of their data.
Keywords: bias, data ethics, taxonomy, computational methods, digital humanities

ABSTRACT (ITALIANO)[1]
Titolo del paper: Dalla paralisi da pregiudizio al pregiudizio come categoria di analisi: Introduzione al Framework Bias-Aware

Poiché le scienze umane digitali (DH) si affidano sempre più a metodi computazionali—dal riconoscimento del testo manoscritto all'estrazione automatica dei metadati—affrontare i pregiudizi diventa sia più urgente che più complesso. I dataset ereditano e perpetuano pregiudizi attraverso molteplici canali: linguaggio discriminatorio negli archivi, rappresentazione diseguale nelle pratiche di raccolta e pregiudizi algoritmici nell'elaborazione assistita dall'intelligenza artificiale. Questi pregiudizi si amplificano durante tutto il processo di ricerca, mentre il termine "pregiudizio" stesso manca di una definizione chiara, causando spesso una "paralisi da pregiudizio." Questo articolo introduce il Framework Bias-Aware, un quadro di ricerca che trasforma il pregiudizio da problema paralizzante a categoria analitica con chiari punti d'azione.

Il framework consiste di tre componenti chiave: (1) un *Thesaurus dei Pregiudizi*, che stabilisce un vocabolario condiviso del 'pregiudizio' tra discipline per affrontare la sua instabilità concettuale; (2) un *Modello del Ciclo di Vita dei Dati* che mostra dove i pregiudizi entrano nel processo di ricerca; e (3) *Linee guida* per documentare, descrivere e mitigare i pregiudizi. Insieme, questi componenti possono trasformare preoccupazioni astratte sui pregiudizi in opportunità concrete per il miglioramento metodologico. Pertanto, affrontiamo il pregiudizio non semplicemente come un errore, ma come una lente analitica rivelatrice che modella la produzione di conoscenza. Descrivendo esplicitamente queste condizioni di produzione, i ricercatori possono migliorare la trasparenza, perfezionare la documentazione dei dataset e consentire un riutilizzo più consapevole dei loro dati.
Parole chiave: pregiudizio, etica dei dati, tassonomia, metodi computazionali, scienze umane digitali

[1] Please note that the translation of the Italian title and abstract at the start of this document were created using generative AI (DeepL), as neither of the authors is a speaker of Italian.

1. INTRODUCTION

The digital humanities (DH) face an increasingly critical challenge: initial enthusiasm for the seemingly infinite possibilities of computational methods has made way for a more critical stance towards their implementation in current DH research (Prescott, 2023). Within this shift, the question of bias has taken on a new urgency - not just as a theoretical concern, but as a practical challenge that shapes every aspect of research. This urgency stems from the recognition that unexamined biases in institutions, (training) data, metadata, and computational methods risk amplifying historical exclusions and further marginalising already underrepresented narratives. These concerns are compounded by the challenge of applying computational methods to historical materials that contain inherently biased perspectives. To illustrate, when using digital infrastructure based on historical sources, descendants of enslaved peoples are being forced to search for their ancestors in lists of 'commodities' or with the use of painful search terms; alternatively, individuals mentioned in sources are anonymised or erased through opaque indexing practices (Luthra et al., 2023). While bias has long been a concept of inquiry in the humanities (McCullagh, 2000; Trouillot, 1995), DH researchers are now having to carefully perform a balancing act, in which both historicity and social justice are valued and attended to in their research practices, while also addressing technical concerns about algorithmic fairness and data representation.

But, the term "bias" itself lacks a clear definition, causing 'bias paralysis' among researchers: after all, if everything is biased, what should we do? This paper proposes that instead of intending to fully 'remove' bias from research - impossible and, for humanities and social sciences research, unwanted - bias should be treated as a productive category of analysis for digital humanities research, a productive lens through which to further enhance research. For this, we have developed a Bias-Aware Framework, consisting of a Bias Thesaurus, a Data Lifecycle Model and Guidelines for bias identification, articulation and reduction. This contribution therefore aims to:

- **Deconstruct and Articulate Bias** in DH research through the explanation and implementation of the Bias-Aware Frameworks.
- **Encourage Ethical Intervention** for mitigating bias through the data life cycle.
- **Promote** bias analysis and mitigation as urgent and necessary inclusion to the (digital humanities') research agenda.

2. BACKGROUND

As computational methods are increasingly entangled with humanities research, the issues of bias take on renewed urgency. While AI and digital technologies enable increased access to historical sources, they also risk perpetuating historical and contemporary biases embedded in archival sources, metadata, tools and technology (Navigli et al., 2023; Thylstrup, 2019). DH researchers must grapple with multiple forms of bias that intersect across different disciplines and stages of research. A single DH project might simultaneously confront archival biases in source selection (Trouillot, 1995), historical power structures in interpretation (van Rossum, 2019), representational biases in digitisation (Kizhner et al., 2021), and algorithmic biases in computational analysis (Buolamwini & Gebru, 2018; Mehrabi et al., 2021; Søgaard et al., 2014; Suresh & Guttag, 2021).

What makes this challenge particularly concerning is that these diverse forms of bias don't simply exist in parallel—they compound and amplify each other throughout the research lifecycle. Historical biases in archival selection get encoded into metadata structures, which then inform algorithmic design, creating feedback loops that can dramatically magnify originally subtle imbalances. This cascading effect of bias—from historical sources through metadata design to algorithmic processing—creates a distinctive challenge that requires a systematic approach rather than isolated interventions. This compounding effect means that seemingly minor biases at early stages can result in significantly skewed outcomes during analysis, interpretation, and beyond.

Additionally, due to the ubiquitous use of the term 'bias', there is no consistent vocabulary to rely on when wishing to address the issue. Even within specific academic fields, the concept of bias proves elusive: Blodgett et al.'s (2020) analysis of 146 papers in the field of natural language processing reveals significant confusion in defining 'bias', while in digital cultural heritage, the characterization of offensive terminology as bias remains unclear.[2] Without a coherent transdisciplinary framework for understanding

[2] For instance the Words Matter (Modest & Lelijveld, 2018), a publication on sensitive words in the museum sector, doesn't use the term 'bias', but projects such as DE-BIAS (Masschelein et al., 2023), based at the Dutch Institute of Sound and Vision, use the term in context of developing a tool to identify harmful language in archives.

and addressing bias, there is a risk of either oversimplifying it or becoming stunned by it - what we term "bias paralysis".

Drawing inspiration from Joan Scott's (1986, 2010) concept of gender as an analytical category and Sherman et al.'s (2024) treatment of algorithmic absences, we argue that bias, like Foucault's concept of power (2008), is relational and dynamic, actively shaping and being shaped by social, cultural, and historical contexts. This perspective shifts our focus from attempting to "solve" bias - an impossible task - to using bias as a critical tool for reflection and analysis.

Unlike existing bias taxonomies in computer science and AI ethics that focus primarily on statistical and algorithmic biases (Blodgett et al., 2020; Navigli et al., 2023), our Bias-Aware Framework specifically addresses the unique challenges faced by digital humanities researchers who must simultaneously consider historical context, cultural sensitivity, and computational aspects. The Framework's distinctiveness lies in its recognition that for humanities scholars, bias is not merely a technical problem to be eliminated but also a historically significant phenomenon that requires careful documentation and contextualization. Where AI ethics frameworks typically aim to remove bias entirely, our approach acknowledges that humanistic inquiry necessitates preserving certain original biases as historical evidence while preventing their harmful amplification through computational methods.

There have been several valuable contributions towards practical mitigation of bias, including templates for transparent documentation and alternatives to harmful terminology (Bender & Friedman, 2018; Chilcott, 2022, 2022; Luthra & Eskevich, 2024; Masschelein et al., 2023; Modest & Lelijveld, 2018; Scheuerman et al., 2020). However, these approaches typically address specific manifestations of bias. DH researchers need tools that can help them address the full spectrum of biases they encounter, from historical biases to technical biases to representation biases. Moreover, they need to understand how these different forms of bias interact and transform across the stages of their work. Furthermore, there is the need to make this approach specific to DH: because unlike the gross of the strategies and taxonomies proposed within the data and computer science fields, humanists necessarily need to 'respect' the original form of their (historical) data, as its biases provide valuable insights. Reflexively supplementing the data or changing the source data is therefore out of the question - to humanists, it's precisely these imbalances that are significant to research, but not to be perpetuated through machine learning models. We believe that bias mitigation in DH currently faces the following fundamental questions as obstacles to identifying, articulating, and mitigating bias:

1. What exactly do we mean by "bias" in digital humanities research?
2. Where does bias occur in the dataset creation process?
3. How can researchers practically address bias with their available resources?

In order to address these challenges, we are developing a "Bias Aware Framework" which has the the three corresponding components:

1. **A Bias Thesaurus**: A comprehensive list of the concepts connected to bias (such as representation, offensive language, FAIR, CARE, silences, etc.) that creates a shared vocabulary for discussing bias across disciplines.
2. **A Bias-Aware Data Lifecycle Model**: Showing where and how bias manifests at different research stages, allowing for targeted interventions at critical points.
3. **Guidelines and Toolkit**: Reflective questions at each stage of the dataset lifecycle, illustrative examples, and "good-better-best" recommendations for bias analysis, description, and mitigation.

Our development of the Bias-Aware Framework is a response to the gap in the DH field, that "a set of guidelines is missing, a serious lack when one might want to think through ethical concerns" (O'Sullivan, 2024). It is an actionable framework that demystifies 'bias' and transforms it into a productive tool for improving current and future research and knowledge production. Our focus on the dataset specifically

3. METHODOLOGIES

Our development of Bias-Aware Framework has followed a three-phase iterative approach combining theoretical analysis with practical validation:

1. **Literature Review:** To gain a better overview of current theories about and strategies against bias, we systematically reviewed literature in the fields of archival studies (Modest & Lelijveld, 2018; Stoler, 2010; Trouillot, 1995), epistemology (Foucault, 2002) and computer sciences (Bender & Friedman, 2018; Gebru et al., 2021; Orr & Crawford, 2023; Scheuerman et al., 2020). These are fields in which (origins of) bias has received much attention. From here, we expanded out to include sources that were not academic and/or text-based, such as video's, art installations,

and fiction. The importance of including these resources was to critically confront our own biases for the written and academic. For each resource, we focused on what forms of bias or strategies to mitigate bias were present, compiling these in an open-to-access list of resources.[3]

2. **Insights from Partner Projects:** The Framework's development is guided by semi-structured interviews with stakeholders from four major digital infrastructure projects focusing on colonial and slavery archives: Slave Voyages[4], GLOBALISE[5], Exploring Slave Trade in Asia[6], and the Historical Database of Suriname and Caribbean[7]. These partnerships provide crucial insights into practical challenges and implementation requirements. We also conduct conversations with (external) advisors, based in different parts of the world and with expertise from a range of disciplines: cultural heritage, critical archival studies, community (memory) work, natural language processing, and the FAIR principles. As a result, the Framework is rooted in plurivocality (Sitzia, 2023). The experiences of our partners and advisors, both in terms of challenges and strategies to navigate these, form the foundation of the Framework.
3. **Framework Validation and Refinement:** We are validating and refining the Bias-Aware Framework through two parallel tracks: expert consultations and interactive workshops with digital humanities projects. Workshops have been organised that serve as practical testing grounds where participants apply the Framework's toolkit to analyse bias in their own datasets. This implementation phase aims to reveal the framework's strengths and limitations in practice and identify potential blind spots. Participant feedback and documented use cases will drive iterative improvements to the Framework's components, ensuring its broader applicability and effectiveness.

4. BIAS-AWARE FRAMEWORK

The Bias-Aware Framework consists of three elements:

1. **The Bias Thesaurus** maps the various *expressions of bias*—concrete forms bias takes in research practices, such as harmful language, uneven descriptive depth, or limiting categorization schemes. The Thesaurus creates a shared vocabulary across disciplines, visualises interconnections between different expressions of bias through network graphs, and provides researchers with a conceptual map for navigating bias-related concerns. For each expression of bias, the Thesaurus includes their definition, where they occur in the life cycle (therefore linked to the Lifecycle Model), what questions to consider regarding the specific expression throughout their research and good-better-best practices (therefore linked to the Guidelines). In this way, the Thesaurus is the theoretical grounding of our Framework (see Figure 1).
2. **Bias-Aware Dataset Lifecycle Model** (Figure 2) forms the structural backbone of our framework, grounding abstract bias considerations in familiar research workflows while addressing a gap in digital humanities methodology. Our model[8] identifies five key stages (Set Up, Collect, Process, Analyse, Preserve & Share) and maps how different expressions of bias defined in the Thesaurus manifest at each stage. A key insight from our research is the "stickiness" of certain bias expressions across multiple stages, though they manifest differently depending on the stage's focus. For example, representation concerns appear throughout the lifecycle: in Set Up, they relate to whose scholarship informs the project; in Collect, they concern whose perspectives are captured in the data; in Process, they involve how categories represent complex realities; and in Analyse, they address whose stories are highlighted in the subsequent story the data tells.

[3] Combatting Bias Resources List: https://docs.google.com/spreadsheets/d/17mAClY06JuPQm9qf3Z7ZmdYi08VrZboCUTOv27J6J-E/edit?usp=sharing
[4] Slave Voyages: https://www.slavevoyages.org/
[5] GLOBALISE: https://globalise.huygens.knaw.nl/
[6] Exploring Slave Trade in Asia (ESTA): https://esta.iisg.nl/
[7] Historische Database Suriname and Caribbean (HDSC): https://www.ru.nl/onderzoek/onderzoeksprojecten/historische-database-van-suriname-en-de-cariben
[8] See https://www.rd-alliance.org/wp-content/uploads/2024/09/D1_The-creation-of-a-harmonised-research-data-lifecycle-RDL-model-and-crosswalkpdf

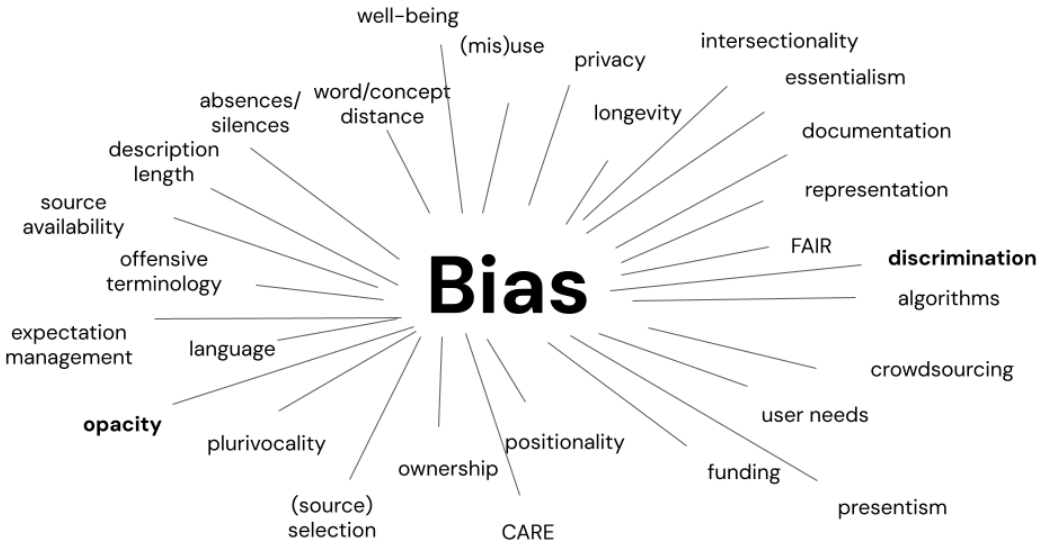


Figure 1: Bias as heuristic for distinct yet interrelated issues. These expressions of bias are explored in the Bias Thesaurus.

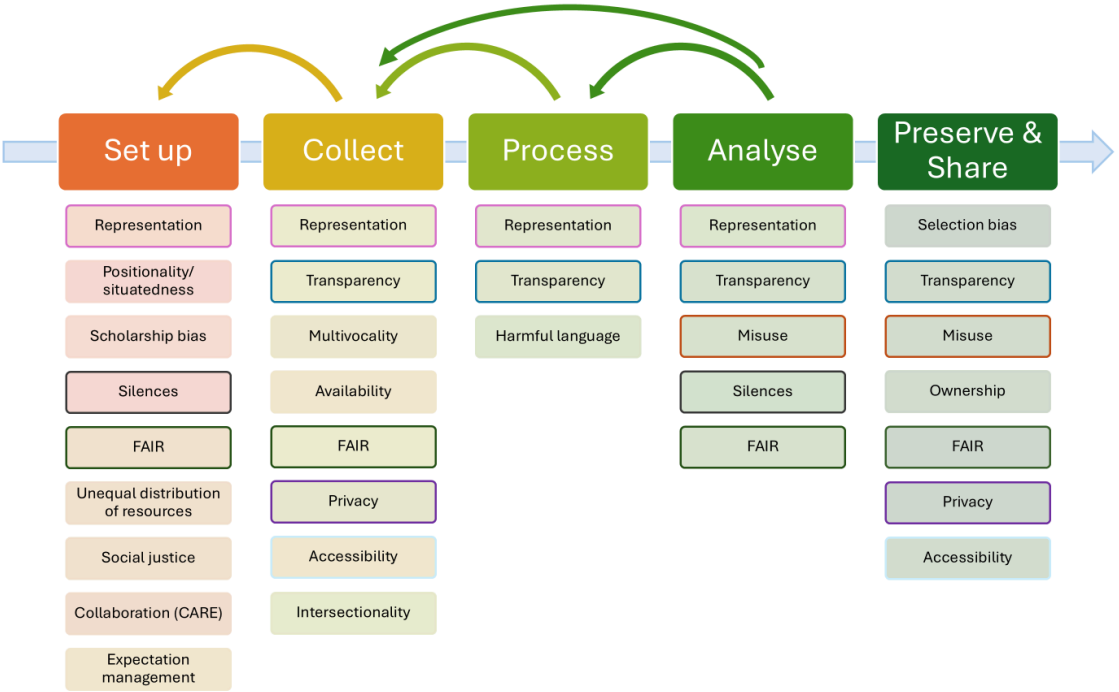


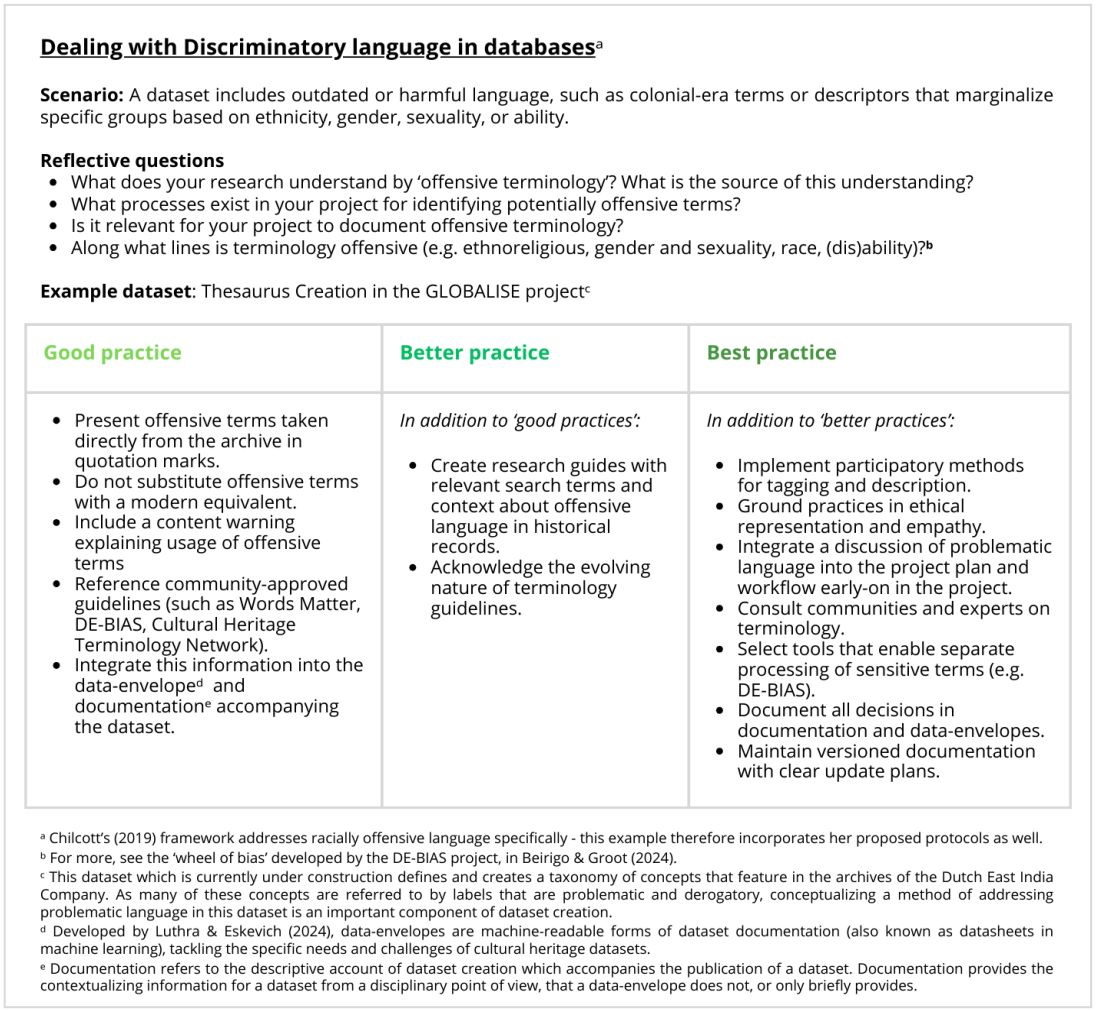
Figure 2: The Bias-Aware Lifecycle Model. The expressions of bias associated with each stage are contextualised in the Thesaurus and Guidelines.

3. **Guidelines and Toolkit:** The final component transforms theoretical understanding into practical action through structured guidelines for each stage of the dataset lifecycle. These guidelines provide reflective questions, curated resources, documentation templates, "good-better-best" recommendations[9] that accommodate varying resource constraints, and example strategies drawn from successful DH projects. Figure 4 illustrates our guideline approach for addressing archival silences, offering tiered intervention strategies from basic documentation to participatory community engagement. The guidelines emphasize that addressing bias is not an all-or-nothing proposition—even resource-constrained projects can implement basic bias-aware practices. This

[9] The good-better-best-framework has been humbly taken over from Alicia Chilcott (2022).

scaffolded approach helps prevent "bias paralysis" by making intervention accessible regardless of project scale or resources.

Essentially, the Framework compiles existing resources in the field - both published (literature) and unpublished (internal project strategies, collected through conversation) - to create a visual and practical toolkit on bias identification, articulation and reduction.



Dealing with Discriminatory language in databases[a]

Scenario: A dataset includes outdated or harmful language, such as colonial-era terms or descriptors that marginalize specific groups based on ethnicity, gender, sexuality, or ability.

Reflective questions

- What does your research understand by 'offensive terminology'? What is the source of this understanding?
- What processes exist in your project for identifying potentially offensive terms?
- Is it relevant for your project to document offensive terminology?
- Along what lines is terminology offensive (e.g. ethnoreligious, gender and sexuality, race, (dis)ability)?[b]

Example dataset: Thesaurus Creation in the GLOBALISE project[c]

Good practice	Better practice	Best practice
• Present offensive terms taken directly from the archive in quotation marks. • Do not substitute offensive terms with a modern equivalent. • Include a content warning explaining usage of offensive terms • Reference community-approved guidelines (such as Words Matter, DE-BIAS, Cultural Heritage Terminology Network). • Integrate this information into the data-envelope[d] and documentation[e] accompanying the dataset.	*In addition to 'good practices':* • Create research guides with relevant search terms and context about offensive language in historical records. • Acknowledge the evolving nature of terminology guidelines.	*In addition to 'better practices':* • Implement participatory methods for tagging and description. • Ground practices in ethical representation and empathy. • Integrate a discussion of problematic language into the project plan and workflow early-on in the project. • Consult communities and experts on terminology. • Select tools that enable separate processing of sensitive terms (e.g. DE-BIAS). • Document all decisions in documentation and data-envelopes. • Maintain versioned documentation with clear update plans.

[a] Chilcott's (2019) framework addresses racially offensive language specifically - this example therefore incorporates her proposed protocols as well.
[b] For more, see the 'wheel of bias' developed by the DE-BIAS project, in Beirigo & Groot (2024).
[c] This dataset which is currently under construction defines and creates a taxonomy of concepts that feature in the archives of the Dutch East India Company. As many of these concepts are referred to by labels that are problematic and derogatory, conceptualizing a method of addressing problematic language in this dataset is an important component of dataset creation.
[d] Developed by Luthra & Eskevich (2024), data-envelopes are machine-readable forms of dataset documentation (also known as datasheets in machine learning), tackling the specific needs and challenges of cultural heritage datasets.
[e] Documentation refers to the descriptive account of dataset creation which accompanies the publication of a dataset. Documentation provides the contextualizing information for a dataset from a disciplinary point of view, that a data-envelope does not, or only briefly provides.

Figure 3: Example *Guideline* for dealing with discriminatory language in datasets.

5. ABOUT COMBATTING BIAS

Combatting Bias[10], is a one-year project focusing on the ethical creation of datasets for the social sciences and humanities. It particularly focuses on the dataset, because creators, users, and digital infrastructures intersect at this point, making it a good unit of analysis. CB's work is enriched by partnerships with leading digitisation initiatives and advisors from different geographies and disciplines such as museum studies, critical archival studies, ethnomusicology, history, computer science, and economics.

ACKNOWLEDGEMENTS

The authors of this paper are part of the *Combatting Bias* project, funded by the NWO via the Thematic Digital Competence Centre Social Sciences & Humanities (TDCC-SSH). We thank the anonymous reviewers for their critical feedback.

REFERENCES

Bender, E. M., & Friedman, B. (2018). Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics*, *6*, 587–604. https://doi.org/10.1162/tacl_a_00041

[10] https://combattingbias.huygens.knaw.nl/

Beirigo, Isabel & Monique Groot (2024). *D2.2 Community Interactions: Scenarios and Results* (No. D2.2; DE-BIAS). Netherlands Institute of Sound and Vision.

Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (2020). Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5454–5476). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.485

Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, 77–91. https://proceedings.mlr.press/v81/buolamwini18a.html

Chilcott, A. (2022). Towards protocols for describing racially offensive language in UK public archives. In V. Frings-Hessami & F. Foscarini (Eds.), *Archives in a Changing Climate—Part I & Part II* (pp. 151–168). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-19289-0_10

Foucault, M. (2002). *Archaeology of Knowledge* (2nd ed.). Routledge. https://doi.org/10.4324/9780203604168

Foucault, M., & Hurley, R. (2008). The history of sexuality. Volume 1, The will to knowledge. Penguin.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., III, H. D., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, *64*(12), 86–92. https://doi.org/10.1145/3458723

Kizhner, I., Terras, M., Rumyantsev, M., Khokhlova, V., Demeshkova, E., Rudov, I., & Afanasieva, J. (2021). Digital cultural colonialism: Measuring bias in aggregated digitized content held in Google Arts and Culture. *Digital Scholarship in the Humanities*, *36*(3), 607–640. https://doi.org/10.1093/llc/fqaa055

Luthra, M., & Eskevich, M. (2024). Data-Envelopes for Cultural Heritage: Going beyond Datasheets. In I. Siegert & K. Choukri (Eds.), *Proceedings of the Workshop on Legal and Ethical Issues in Human Language Technologies @ LREC-COLING 2024* (pp. 52–65). ELRA and ICCL. https://aclanthology.org/2024.legal-1.9

Masschelein, A., Truyen, F., Taes, S., van Mulder, J., Stynen, A., & Pireddu, R. (2023). *Report on research into bias types and patterns, including a typology applied to Europeana use cases and a vocabulary co-created with communities* (D2.1; DE-BIAS). KU Leuven.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys*, *54*(6), 115:1-115:35. https://doi.org/10.1145/3457607

Modest, W., & Lelijveld, R. (2018). *Words Matter: An unfinished guide to word choices in the cultural sector*. The National Museum for World Cultures (Tropenmuseum, Afrikamuseum, Museum Volkenkunde, Wereldmuseum). https://amsterdam.wereldmuseum.nl/en/about-wereldmuseum-amsterdam/research/words-matter-publication

Navigli, R., Conia, S., & Ross, B. (2023). Biases in Large Language Models: Origins, Inventory, and Discussion. *Journal of Data and Information Quality*, *15*(2), 10:1-10:21. https://doi.org/10.1145/3597307

Orr, W., & Crawford, K. (2023). *The social construction of datasets: On the practices, processes and challenges of dataset creation for machine learning*. OSF. https://doi.org/10.31235/osf.io/8c9uh

O'Sullivan, J. (2024). *The Bloomsbury Handbook to the Digital Humanities*. Bloomsbury Publishing. http://www.bloomsburycollections.com/collections/monograph

Prescott, A. (2023). Bias in Big Data, Machine Learning and AI: What Lessons for the Digital Humanities? *Digital Humanities Quarterly*, *17*(2). https://www.digitalhumanities.org/dhq/vol/17/2/000689/000689.html

Scheuerman, M. K., Spiel, K., Haimson, O. L., Hamidi, F., & Branham, S. M. (2020). *HCI Guidelines for Gender Equity and Inclusivity*. https://doi.org/10.13016/M2NW1F-P0JX

Scott, J. W. (1986). Gender: A Useful Category of Historical Analysis. *The American Historical Review*, *91*(5), 1053–1075. https://doi.org/10.2307/1864376

Scott, J. W. (2010). Gender: Still a Useful Category of Analysis? *Diogenes*, *57*(1), 7–14. https://doi.org/10.1177/0392192110369316

Sherman, J., Morrison, R., Klein, L., & Rosner, D. (2024). The Power of Absence: Thinking with Archival Theory in Algorithmic Design. *Designing Interactive Systems Conference*, 214–223.

https://doi.org/10.1145/3643834.3660690
Sitzia, E. (2023). Multiple Narratives and Polyvocality as Strategies of Inclusive Public Participation: Challenges and Disruption in the History Museum. *Muséologies. Les Cahiers d'études Supérieures*, *10*(2), 51–63. https://doi.org/10.7202/1108037ar
Søgaard, A., Plank, B., & Hovy, D. (2014). Selection bias, label bias, and bias in ground truth. *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Tutorial Abstracts*, 11–13. https://aclanthology.org/C14-3005.pdf
Stoler, A. L. (2010). *Along the Archival Grain: Epistemic Anxieties and Colonial Common Sense*. Princeton University Press.
Suresh, H., & Guttag, J. (2021). A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. *Equity and Access in Algorithms, Mechanisms, and Optimization*, 1–9. https://doi.org/10.1145/3465416.3483305
Thylstrup, N. B. (2019). *The politics of mass digitization*. MIT Press.
Trouillot, M.-R. (1995). *Silencing the Past: Power and the Production of History*. Beacon Press.
van Rossum, M. (2019). Labouring Transformations of Amphibious Monsters: Exploring Early Modern Globalization, Diversity, and Shifting Clusters of Labour Relations in the Context of the Dutch East India Company (1600–1800). *International Review of Social History*, *64*(S27), 19–42. Cambridge Core. https://doi.org/10.1017/S0020859019000014

Supporting Children with Linguistic Vulnerabilities Through Advanced, Theory-Driven Technological Solutions: The TELMÌ Approach for Italian children with DLD and Children with Italian as L2

Arianna Compostella[1], Giulia Valcamonica[2], Mattia Gianotti[3], Matteo Secco[4], Silvia Silleresi[5], Fabrizio Arosio[6], Franca Garzotto[7], Maria Teresa Guasti[8]

[1] University of Milano – Bicocca, Italy – arianna.compostella@unimib.it
[2] Politecnico di Milano, Italy – giulia.valcamonica@polimi.it
[3] Politecnico di Milano, Italy – mattia.gianotti@polimi.it
[4] Politecnico di Milano, Italy – matteo.secco@polimi.it
[5] University of Milano – Bicocca, Italy – silvia.silleresi@unimib.it
[6] University of Milano – Bicocca, Italy – fabrizio.arosio@unimib.it
[7] University of Milano – Bicocca, Italy; Politecnico di Milano, Italy – franca.garzotto@polimi.it
[8] University of Milano – Bicocca, Italy – mariateresa.guasti@unimib.it

ABSTRACT (ENGLISH)

This paper presents the early stages of a research project employing advanced technological solutions to improve language skills in two groups of 5– to 8-year-old children with linguistic vulnerabilities: those with Developmental Language Disorder (DLD) and those with Italian as a second language (L2). Although both groups may exhibit similar linguistic difficulties, the underlying causes are distinct, necessitating tailored interventions. Currently, the lack of appropriate linguistic support interventions and normative data for L2 learners underscore the urgent need to investigate how learning processes differ between the two groups to develop more targeted training programs. Another major challenge in conventional speech therapy is the reliance on traditional, paper-based methods, often perceived as stressful and unengaging by children. These methods frequently overlook the broader cognitive, social, and emotional context of language, and can be costly and logistically demanding for many families. To address these limitations, this study introduces a multi-week training program delivered through an innovative technological platform called TELMÌ, which is grounded in solid theoretical principles from speech therapy and psycholinguistics. TELMÌ aims to foster children's morphosyntactic skills through interactive storytelling deliverable through two distinct modalities: a tablet-based web application and an immersive, interactive multisensory system known as the Magic Room. This study has two main objectives: (1) to assess whether children in the DLD and L2 groups show improvements in their linguistic skills and, if so, to examine the qualitative and quantitative differences in their learning; (2) to explore whether the delivery format (tablet vs. Magic Room) yields distinct outcomes. Findings will inform the design of specialized, context-aware interventions for each population and shed light on the potential benefits of integrating technology into speech therapy.
Keywords: Developmental Language Disorders (DLD); L2 Italian Learners; Technology in speech therapy; smart spaces; interactive storytelling.

ABSTRACT (ITALIANO)

Offrire un supporto ai bambini con vulnerabilità linguistiche attraverso soluzioni tecnologiche avanzate e teoricamente motivate: l'approccio TELMÌ per i bambini italiani con DPL e per quelli con italiano come L2.

Questo articolo presenta le fasi iniziali di un progetto di ricerca che impiega soluzioni tecnologiche avanzate per migliorare le competenze linguistiche in due gruppi di bambini di età compresa tra i 5 e gli 8 anni con vulnerabilità linguistiche: quelli con Disturbo Primario del Linguaggio (DPL) e quelli che usano l'italiano come seconda lingua (L2). Nonostante i due gruppi possano presentare difficoltà linguistiche simili, le cause sottostanti le fragilità linguistiche nei due gruppi sono diverse, richiedendo pertanto interventi mirati. Al momento, la mancanza di interventi adeguati di supporto linguistico e di dati normativi per i bambini con italiano L2 evidenzia l'urgente necessità di indagare come i processi di apprendimento differiscano tra questi due gruppi, al fine di sviluppare programmi di supporto più specifici. Un'ulteriore sfida nelle terapie logopediche convenzionali è rappresentata dall'uso di strumenti cartacei spesso percepiti dai bambini come stressanti e poco coinvolgenti. Tali metodi trascurano di frequente il più ampio contesto cognitivo, sociale ed emotivo del linguaggio e possono risultare costosi e impegnativi dal punto di vista

logistico per molte famiglie. Per superare tali limitazioni, questo studio introduce un programma di training linguistico, basato su solidi principi teorici della logopedia e della psicolinguistica, della durata di alcune settimane, erogato attraverso una piattaforma tecnologica innovativa chiamata TELMÌ. TELMÌ mira a potenziare le abilità morfosintattiche dei bambini attraverso la narrazione interattiva, proposta in due modalità: un'applicazione web su tablet e un sistema immersivo e multisensoriale chiamato "Magic Room". Lo studio ha due obiettivi principali: (1) valutare se i bambini italiani con DPL e quelli con italiano L2 mostrino miglioramenti nelle loro competenze linguistiche e, in caso affermativo, esaminare le differenze qualitative e quantitative nei processi di apprendimento tra i due gruppi; (2) indagare se diverse modalità di erogazione (tablet vs. Magic Room) producano esiti differenti. I risultati forniranno indicazioni utili alla progettazione di interventi specializzati e sensibili al contesto per ciascuna popolazione, facendo luce sui potenziali benefici dell'integrazione tecnologica nella terapia logopedica.
Parole chiave: Disturbo Primario del Linguaggio (DPL); bambini con Italiano L2; Soluzioni tecnologiche in logopedia; spazi intelligenti; narrazione interattiva.

1. INTRODUCTION

Possessing solid language skills is fundamental for social integration and for individuals to fully realize their potential in both educational and professional arenas, as well as in maintaining a fulfilling social life. Effective linguistic abilities are paramount, serving as the cornerstone for interacting with others and navigating various social contexts. Indeed, language is a transversal skill that significantly influences a wide range of other competencies, including cognitive development, academic achievement, and emotional well-being. However, when these abilities are compromised—as in the case of children with Developmental Language Disorder (DLD), a disorder that specifically affects language acquisition and use—such challenges can hinder not only academic performance, but also social and emotional development[1] . Children learning Italian as a second language (L2) may also experience language difficulties, especially when compared to typically developing peers who speak Italian as their first (L1) language[2]. Interestingly, it has been found that children with DLD and those learning Italian as a L2 may share common difficulties, at least superficially, especially during the early years of L2 acquisition[3]. Despite this, while Italian L2 children may initially appear similar to those with DLD, a detailed examination reveals that their language profiles differ significantly in both quality and quantity (Vender et al., 2016). Indeed, it is important to emphasize that the linguistic difficulties experienced by children with DLD and those learning Italian as L2 have different causes and origins—DLD is believed to result from various factors, including genetic, neurological, and environmental influences, while the challenges faced by children learning Italian as an L2 are associated with limited language exposure. Because their language difficulties arise from different sources, both populations need tailored support and training programs. Providing targeted support to both groups is essential to ensure their effective integration and overall development (Guasti, 2017). Currently, there are substantial issues concerning both the support for L2 children with language fragilities and the language rehabilitation of children with DLD. Children with Italian as L2 do not always receive the necessary or adequate attention because their need for support is frequently underestimated. On the other hand, they are sometimes treated as if they have DLD, due to the lack of specific tools to differentiate deficiencies caused by insufficient exposure from those originating from a

1 DLD is a highly heterogeneous disorder and includes various cases. In general, children with DLD start speaking later than their peers and exhibit slow language development (Guasti, 2017). Their difficulties can selectively affect different linguistic areas. They often show phonological challenges (Bortolini et al., 2006), a vocabulary below the typical range, difficulties in the correct use of morphosyntactic structures, and challenges with functional morphemes such as third person plural inflection (Bortolini et al., 2006) and clitics (Arosio et al., 2014; Bortolini et al., 2006), among others. However, some researchers argue that these language disorders can co-occur with non-purely linguistic disorders. Sometimes, difficulties are also observed in non-linguistic areas such as verbal memory (Gathercole, 2006) and fine motor skills (Adi-Japha et al., 2011; Hill, 2001).

2 The proficiency of Italian L2 children is strongly influenced by the timing of their first exposure to the L2 language. The literature distinguishes different types of bilingualism based on the period of language exposure (Abdelilah-Bauer, 2008; Contento, 2010; Guasti, 2017). In general, bilingualism is categorized as simultaneous or sequential, depending on whether exposure to L2 occurs from birth or begins after birth, respectively.

3 Children with Italian as L2 may have a more limited vocabulary and perform worse in lexical comprehension compared to their monolingual peers (Bialystok et al., 2010; Guasti, 2017; Oller et al., 2007). Additionally, they may exhibit reduced average sentence length (Guasti, 2017) and face challenges in using morphosyntactic elements and functional words (Paradis et al., 2008; Paradis & Crago, 2000; Vender et al., 2016).

linguistic problem, as well as the absence of normative data on L2 acquisition (Bedore & Peña, 2008; Guasti, 2017). Therefore, there is an urgent need to clearly distinguish between the two conditions in order to develop tailored training approaches for each distinct group of children. As for the conventional speech-language therapy for DLD children, it commonly relies on paper-based materials. While these materials successfully target some of the most challenging linguistic skills for these children, the methods of their administration present several drawbacks (Spitale et al., 2021). Firstly, therapy is often perceived by children as stressful and frustrating. These exercises are typically administered in a setting where the child is required to respond directly to the therapist, which can cause pressure and anxiety for some children. Another issue is the lack of engagement and appeal in these exercises, which children often perceive as boring, leading to limited effectiveness for training. Lack of engagement can lower children's motivation and participation. When activities seem monotonous or irrelevant to their everyday lives, their willingness to participate and practice declines, reducing the overall effectiveness of the intervention. Moreover, many of these exercises are overly artificial, straying from the way in which language acquisition naturally occurs, involving multiple aspects of communication, including social and relational elements. The artificial nature of conventional speech-language therapy can lead to negative outcomes: by concentrating narrowly on isolated markers, these sessions often fail to support the holistic language development children need. Because language is inherently tied to cognitive, social, and emotional domains, ignoring its broader context hinders children's ability to integrate language skills into everyday communication, ultimately limiting both their overall proficiency and practical use of language. Another important aspect of conventional therapy to consider is cost and accessibility. Therapy sessions come with expenses that, unfortunately, many families find difficult to afford, especially considering those children who come from low socio-economic backgrounds. Additionally, therapy often requires traveling to the therapist's location and organizing schedules. As a result, the process is frequently demanding in terms of both cost and time. Furthermore, extending interventions to environments beyond the clinical setting, such as homes and schools, presents significant difficulties (Spitale et al., 2021). As emphasized by Spitale and colleagues (2021), a more holistic and ecological approach is necessary. This would foster children's motivation, which appears to mediate the cognitive factors involved in learning (Moreno & Mayer, 2007). Higher motivation would in turn lead to better performance and lower dropout rates during training. This is where technology can address emerging needs. Technology can be leveraged to promote inclusivity by providing accessible, personalized, and engaging tools that cater to the needs of children with linguistic fragilities.

2. TECHNOLOGY IN SPEECH THERAPY

Recent studies have explored the integration of interactive technologies into speech-language therapy to address some of the previously mentioned challenges (Spitale et al., 2021). Technological solutions providing game immersion experiences have proven effective in engaging and captivating children more successfully than traditional intervention methods (Cheng et al., 2015). Moreover, the use of technology in training solutions appears to enhance the acquisition of language skills more rapidly and support longer retention of these skills (Durrleman et al., 2019; Edwards & Dukhovny, 2017). There are several technological solutions that can be adopted for this purpose, including computer devices (Adrian et al., 2003; Clendon et al., 2003; Engwall et al., 2006; Hair et al., 2020; Nanavati et al., 2018; Saz et al., 2009; Spitale et al., 2021; Stark & Warburton, 2018), robots (Buchwald et al., 2018; Charron et al., 2017; Robles-Bykbaev, Lopez-Nores, Pazos-Arias, et al., 2015; Robles-Bykbaev, Lopez-Nores, Ochoa-Zambrano, et al., 2015; Robles-Bykbaev et al., 2016; Spitale et al., 2021, 2023) and smart spaces created with augmented reality (AR) (Da Silva et al., 2015; Wanti et al., 2024) or virtual reality (VR) (Cappadona et al., 2023; Devane et al., 2023; Giachero et al., 2020; Vaezipour et al., 2022, 2023). By integrating these technologies into speech-language training interventions, it is possible to develop more dynamic, engaging, and personalized intervention strategies, ultimately enhancing the overall effectiveness of the training. Each of these technological solutions offers its own advantages. The affordability and widespread availability of computer devices make therapy accessible to more children in need and provide content beyond the therapist's office, allowing sessions to be undertaken at more convenient times while reducing financial and logistical obstacles. Robotic systems can engage children through interactive activities that promote skills important for language, such as turn-taking, eye contact, attention and executive functions (Spitale et al., 2023). Additionally, interacting with a robot might make children feel more comfortable, encouraging them to participate more freely without fear of being judged compared to interacting with a speech therapist. Augmented reality (AR) and virtual reality (VR) provide immersive experiences that

blend digital content with the physical world, making learning more engaging and effective by replicating and simulating real-life learning situations. According to the theoretical foundations of Embodied Cognition, the development of all human cognitive functions is related to our sensorimotor experiences (Wilson & Golonka, 2013). These theories have gained significant traction in the field of language acquisition and processing (Kontra et al., 2012). It appears that sensorimotor interactions with the environment help children process linguistic rules and acquire concepts (Gibbs, 2005). There has long been evidence supporting the beneficial effects of multisensory procedures on language development (Ayres & Mailloux, 1981). The benefits are not limited to strictly language-specific abilities but also include improved academic performance and greater ease of interaction in social contexts (Mauer, 1999).

3. CASE STUDY

The objective of this study is to adopt an innovative methodology that employs an advanced technological solution to overcome the limitations of traditional therapy in providing linguistic training and support to two populations with language fragilities: children aged between 5 and 8 years diagnosed with DLD and those with Italian as a second language (Italian L2). The training program aims to enhance the morphosyntactic skills of the two groups of children effectively. The training activities consist of interactive stories featuring selected morphosyntactic structures which are delivered in two formats: through tablet (web app) and by utilizing a more sophisticated immersive system: a multisensory room called the *Magic Room (*Garzotto & Gelsomini, 2018). With this study we aim to evaluate the effectiveness of training based on these technological solutions by assessing whether participants show improvements in their linguistic skills after the training and whether the proposed activities are engaging and interesting for them. Moreover, we aim to determine whether there are learning differences between the two modalities (tablet vs. Magic Room) and between the two populations (Italian children with DLD vs. children with Italian as L2). With this study, we aim to build upon and extend the work of Garzotto and Gelsomini (2018) by addressing some of the limitations identified in their previous research.

4. METHODOLOGY

The study will involve approximately 120 children aged 5 to 8, divided into two main groups: 60 native Italian-speaking children with DLD and 60 typically developing children with Italian as L2. Each main group will be further divided into three subgroups of 20 children each, with the following participation modalities: one group will carry out the training in the Magic Room, a second group will train using the tablet, a third group will passively listen to the stories without performing specific activities during the narration. This group will serve as the control group: although exposed to the same stories as the other groups, they will not actively participate. The presence of this group is crucial to demonstrating the effectiveness of the active training. The study comprises three primary phases: an initial assessment of linguistic abilities prior to training (pre-training), the training intervention itself (conducted either in the multisensory room or via tablet), and a subsequent assessment of linguistic abilities following the training (post-training). For both the pre-training and post-training assessments, standardized tests from the BVL battery (Marini et al., 2015) will be administered to evaluate language production and comprehension. Additionally, during the pre-training assessment, the standardized Raven's Progressive Matrices (Raven, 1956) will be administered to evaluate non-verbal intelligence. The training phase consists of ten sessions, each lasting approximately one hour. In each session, children are presented with interactive stories that incorporate target sentences designed to train morphosyntactic abilities in comprehension. These include: passive constructions (Reggiani, 2010; Leonard et al., 2003, 2006; Riches, 2013; Robertson & Joanisse, 2010); relative clauses (Arosio et al., 2017, 2024; Contemori & Garraffa, 2010; Håkansson & Hansson, 2000; Novogrodsky & Friedmann, 2006; Rakhlin et al., 2016; Riches et al., 2010; Stavrakaki, 2002; Wada et al., 2020; Wang & Yu, 2021, 2022); clitic pronouns (Arosio et al., 2014; Guasti, 2017); inflectional morphology, including the pluralization of both real words and non-words, and subject-verb agreement with auxiliaries in compound tenses (Chondrogianni & John, 2018; Deevy & Leonard, 2018; Devane et al., 2023; Hadley & Rice, 1996; Leonard, 2014; Niemi, 1999; Oetting & Rice, 1993; Rice & Wexler, 1996). The interactive stories have been designed to be engaging, stimulating, instructive, enjoyable, and appropriate for the target age group of the children. We adopted the storytelling format, as in Spitale and colleagues (2021), since extensive research has shown that storytelling can effectively boost and improves children's language skills and their syntactic comprehension (Isbell et al., 2004). The stories were designed to be scalable and maintain a consistent structure. The entire development process, encompassing both the UX design of activities and the coding process, leverages an innovative pattern-oriented approach with

multiple levels of abstraction, which supports the reuse and scalability, both at the level of UX design and implementation phase (Borchers, 2001; Schmidt et al., 2000; Pree & Sikora, 1997; Spitale et al., 2021; Van Welie et al., 2001). The interactive stories are delivered through a web app called TELMÌ, which supports language training activities in two formats: through tablet (web app) (Figure 1B) and by utilizing a more sophisticated immersive system: a multisensory room called the *Magic Room* (Figure 1A). The Magic Room employs innovative software and hardware technologies to transform indoor spaces into interactive, multisensory, and highly ecological environments. Users can be exposed to a variety of visual and auditory stimuli that can be integrated and controlled through gestures and movements, thereby facilitating the use of language by recreating everyday situations. The two versions of TELMÌ (tablet and Magic Room) work in the same way, with a difference only in how participants interact with the systems. In both versions, each segment of the story is accompanied by three different images (a target, a competitor, and a distractor), following the principle of the picture-matching task. Only one of these three images (i.e., the target image) accurately reflects the content described by the pre-recorded narration. The child's task is to select the image that they believe correctly corresponds to the heard sentence. In the web app the participants interact by clicking on the image on the tablet (Figure 2B), while in the Magic Room users must physically move, by either touching the projected image on the wall, stepping onto it, or placing an object on the image if it is projected on the floor (Figure 2A).

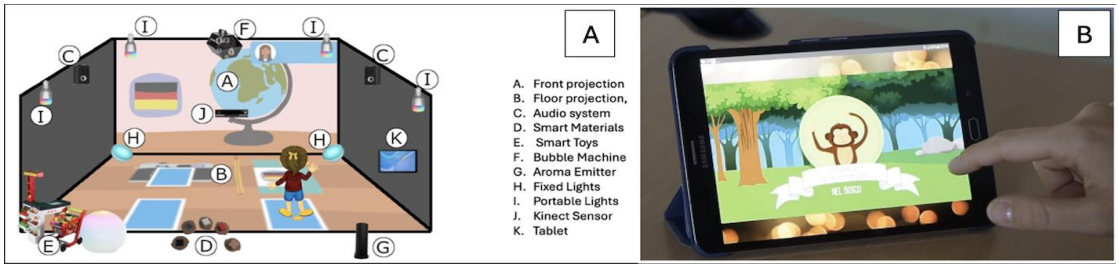


Figure 1A: The Magic Room and its components.

Figure 1B: TELMÌ tablet application.

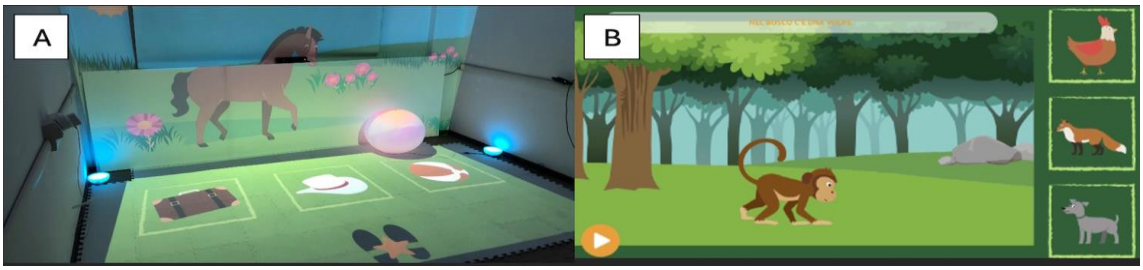


Figure 2: A story segment featuring three different images, with only one image accurately reflecting the content presented by the pre-recorded narration. Figure 2A represents the Magic Room version. In this case, the child has to step onto the image projected on the floor which corresponds to the heard sentence.

Figure 2B: The tablet version. The child's task is to select the image that they believe correctly corresponds to the heard sentence by touching it.

In the Magic Room responses are recorded via sensors – a Kinect device to track participants' movements and positions – and no video recordings are made. In both the tablet and Magic Room versions, if a child selects an incorrect image, they receive immediate feedback and the opportunity to attempt the response again. The response time and the number of attempts are recorded for each trial.
Consistent with findings in the literature on the use of technology in language therapy, we expect both types of training to yield positive effects. Furthermore, we hypothesize that the Magic Room may be more engaging than the tablet, leading to greater improvements in performance. Indeed, the greater sensorimotor engagement provided by the multisensory room could address the issue of "non-ecological" approaches typical of traditional clinical protocols and training methods. This in turn would foster higher motivation, which mediates the cognitive factors involved in learning (Moreno & Mayer, 2007), thus promoting better performance and reducing the risk of dropouts during training. Another objective of our study is to examine whether there are differences in training between the two groups of children and, if so, to identify what those differences are. As previously explained, the origins and causes of linguistic difficulties differ between the two populations; therefore, we expect to observe differences in their learning

processes. Specifically, we hypothesize that the two groups will exhibit distinct linguistic behaviors and that their learning pace (measured in terms of training benefits) will differ. More precisely, we anticipate that children with Italian L2 will demonstrate a faster learning pace during the sessions compared to Italian children with DLD.

5. CONCLUSIONS

We presented the design of an experimental study that harnesses cutting-edge technology to enhance language abilities in two populations with fragilities: children with DLD and those with Italian as a second language. Our research has the potential to guide future studies, drive the development of targeted and tailored linguistic training interventions for the two groups, and illuminate both the benefits and possible limitations of using technology in speech therapy.

ACKNOWLEDGEMENTS

This work was supported by the Italian National Recovery and Resilience Plan (PNRR), Mission 4, Component 2, Investment 1.1, funded by the European Union – NextGenerationEU, under the PRIN 2022 project "A Technology-driven linguistic intervention for children with language and literacy weaknesses", CUP: B53D23014380006.

REFERENCES

Abdelilah-Bauer, B. (2008). Il bambino bilingue: crescere parlando più di una lingua. R. Cortina.

Adi-Japha, E., Strulovich-Schwartz, O., & Julius, M. (2011). Delayed motor skill acquisition in kindergarten children with language impairment. Research in Developmental Disabilities, 32(6), 2963–2971. https://doi.org/10.1016/j.ridd.2011.05.005

Adrian, J., Gonzalez, M., & Buiza, J. (2003). CASE STUDY The use of computer-assisted therapy in anomia rehabilitation: A single-case report. Aphasiology, 17(10), 981–1002. https://doi.org/10.1080/02687030344000256

Arosio, F., Branchini, C., Barbieri, L., & Guasti, M. T. (2014). Failure to produce direct object clitic pronouns as a clinical marker of SLI in school-aged Italian speaking children. Clinical Linguistics & Phonetics, 28(9), 639–663. https://doi.org/10.3109/02699206.2013.877081

Arosio, F., Panzeri, F., Molteni, B., Magazù, S., & Guasti, M. T. (2017). The comprehension of Italian relative clauses in poor readers and in children with Specific Language Impairment. Glossa: A Journal of General Linguistics, 2(1). https://doi.org/10.5334/gjgl.107

Arosio, F., Silleresi, S., & Guasti, M. T. (2024). The production of relative clauses in Italian-speaking children with DLD. Isogloss. Open Journal of Romance Linguistics, 10(1), 1–23. https://doi.org/10.5565/rev/isogloss.306

Ayres, A. J., & Mailloux, Z. (1981). Influence of Sensory Integration Procedures on Language Development. The American Journal of Occupational Therapy, 35(6), 383–390. https://doi.org/10.5014/ajot.35.6.383

Bedore, L. M., & Peña, E. D. (2008). Assessment of Bilingual Children for Identification of Language Impairment: Current Findings and Implications for Practice. International Journal of Bilingual Education and Bilingualism, 11(1), 1–29. https://doi.org/10.2167/beb392.0

Bialystok, E., Luk, G., Peets, K. F., & Yang, S. (2010). Receptive vocabulary differences in monolingual and bilingual children. Bilingualism: Language and Cognition, 13(4), 525–531. https://doi.org/10.1017/S1366728909990423

Borchers, J. O. (2001). A pattern approach to interaction design. AI & Society, 15(4), 359–376. https://doi.org/10.1007/BF01206115

Bortolini, U., Arfé, B., Caselli, C. M., Degasperi, L., Deevy, P., & Leonard, L. B. (2006). Clinical markers for specific language impairment in Italian: The contribution of clitics and non-word repetition. International Journal of Language & Communication Disorders, 41(6), 695–712. https://doi.org/10.1080/13682820600570831

Buchwald, A., Falconer, C., Rykman-Peltz, A., Cortes, M., Pascual-Leone, A., Thickbroom, G. W., Krebs, H. I., Fregni, F., Gerber, L. M., Oromendia, C., Chang, J., Volpe, B. T., & Edwards, D. J. (2018). Robotic Arm Rehabilitation in Chronic Stroke Patients With Aphasia May Promote Speech and Language

Recovery (but Effect Is Not Enhanced by Supplementary tDCS). Frontiers in Neurology, 9, 853. https://doi.org/10.3389/fneur.2018.00853

Cappadona, I., Ielo, A., La Fauci, M., Tresoldi, M., Settimo, C., De Cola, M. C., Muratore, R., De Domenico, C., Di Cara, M., Corallo, F., Tripodi, E., Impallomeni, C., Quartarone, A., & Cucinotta, F. (2023). Feasibility and Effectiveness of Speech Intervention Implemented with a Virtual Reality System in Children with Developmental Language Disorders: A Pilot Randomized Control Trial. Children, 10(8), 1336. https://doi.org/10.3390/children10081336

Charron, N., Lewis, L., & Craig, M. (2017). A Robotic Therapy Case Study: Developing Joint Attention Skills With a Student on the Autism Spectrum. Journal of Educational Technology Systems, 46(1), 137–148. https://doi.org/10.1177/0047239516687721

Cheng, M. -T., She, H. -C., & Annetta, L. A. (2015). Game immersion experience: Its hierarchical structure and impact on game-based science learning. Journal of Computer Assisted Learning, 31(3), 232–253. https://doi.org/10.1111/jcal.12066

Chondrogianni, V., & John, N. (2018). Tense and plural formation in Welsh–English bilingual children with and without language impairment. International Journal of Language & Communication Disorders, 53(3), 495–514. https://doi.org/10.1111/1460-6984.12363

Clendon, S., Flynn, M. C., & Coombes, T. (2003). Facilitating speech and language development in children with cochlear implants using computer technology. Cochlear Implants International, 4(3), 119–136. https://doi.org/10.1179/cim.2003.4.3.119

Contemori, C., & Garraffa, M. (2010). Comparison of modalities in SLI syntax: A study on the comprehension and production of non-canonical sentences. Lingua, 120(8), 1940–1955. https://doi.org/10.1016/j.lingua.2010.02.011

Contento, S. (a cura di.) (2010). Crescere nel bilinguismo. Aspetti cognitivi, linguistici ed emotivi. Carocci.

Da Silva, C. A., Fernandes, A. R., & Grohmann, A. P. (2015). STAR: Speech Therapy with Augmented Reality for Children with Autism Spectrum Disorders. In J. Cordeiro, S. Hammoudi, L. Maciaszek, O. Camp, & J. Filipe (Eds.), Enterprise Information Systems (Vol. 227, pp. 379–396). Springer International Publishing. https://doi.org/10.1007/978-3-319-22348-3_21

Deevy, P., & Leonard, L. B. (2018). Sensitivity to Morphosyntactic Information in Preschool Children With and Without Developmental Language Disorder: A Follow-Up Study. Journal of Speech, Language, and Hearing Research, 61(12), 3064–3074. https://doi.org/10.1044/2018_JSLHR-L-18-0038

Devane, N., Behn, N., Marshall, J., Ramachandran, A., Wilson, S., & Hilari, K. (2023). The use of virtual reality in the rehabilitation of aphasia: A systematic review. Disability and Rehabilitation, 45(23), 3803–3822. https://doi.org/10.1080/09638288.2022.2138573

Durrleman, S., Burnel, M., De Villiers, J. G., Thommen, E., Yan, R., & Delage, H. (2019). The Impact of Grammar on Mentalizing: A Training Study Including Children With Autism Spectrum Disorder and Developmental Language Disorder. Frontiers in Psychology, 10, 2478. https://doi.org/10.3389/fpsyg.2019.02478

Edwards, J., & Dukhovny, E. (2017). Technology Training in Speech-Language Pathology: A Focus on Tablets and Apps. Perspectives of the ASHA Special Interest Groups, 2(10), 33–48. https://doi.org/10.1044/persp2.SIG10.33

Engwall, O., Bälter, O., Öster, A.-M., & Kjellström, H. (2006). Designing the user interface of the computer-based speech training system ARTUR based on early user tests. Behaviour & Information Technology, 25(4), 353–365. https://doi.org/10.1080/01449290600636702

Garzotto, F., & Gelsomini, M. (2018). Magic Room: A Smart Space for Children with Neurodevelopmental Disorder. IEEE Pervasive Computing, 17(1), 38–48. https://doi.org/10.1109/MPRV.2018.011591060

Gathercole, S. E. (2006). Nonword repetition and word learning: The nature of the relationship. Applied Psycholinguistics, 27(4), 513–543. https://doi.org/10.1017/S0142716406060383

Giachero, A., Calati, M., Pia, L., La Vista, L., Molo, M., Rugiero, C., Fornaro, C., & Marangolo, P. (2020). Conversational Therapy through Semi-Immersive Virtual Reality Environments for Language Recovery and Psychological Well-Being in Post Stroke Aphasia. Behavioural Neurology, 2020, 1–15. https://doi.org/10.1155/2020/2846046

Gibbs, J. R. R. W. (2005). Embodiment and Cognitive Science. Cambridge University Press. https://doi.org/10.1017/CBO9780511805844

Guasti, M.T. (2017). I disturbi specifici del linguaggio nei bambini con italiano L2. Giornale Di Neuropsichiatria Dell'età Evolutiva, 37, 83–91.

Hadley, P. A., & Rice, M. L. (1996). Emergent Uses of BE and DO: Evidence From Children With Specific Language Impairment. Language Acquisition, 5(3), 209–243. https://doi.org/10.1207/s15327817la0503_2

Hair, A., Markoulli, C., Monroe, P., McKechnie, J., Ballard, K. J., Ahmed, B., & Gutierrez-Osuna, R. (2020). Preliminary Results From a Longitudinal Study of a Tablet-Based Speech Therapy Game. Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems, 1–8. https://doi.org/10.1145/3334480.3382886

Håkansson, G., & Hansson, K. (2000). Comprehension and production of relative clauses: A comparison between Swedish impaired and unimpaired children. Journal of Child Language, 27(2), 313–333. https://doi.org/10.1017/S0305000900004128

Hill, E. L. (2001). Non-specific nature of specific language impairment: A review of the literature with regard to concomitant motor impairments. International Journal of Language & Communication Disorders, 36(2), 149–171. https://doi.org/10.1080/13682820010019874

Isbell, R., Sobol, J., Lindauer, L., & Lowrance, A. (2004). The Effects of Storytelling and Story Reading on the Oral Language Complexity and Story Comprehension of Young Children. Early Childhood Education Journal, 32(3), 157–163. https://doi.org/10.1023/B:ECEJ.0000048967.94189.a3

Kontra, C., Goldin-Meadow, S., & Beilock, S. L. (2012). Embodied Learning Across the Life Span. Topics in Cognitive Science, 4(4), 731–739. https://doi.org/10.1111/j.1756-8765.2012.01221.x

Leonard, L. B. (2014). Children with Specific Language Impairment (2nd ed.). The MIT Press. https://doi.org/10.7551/mitpress/9152.001.0001

Leonard, L. B., Deevy, P., Miller, C. A., Rauf, L., Charest, M., & Kurtz, R. (2003). Surface Forms and Grammatical Functions: Past Tense and Passive Participle Use by Children With Specific Language Impairment. Journal of Speech, Language, and Hearing Research, 46(1), 43–55. https://doi.org/10.1044/1092-4388(2003/004)

Leonard, L. B., Wong, A. M.-Y., Deevy, P., Stokes, S. F., & Fletcher, P. (2006). The production of passives by children with specific language impairment: Acquiring English or Cantonese. Applied Psycholinguistics, 27(2), 267–299. https://doi.org/10.1017/S0142716406060280

Marini, A., Marotta, L., Bulgheroni, S., & Fabbro, F. (2015). Batteria per la Valutazione del linguaggio in bambini dai 4 ai 12 anni (BVL_4-12). Giunti O.S.

Mauer, D. M. (1999). Issues and Applications of Sensory Integration Theory and Treatment With Children With Language Disorders. Language, Speech, and Hearing Services in Schools, 30(4), 383–392. https://doi.org/10.1044/0161-1461.3004.383

Moreno, R., & Mayer, R. (2007). Interactive Multimodal Learning Environments: Special Issue on Interactive Learning Environments: Contemporary Issues and Trends. Educational Psychology Review, 19(3), 309–326. https://doi.org/10.1007/s10648-007-9047-2

Nanavati, A., Dias, M. B., & Steinfeld, A. (2018). Speak Up: A Multi-Year Deployment of Games to Motivate Speech Therapy in India. Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, 1–12. https://doi.org/10.1145/3173574.3173892

Niemi, J. (1999). Production of Grammatical Number in Specific Language Impairment: An Elicitation Experiment on Finnish. Brain and Language, 68(1–2), 262–267. https://doi.org/10.1006/brln.1999.2100

Novogrodsky, R., & Friedmann, N. (2006). The production of relative clauses in syntactic SLI: A window to the nature of the impairment. Advances in Speech Language Pathology, 8(4), 364–375. https://doi.org/10.1080/14417040600919496

Oetting, J. B., & Rice, M. L. (1993). Plural Acquisition in Children With Specific Language Impairment. Journal of Speech, Language, and Hearing Research, 36(6), 1236–1248. https://doi.org/10.1044/jshr.3606.1236

Oller, D. K., Pearson, B. Z., & Cobo-Lewis, A. B. (2007). Profile effects in early bilingual language and literacy. Applied Psycholinguistics, 28(2), 191–230. https://doi.org/10.1017/S0142716407070117

Paradis, J., & Crago, M. (2000). Tense and Temporality: A Comparison Between Children Learning a Second Language and Children With SLI. Journal of Speech, Language, and Hearing Research, 43(4), 834–847. https://doi.org/10.1044/jslhr.4304.834

Paradis, J., Rice, M. L., Crago, M., & Marquis, J. (2008). The acquisition of tense in English: Distinguishing child second language from first language and specific language impairment. Applied Psycholinguistics, 29(4), 689–722. https://doi.org/10.1017/S0142716408080296

Pree, W., & Sikora, H. (1997). Design patterns for object-oriented software development (tutorial). Proceedings of the 19th International Conference on Software Engineering - ICSE '97, 663–664. https://doi.org/10.1145/253228.253810

Rakhlin, N., Kornilov, S. A., Kornilova, T. V., & Grigorenko, E. L. (2016). Syntactic Complexity Effects of Russian Relative Clause Sentences in Children with and without Developmental Language Disorder. Language Acquisition, 23(4), 333–360. https://doi.org/10.1080/10489223.2016.1179312

Raven, J. C. (1956). Guide to using the coloured progressive matrices: sets A, Ab, B (revised order, 1956). *H.K. Lewis.*

Reggiani, D. (2010). Dyslexia and the Acquisition of Syntax. Passive and Control. Doctoral Dissertation. Università degli Studi di Verona.

Rice, M. L., & Wexler, K. (1996). Toward Tense as a Clinical Marker of Specific Language Impairment in English-Speaking Children. Journal of Speech, Language, and Hearing Research, 39(6), 1239–1257. https://doi.org/10.1044/jshr.3906.1239

Riches, N. G. (2013). Treating the passive in children with specific language impairment: A usage-based approach. Child Language Teaching and Therapy, 29(2), 155–169. https://doi.org/10.1177/0265659012466667

Riches, N. G., Loucas, T., Baird, G., Charman, T., & Simonoff, E. (2010). Sentence repetition in adolescents with specific language impairments and autism: An investigation of complex syntax. International Journal of Language & Communication Disorders, 45(1), 47–60. https://doi.org/10.3109/13682820802647676

Robertson, E. K., & Joanisse, M. F. (2010). Spoken sentence comprehension in children with dyslexia and language impairment: The roles of syntax and working memory. Applied Psycholinguistics, 31(1), 141–165. https://doi.org/10.1017/S0142716409990208

Robles-Bykbaev, V., López-Nores, M., Pazos-Arias, J. J., & Garcia-Duque, J. (2015). RAMSES: A robotic assistant and a mobile support environment for speech and language therapy. Fifth International Conference on the Innovative Computing Technology (INTECH 2015), 1–4. https://doi.org/10.1109/INTECH.2015.7173473

Robles-Bykbaev, V., López-Nores, M., Ochoa-Zambrano, J., García-Duque, J., & Pazos-Arias, J. J. (2015). SPELTRA: A Robotic Assistant for Speech-and-Language Therapy. In M. Antona & C. Stephanidis (Eds.), Universal Access in Human-Computer Interaction. Access to Learning, Health and Well-Being (Vol. 9177, pp. 525–534). Springer International Publishing. https://doi.org/10.1007/978-3-319-20684-4_51

Robles-Bykbaev, V., Ochoa-Guaraca, M., Carpio-Moreta, M., Pulla-Sanchez, D., Serpa-Andrade, L., Lopez-Nores, M., & Garcia-Duque, J. (2016). Robotic assistant for support in speech therapy for children with cerebral palsy. 2016 IEEE International Autumn Meeting on Power, Electronics and Computing (ROPEC), 1–6. https://doi.org/10.1109/ROPEC.2016.7830603

Saz, O., Yin, S.-C., Lleida, E., Rose, R., Vaquero, C., & Rodríguez, W. R. (2009). Tools and Technologies for Computer-Aided Speech and Language Therapy. Speech Communication, 51(10), 948–967. https://doi.org/10.1016/j.specom.2009.04.006

Schmidt, D., Stal, M., Rohnert, H., & Buschmann, F. (2000). Pattern-oriented software architecture, Patterns for Concurrent and Networked Objects (Vol. 2). J. Wiley and sons.

Spitale, M., Silleresi, S., Garzotto, F., & Matarić, M. J. (2023). Using Socially Assistive Robots in Speech-Language Therapy for Children with Language Impairments. International Journal of Social Robotics, 15(9–10), 1525–1542. https://doi.org/10.1007/s12369-023-01028-7

Spitale, M., Silleresi, S., Leonardi, G., Arosio, F., Giustolisi, B., Guasti, M. T., & Garzotto, F. (2021). Design Patterns of Technology-based Therapeutic Activities for Children with Language Impairments: A Psycholinguistic-Driven Approach. Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems, 1–7. https://doi.org/10.1145/3411763.3451775

Stark, B. C., & Warburton, E. A. (2018). Improved language in chronic aphasia after self-delivered iPad speech therapy. Neuropsychological Rehabilitation, 28(5), 818–831. https://doi.org/10.1080/09602011.2016.1146150

Stavrakaki, S. (2002). A-bar movement constructions in Greek children with SLI: Evidence for deficits in the syntactic component of language. In E. Fava (Ed.), Current Issues in Linguistic Theory (Vol. 227, pp. 131–153). John Benjamins Publishing Company. https://doi.org/10.1075/cilt.227.12sta

Vaezipour, A., Aldridge, D., Koenig, S., Burns, C., Baghaei, N., Theodoros, D., & Russell, T. (2023). Rehabilitation Supported by Immersive Virtual Reality for Adults With Communication Disorders: Semistructured Interviews and Usability Survey Study. JMIR Rehabilitation and Assistive Technologies, 10, e46959. https://doi.org/10.2196/46959

Vaezipour, A., Aldridge, D., Koenig, S., Theodoros, D., & Russell, T. (2022). ”It’s really exciting to think where it could go": a mixed-method investigation of clinician acceptance, barriers and enablers of virtual reality technology in communication rehabilitation. Disability and Rehabilitation, 44(15), 3946–3958. https://doi.org/10.1080/09638288.2021.1895333

Van Welie, M., Van Der Veer, G. C., & Eliëns, A. (2001). Patterns as Tools for User Interface Design. In J. Vanderdonckt & C. Farenc (Eds.), Tools for Working with Guidelines (pp. 313–324). Springer London. https://doi.org/10.1007/978-1-4471-0279-3_30

Vender, M., Garraffa, M., Sorace, A., & Guasti, M. T. (2016). How early L2 children perform on Italian clinical markers of SLI: A study of clitic production and nonword repetition. Clinical Linguistics & Phonetics, 30(2), 150–169. https://doi.org/10.3109/02699206.2015.1120346

Wada, R., Gillam, S. L., & Gillam, R. B. (2020). The Use of Structural Priming and Focused Recasts to Facilitate the Production of Subject- and Object-Focused Relative Clauses by School-Age Children With and Without Developmental Language Disorder. American Journal of Speech-Language Pathology, 29(4), 1883–1895. https://doi.org/10.1044/2020_AJSLP-19-00090

Wang, H., & Yu, H. (2021). The Production of Relatives in Mandarin Children With Specific Language Impairment–From the Perspective of Edge Feature Underspecification Hypothesis. Frontiers in Psychology, 12, 705526. https://doi.org/10.3389/fpsyg.2021.705526

Wang, H., & Yu, H. (2022). The repetition of relative clauses in Mandarin children with Developmental Language Disorder. Language Acquisition, 30(2), 139–162. https://doi.org/10.1080/10489223.2022.2147839

Wanti, L. P., Somantri, O., & Kartiyani, T. (2024). Utilization of Augmented Reality Technology in Independent Speech Therapy Applications. Lontar Komputer: Jurnal Ilmiah Teknologi Informasi, 15(1), 1. https://doi.org/10.24843/LKJITI.2024.v15.i01.p01

Wilson, A. D., & Golonka, S. (2013). Embodied Cognition is Not What you Think it is. Frontiers in Psychology, 4. https://doi.org/10.3389/fpsyg.2013.00058

Tecnologie AI per la didattica

Gabriele Prosperi[1], Giulia Miglietta[2], Eleonora Miccoli[3], Mario Bochicchio[4]
[1] Università del Salento, Italia - gabriele.prosperi@unisalento.it
[2] Università del Salento, Italia - giulia.miglietta@unisalento.it
[3] Università degli Studi di Bari, Italia - eleonora.miccoli@uniba.it
[4] Università degli Studi di Bari, Italia - mario.bochicchio@uniba.it

ABSTRACT (ITALIANO)

Il paper presenta un sistema basato sull'Intelligenza Artificiale (IA) come supporto alla didattica. Utilizzando un modello Retrieval-Augmented Generation (RAG), il sistema integra fonti storiche selezionate su un argomento specifico, per fornire risposte accurate e pertinenti. Attraverso una Web App, studenti e docenti interagiscono con una piattaforma che supera i limiti dei Large Language Models (LLM) in contesti specialistici, offrendo un livello di approfondimento e un registro linguistico adattabili alle esigenze dell'utente. Il progetto sperimenta l'integrazione del sistema nella didattica con la metodologia della flipped classroom, promuovendo l'autonomia, la ricerca critica e lo sviluppo di competenze digitali. Inoltre, integra strumenti compensativi pensati per studenti con Disturbi Specifici dell'Apprendimento (DSA), garantendo accessibilità e inclusività. I risultati attesi mirano a raggiungere un equilibrio tra innovazione tecnologica e obiettivi educativi, migliorando sia la qualità dell'insegnamento sia la fruizione dei contenuti storici, con un impatto positivo sull'esperienza formativa complessiva.
Parole chiave: IA; RAG; didattica inclusiva; Media Education; flipped classroom

ABSTRACT (ENGLISH)

AI Technologies for Education.

The paper presents an Artificial Intelligence (AI)-based system designed to support education. Using a Retrieval-Augmented Generation (RAG) model, the system integrates selected historical sources on a specific topic to provide accurate and relevant responses. Through a Web App, students and teachers interact with a platform that overcomes the limitations of Large Language Models (LLMs) in specialized contexts, offering a customizable depth of content and linguistic register tailored to the user's needs. The project explores the integration of the system into education using the flipped classroom methodology, fostering autonomy, critical inquiry, and the development of digital skills. Additionally, it incorporates compensatory tools designed for students with Specific Learning Disorders (SLD), ensuring accessibility and inclusivity. The expected outcomes aim to achieve a balance between technological innovation and educational goals, enhancing both the quality of teaching and the accessibility of historical content, with a positive impact on the overall learning experience.
Keywords: AI; RAG; inclusive education; Media Education; flipped classroom.

1. INTRODUZIONE

L'impiego dell'intelligenza artificiale (AI) generativa e dei *Large Language Model* (LLM) nelle attività scolastiche è promettente in aree come il supporto agli studenti, l'assistenza agli insegnanti e la creazione di percorsi e/o attività di apprendimento personalizzate, ma presenta anche limiti e potenziali rischi. Tra i limiti piú significativi per la didattica è noto il basso livello di conoscenze normalmente disponibile agli LLM su temi specifici come la storia, le tradizioni e la cultura locale [Theophilou et al., 2023] che, peraltro, rappresentano un elemento identitario cardine per molti ambiti didattici. Tra i rischi più rilevanti è sicuramente da considerare la possibilità che gli LLM incorrano in allucinazioni o che generino informazioni errate [Groza & Marginean, 2023]. Per avviare una riflessione critica e consapevole sull'impiego degli LLM per scopi didattici, nell'ambito di una collaborazione tra l'Università del Salento ed una scuola pugliese è in corso di svolgimento una sperimentazione sull'impiego di una collezione di testi e materiale iconografico sulla storia del Salento in periodo alto-medioevale, per lo sviluppo di attività curricolari e di approfondimento. La sperimentazione prevede sia attività in aula, guidate dal docente e dai ricercatori, sia attività a casa, in cui gli studenti interagiscono autonomamente o in piccoli gruppi con il LLM. Il prototipo impiegato per la sperimentazione integra un LLM commerciale, gpt-3.5-turbo nel nostro caso, con un sistema di *information retrieval*, ovvero *vectorize.io*, che integra funzionalità di *chunking*, *embedding* e

vector store, per costituire un sistema di Retrieval Augmented Generation (RAG), che consente di superare alcuni dei limiti e ridurre alcuni dei rischi più noti dei LLM, proteggendo al tempo stesso la *privacy* e la riservatezza degli studenti e dei docenti inclusi nella sperimentazione. Nel caso specifico, *vectorize* consente di incorporare altre fonti bibliografiche (testi digitalizzati) rilevanti per la didattica curricolare e non disponibili agli LLM commerciali. Questa tecnica permette al LLM di accedere in modo controllato alla collezione documentale precedentemente inserita in *vectorize* dai ricercatori e dagli insegnanti, e non inclusa tra le fonti di addestramento del LLM, consentendo la generazione di risposte pertinenti e approfondite. Utilizzando *vectorize*, l'estensione della collezione documentale può essere effettuata direttamente dai ricercatori e dai docenti senza richiedere competenze di programmazione. Detta collezione documentale è derivata da un processo di selezione, digitalizzazione ed elaborazione di contenuti testuali e iconografici relativi alla storia del Salento. Una volta creata la collezione documentale si è concentrata l'attenzione sulla scrittura di prompt efficaci per ottenere uno stile di risposta adatto agli studenti e insegnanti inclusi nella sperimentazione. Per consentire agli studenti di interagire facilmente da casa con il sistema, si è scelto di utilizzare una Web App connessa al RAG. A valle dell'attuale fase di codesign, che vede la stretta collaborazione di ricercatori, insegnanti e studenti, il sistema sarà utilizzato per le attività curricolari oggetto della sperimentazione.

2. TECNOLOGIE UTILIZZATE

Come anticipato nell'introduzione, l'obiettivo è quello di creare un sistema in grado di rispondere alle domande degli utenti (studenti o docenti) incrociando fonti di conoscenza autorevoli. Per far questo, il sistema sfrutta la Retrieval Augmented Generation (RAG) [Alsafari et al., 2024] per generare risposte su fatti e avvenimenti della storia medievale salentina, che siano rilevanti e corrette dal punto di vista storico.
I modelli linguistici di grandi dimensioni (LLM) sono addestrati su vasti volumi di dati e utilizzano miliardi di parametri per generare output testuali per attività di elaborazione del linguaggio naturale come ad esempio rispondere a domande, riassumere o tradurre testi in diverse lingue.
Tuttavia, hanno carenze su molti argomenti relativi a domini specifici, possono presentare informazioni false quando non hanno la risposta, o presentare informazioni obsolete o generiche quando l'utente si aspetta una risposta specifica e attuale. Possono inoltre creare una risposta da fonti non autorevoli, o creare risposte imprecise a causa della confusione terminologica, in cui diverse fonti di formazione utilizzano la stessa terminologia per parlare di cose diverse.
Per tali ragioni in questo contesto utilizzare un LLM non è la soluzione ottimale per lo studio e l'approfondimento di tali temi. Un'alternativa valida è rappresentata dai sistemi RAG che estendono le capacità degli LLM senza la necessità di riaddestrare il modello.
I sistemi RAG consentono di migliorare l'output degli LLM in modo che rimanga pertinente, accurato e utile in vari contesti. Per far questo, recuperano informazioni da fonti di conoscenza autorevoli e predeterminate, passandole poi al LLM per fargli generare una risposta, consentendo di avere un maggiore controllo sull'output generato.
In generale i sistemi RAG sono basati su tre componenti fondamentali: un LLM, un database vettoriale e un componente per l'inserimento e il recupero delle informazioni nel database vettoriale che, a partire dalla richiesta dell'utente, estrae le informazioni dal database sulla base di criteri di prossimità semantica.
Operativamente, la collezione documentale, composta da testi digitalizzati selezionati, viene suddivisa in frammenti più brevi (indicativamente: 1000 caratteri) detti *chunks*, che sono poi associati a rappresentazioni vettoriali, dette *embeddings*, che consentono mediante calcoli opportuni (distanze vettoriali) eseguiti dal database vettoriale (*vector store*) di valutare la prossimità semantica tra le domande testuali poste dagli studenti al sistema RAG e i *chunks* di testo (le fonti documentali) semanticamente più vicini a tale domanda. Nell'esperimento qui descritto, queste funzioni sono già integrate in *vectorize* e non richiedono la scrittura di software per il loro utilizzo.
Un sistema così strutturato può sfruttare le competenze linguistiche del LLM per "conversare" con studenti e docenti fondando le proprie risposte su documenti verificabili e precedentemente selezionati dai docenti stessi (o dagli studenti opportunamente guidati da esperti del settore), anziché sulla base del solo patrimonio di conoscenze disponibile al LLM in base al processo di addestramento del LLM stesso. Nel ciclo domanda-risposta, infatti, le domande vengono convertite nella corrispondente rappresentazione vettoriale, questa viene utilizzata per eseguire una ricerca di pertinenza nel database vettoriale che individua e ordina le fonti in base alla rilevanza (ovvero alla distanza vettoriale tra domanda e risposta) e restituisce solo i frammenti più rilevanti al LLM, che si occupa della trasposizione discorsiva corretta dei concetti contenuti nei frammenti.

Opportune tecniche di progettazione dei *prompt* (*prompt engineering*) consentono al LLM di adeguare il registro comunicativo dei frammenti al contesto didattico e al tipo di attività richiesta dal docente.

3. SELEZIONE DEI CONTENUTI E PROMPT ENGINEERING

L'efficacia del sistema RAG sviluppato dipende direttamente dalla combinazione tra la struttura dei prompt inseriti e la selezione accurata dei testi di riferimento, inseriti in un *repository* contestuale entro cui il sistema rintraccia le informazioni da restituire agli utenti.
Per garantire la qualità delle risposte, il *repository* è stato accuratamente composto da fonti storiche selezionate dal team sulla base di criteri quali l'autenticità, l'affidabilità e l'accuratezza storica, la rilevanza culturale e tematica. Il database include al suo interno fonti primarie e ricerche accademiche focalizzate sull'oggetto di studio della sperimentazione, ovvero sulla storia del Salento, dei suoi castelli e delle figure storiche più rilevanti.
Oltre alla costruzione di un archivio di contenuti, grande importanza ha avuto la fase di *prompt engineering* – ovvero di creazione di prompt progettati per migliorare le prestazioni del sistema – che comporta comunemente l'uso di conoscenze specifiche del dominio di riferimento, la fornitura di esempi del risultato desiderato e l'utilizzo di termini chiave per rendere la risposta efficace. Sono stati condotti dei test preliminari sul sistema RAG per verificare che lo stile delle risposte fosse quello desiderato e assicurarsi che le informazioni restituite fossero affidabili.
Durante il processo di generazione, il sistema RAG recupera fino a cinque (nel nostro caso) documenti rilevanti dal *repository*, che vengono poi utilizzati come contesto per costruire una risposta articolata e pertinente. Questo approccio fa sì che la capacità di sintesi del sistema sia bilanciata con la profondità delle informazioni, riducendo al minimo il rischio di includere contenuti non rilevanti, falsi o fuorvianti.
I prompt sono elaborati per rispondere in modo specifico alle esigenze degli utenti, siano essi studenti o docenti, e si differenziano in base alle possibili domande inviate al sistema.
Il processo non è solo tecnico, ma coinvolge anche una dimensione pedagogica: l'utilizzo delle tecnologie digitali nei contesti educativi non può prescindere dalle dinamiche di soggettivazione e dalle pratiche peer-to-peer che si sviluppano tra gli studenti [Tirocchi & Serpieri, 2020]. Similmente, una user experience efficace nei sistemi AI per l'educazione dovrà bilanciare aspetti tecnologici e pedagogici, ovvero creare interfacce che non siano solo user-friendly, ma che facilitino anche un'esperienza di apprendimento consapevole [Farinacci, 2024]. A questo livello sono stati perciò definiti il linguaggio e il tono che il sistema deve adottare, specificando nel prompt alcuni esempi di risposta, suggerendo l'approccio narrativo, orientando il sistema verso una contestualizzazione storica e tematica, e richiedendo, all'interno della risposta, la proposta di ulteriori approfondimenti (Figura 1).

Rispondi alle domande specifiche sull'epoca o sui personaggi storici legati ai castelli, in base al contesto fornito, adottando uno stile narrativo che bilanci precisione storica e coinvolgimento emotivo, come se stessi guidando un ascoltatore in un viaggio nel passato. Offri un'introduzione che contestualizzi il tema storico o il personaggio all'interno dell'epoca di riferimento, trasmettendo l'atmosfera del tempo. Ad esempio: "Era l'anno del Signore 1370 quando Maria d'Enghien nacque in una terra segnata da guerre e alleanze, destinata a diventare una delle figure più influenti del suo tempo." Fornisci informazioni dettagliate sugli eventi, le figure storiche o i luoghi coinvolti,

FIGURA 1 - parte di prompt per suggerire stile narrativo e contestualizzazione storica

A titolo d'esempio, uno dei prompt realizzati comunica al sistema di assumere il ruolo di un personaggio storico per rispondere alla richiesta dell'utente (ovvero alla domanda di uno studente o al dialogo con un attore in scena) in modo interattivo, immergendo quindi il fruitore in una narrazione amodale [Ryan, 2006]. Similmente, il prompt adottato per le risposte da fornire ai docenti include l'indicazione di dare suggerimenti su come applicare i contenuti alla didattica in classe, attraverso attività laboratoriali o domande da porre agli studenti per stimolare il dibattito (Figura 2).
La personalizzazione dei prompt si estende, infatti, alla tipologia di utente, come sarà meglio considerato nella descrizione della user experience: sono stati progettati prompt specifici per bambini sotto i 14 anni, studenti delle scuole superiori (14-19 anni) e docenti, permettendo così a ogni categoria di ricevere contenuti con un registro linguistico e un livello di approfondimento adattati all'utilizzatore. L'applicazione adatta, in questo modo, il livello di complessità e il registro linguistico in funzione dell'utente, con la

possibilità di accedere a spiegazioni più dettagliate e focalizzate sull'approfondimento critico e interdisciplinare per gli studenti o, per i docenti, di accedere a funzioni avanzate per la didattica. In entrambi i casi, i prompt vanno incontro all'esigenza specifica individuata: se, ad esempio, si corre il rischio di un mancato approfondimento degli argomenti da parte dello studente, ciò viene mitigato grazie a un prompt progettato per incoraggiare la prosecuzione dell'utilizzo, indicando al sistema di porre nuove domande nella sua risposta, o lasciando aperto il dialogo per ulteriori connessioni tematiche.

Rispondi alla domanda in base al contesto sopra descritto utilizzando un linguaggio formale e ricco di dettagli, adatto a un pubblico adulto di docenti. Offri una spiegazione storicamente accurata e ben strutturata, combinando descrizioni analitiche con approfondimenti che evidenzino le connessioni tra il tema trattato e i concetti più ampi di storia, arte, filosofia o scienza.

Ad esempio: "Oltre a essere un esempio di fortificazione, il Castello di Acaya è anche un simbolo dell'ingegneria rinascimentale, dove la funzionalità difensiva si intreccia con la visione estetica tipica dell'epoca."

Analizza il tema in modo critico, mettendo in evidenza eventuali controversie storiografiche o prospettive interdisciplinari che possano stimolare la riflessione e il dialogo tra colleghi.

Collega la spiegazione a esempi concreti o esperienze pratiche, su come il tema possa essere utilizzato per coinvolgere gli studenti in aula, attraverso attività interdisciplinari o metodologie didattiche e proponi un esempio di attività per la classe.

Ad esempio: "Un'attività didattica potrebbe prevedere che gli studenti analizzino le mappe storiche dei castelli e le confrontino con quelle moderne, discutendo come l'evoluzione tecnologica abbia trasformato le strategie di difesa. Questo esercizio potrebbe includere una ricerca sulle differenze nei materiali utilizzati e una riflessione sull'uso contemporaneo degli stessi luoghi, come musei o attrazioni culturali."

Concludi con una riflessione aperta, che incoraggi ulteriori approfondimenti, lasciando spazio per collegare il tema storico al dibattito in aula.

FIGURA 2 – frammento del prompt con richiesta di attività didattiche

La sperimentazione si propone, così, di sviluppare nei giovani utenti una comprensione critica della natura di questo nuovo medium. In linea con i recenti studi sulla Media Education [Buckingham, 2020], l'intento è di promuovere la consapevolezza e la riflessione sull'uso dei media, integrando l'educazione *con* i media e l'educazione *ai* media. Nell'ambito dell'esperimento, ciò si concretizza attraverso l'uso di tecnologie che trasformano l'apprendimento in un processo dinamico e narrativo, consentendo agli studenti, contemporaneamente, di interagire con i contenuti storici in modo critico e sviluppare nuove competenze digitali. Similmente, per i docenti, l'integrazione di questo nuovo strumento offre l'opportunità di una meta-riflessione sulle metodologie didattiche adottabili.

4. DESCRIZIONE DELLA USER EXPERIENCE: STRUMENTI E METODI

Per l'integrazione del sistema nella didattica si è scelto di utilizzare la cornice metodologica della *flipped classroom* (classe capovolta). Come è noto, si tratta di una strategia didattica che trasforma il modello tradizionale di insegnamento-apprendimento nelle scuole, rendendo le studentesse e gli studenti protagonisti del loro percorso di conoscenza e ristrutturando l'ambiente di apprendimento alla luce dell'innovazione tecnologica [INDIRE, 2016]. In particolare, l'attività proposta dal progetto RASTA a una classe di studenti liceali ha come obiettivo l'apprendimento di personaggi, luoghi ed eventi storici inerenti al medioevo salentino, con particolare riferimento alla storia medievale e rinascimentale del Castello Carlo V di Lecce, del Castello di Acaya e del Castello di Corigliano d'Otranto, i medesimi temi trattati nei documenti che costituiscono la collezione documentale utilizzata per il sistema RAG.
A tal fine, si è pensato di strutturare l'attività didattica in più fasi. In una prima lezione, il docente ha il compito di fornire agli studenti un'introduzione generale dell'argomento nonché della metodologia che sarà utilizzata nello svolgimento dell'attività. Successivamente, il docente ha il compito di dividere la classe in gruppi, assegnando a ciascuno di essi un tema specifico da studiare in autonomia e al di fuori dell'orario scolastico. In questo caso, si è scelto di formare cinque gruppi, ciascuno dei quali si è occupato della ricerca, dello studio e dell'approfondimento di un argomento specifico correlato alla storia del Castello Carlo V. A distanza di un tempo definito e concordato durante la lezione, le conoscenze acquisite sono infine restituite in forma di presentazione scritta e orale da ciascun gruppo all'intera classe, secondo le modalità scelte dai gruppi stessi, ma ricorrendo senz'altro alle tecnologie dell'informazione e della

comunicazione, come strumenti in grado di integrare e ampliare le possibilità offerte dai tradizionali strumenti di studio e ricerca.
Nello specifico, come ausilio all'attività di ricerca, ogni gruppo ha a disposizione la Web App, attraverso la quale è possibile avviare una chat con il sistema RAG che risponde in tempo reale a domande poste dall'utente/studente sui temi di storia locale di cui sopra.
Il sistema è in grado di adattare il registro in funzione del grado di formazione dell'utente (studente o docente) e dell'età dello studente (ad esempio età compresa tra i 14 e i 19 anni o meno di 14 anni). Rispetto al recupero dei dati per mezzo di un qualunque motore di ricerca, il sistema consente agli studenti di reperire quante più informazioni possibili sul loro oggetto di studio specifico, con il vantaggio di fornire risposte basate su fonti già selezionate *ab origine* da un team di storici e, dunque, dotate di un grado maggiore di affidabilità rispetto alla ricerca, il più delle volte randomica, operata da uno studente sul web. Inoltre, le risposte sono corredate dai riferimenti bibliografici ai quali il sistema attinge per formulare il testo restituito all'utente.
In questo contesto, la metodologia didattica della *flipped classroom*, oltre a favorire nelle studentesse e negli studenti lo sviluppo e il consolidamento di tutta una serie di abilità per le quali tale strategia è nota in ambito scolastico (autonomia nello studio, public speaking, team working e digital literacy), è arricchita dai benefici specifici che trae dall'utilizzo della Web App RASTA. Difatti, l'utilizzo di tale sistema, da un lato funge da filtro per le informazioni a cui lo studente può avere accesso, agevolando la raccolta di informazioni utili per l'apprendimento in autonomia, e dall'altro lato si presta a essere uno strumento in grado di fornire alle studentesse e agli studenti una prima formazione sulla ricerca sulle fonti e, più in generale, di introduzione alla ricerca scientifica. Le risposte possono includere anche approfondimenti interdisciplinari, arricchendo il racconto storico con informazioni relative all'arte, all'architettura, alla musica e alla letteratura. Inoltre, il sistema funge da supporto anche per i docenti, offrendo un ausilio nell'organizzazione e nella pianificazione delle lezioni. In particolare, è in grado di suggerire proposte e strategie didattiche innovative che possono aiutare i docenti ad affrontare l'argomento trattato in modo creativo, rendendo la lezione in classe più dinamica e interattiva, creando un ambiente di apprendimento più stimolante e partecipativo per gli studenti, e dunque un processo di apprendimento più efficace. Inoltre, l'applicazione favorisce l'inclusività, in quanto è pensata anche come strumento compensativo per studenti che presentano disturbi specifici dell'apprendimento (DSA). Le risposte, infatti, possono essere elaborate in vari formati – sintesi visiva e sintesi vocale – oltre quello testuale. La storia può essere raccontata tramite immagini (fornite anch'esse dagli storici e conservate nel *repository* o create con l'intelligenza artificiale) o fumetti, generati anch'essi in automatico con l'IA. È possibile, inoltre, fornire all'utente un racconto schematico o sotto forma di mappa concettuale, per facilitare l'apprendimento alle studentesse e agli studenti che presentano difficoltà nella lettura o nella memorizzazione. Allo stesso modo, la sintesi vocale dell'applicazione consente di ascoltare il racconto piuttosto che leggerlo, agevolando così l'acquisizione delle conoscenze degli alunni con dislessia.

5. CONCLUSIONI E SVILUPPI FUTURI

Il progetto propone un approccio innovativo per integrare le tecnologie di AI nella didattica, puntando a superare i limiti dei Large Language Models grazie al sistema RAG. La fase di progettazione ha permesso di definire un framework che bilancia rigore tecnologico e obiettivi educativi, garantendo flessibilità, accessibilità e inclusività. Il progetto prevede perciò una valutazione del grado di efficacia della sperimentazione all'interno del percorso di apprendimento. Nello specifico, si intende valutare, da un lato, il grado di apprezzamento del corpo docente in relazione all'implementazione di tale tecnologia nella costruzione dell'attività didattica e, dall'altro lato, la qualità dell'acquisizione dei contenuti da parte delle studentesse e degli studenti. Sono stati perciò presi in considerazione l'accessibilità delle informazioni in relazione alla storia locale veicolate dall'applicazione, la flessibilità della stessa nel fornire i contenuti diversificando le risposte agli utenti, l'inclusività nel rispetto delle diverse modalità di apprendimento e l'efficacia dello strumento in termini di sviluppo di una maggiore consapevolezza nell'uso di tecnologie di AI e di un crescente senso critico nel condurre ricerche in autonomia. A tal fine, al termine delle attività, sono stati somministrati questionari alla docente e agli studenti per la valutazione dell'esperienza.
La prima sperimentazione, condotta in collaborazione con il Liceo classico e musicale "G. Palmieri" di Lecce, ha dimostrato l'efficacia dell'integrazione del sistema RAG nella didattica. A seguito di una prima giornata introduttiva e dopo una settimana di utilizzo della Web App, l'incontro di restituzione ha avuto luogo nel Castello Carlo V di Lecce, oggetto delle ricerche degli studenti. Quella che inizialmente era stata proposta come *flipped classroom* si è, di fatto, trasformata in un "*flipped tour*", in cui gli studenti hanno

svolto il ruolo di guide turistiche, moderate e supervisionate dai ricercatori presenti, dalla docente e da una guida esperta che ha accompagnato la classe durante l'esperienza.
Nel sondaggio, sebbene gli studenti e la docente abbiano riscontrato alcuni limiti nel livello di approfondimento dei contenuti forniti dalle risposte, gli utenti hanno interpretato questa mancanza come uno stimolo alla ricerca autonoma, trasformando il limite in un'opportunità di apprendimento. I futuri sviluppi dell'applicazione terranno fortemente conto di questa dinamica, sia definendo consapevolmente un confine informativo che favorisca la curiosità e l'approfondimento degli utenti, sia incentivando e migliorando negli studenti la capacità di porre al sistema domande precise, accurate e dettagliate. Questa prima valutazione conferma la validità dell'integrazione di strumenti innovativi in contesti educativi e apre la ricerca a sviluppi futuri, talvolta inaspettati.

RINGRAZIAMENTI

Le attività di ricerca descritte nell'articolo sono parzialmente supportate dal progetto RASTA, finanziato dal Ministero dell'Università e della Ricerca con risorse del PON "Ricerca e Innovazione" 2014–2020 e con Fondi FSC ai sensi e per effetto dell'Art. 13 dell'Avviso per la presentazione di progetti di Ricerca Industriale e Sviluppo Sperimentale nelle 12 aree di specializzazione intelligente individuate dal PNR 2015–2020 e dell'Art. 1 del DD n. 551 del 27 aprile 2020.

BIBLIOGRAFIA

Alsafari, B., Atwell, E., Walker, A., & Callaghan, M. (2024). Towards effective teaching assistants: From intent-based chatbots to LLM-powered teaching assistants. Natural Language Processing Journal, 8, 100101.

Buckingham, D. (2020). Un manifesto per la media education. Milano: Mondadori Università.

Farinacci, E. (2024). Film and Audiovisual Education in the Artificial Intelligence Era. Cinergie – Il cinema e le altre arti, n. 26, 121-133.

Groza, A., & Marginean, A. (2023). Brave new world: ai in teaching and learning. In ICERI2023 Proceedings. IATED, 8706-8713.

INDIRE. Linee guida "Flipped classroom", (2016). https://pheegaro.indire.it/uploads/attachments/1963.pdf.

Ryan, M. L. (2006). Avatars of Story. Minneapolis: University of Minnesota Press.

Theophilou, E., Koyutürk, C., Yavari, M., Bursic, S., Donabauer, G., Telari, A., ... & Ognibene, D. (2023, November). Learning to prompt in the classroom to understand AI limits: a pilot study. In International Conference of the Italian Association for Artificial Intelligence. Cham: Springer Nature Switzerland, 481-496.

Tirocchi, S., & Serpieri, R. (2020). Piattaforme, adolescenti e forme di soggettivazione. Mediascapes journal 16/2020, 207-222.

DEA - An Innovative Technological Tool for Personalized Linguistic Training for Italian Children with Developmental Dyslexia

Marta Tagliani[1], Maria Vender[2], Giulia Valcamonica[3], Giovanni Caleffi[4], Franca Garzotto[5], Denis Delfitto[6]
[1]Università di Verona, Italy, marta.tagliani@univr.it
[2]Università di Verona, Italy, maria.vender@univr.it
[3]Politecnico di Milano, Italy, giulia.valcamonica@polimi.it
[4]Politecnico di Milano, Italy, giovanni.caleffi@polimi.it
[4]Politecnico di Milano, Università degli Studi di Milano-Bicocca, Italy, franca.garzotto@polimi.it
[5]Università di Verona, Italy, denis.delfitto@univr.it

ABSTRACT (ENGLISH)

In this paper, we illustrate a research project using advanced technological tools to improve the linguistic and reading skills of Italian children diagnosed with developmental dyslexia (DD) aged between 8 and 11. Besides the well-known reading and spelling difficulties, DD is characterized by marked linguistic deficits especially at phonological and morphological level, negatively impacting the child's reading abilities and directly affecting text comprehension. Recent studies have shown the effectiveness of linguistic interventions in enhancing phonological and morphological awareness in DD children and improving their reading skills. However, access to such linguistic programs is not free from important barriers; in particular, high costs and difficulties in customizing training programs make it imperative to implement new accessible technological tools. In this context, we developed the DEA (Dyslexia Exercises Application) training program, a modular web application that offers interactive and highly personalized linguistic activities. The ongoing study involves 30 DD Italian children who participate in a 10-week training three times a week for 20-minute sessions. The program includes phonological and morphological exercises with increasing difficulty created with AI support. It allows the therapist to monitor the child's progress and provide a tailored reinforcement program based on their needs. DEA provides individualized and highly accessible training programs to support reading development in children with DD, helping the therapist create specific learning tracks based on the participant's linguistic and cognitive skills.
Keywords: Developmental Dyslexia; Linguistic Training; Morphological and Phonological Awareness; Technology in Dyslexia Training; Interactive Exercises.

ABSTRACT (ITALIANO)

DEA – Uno strumento tecnologico innovativo per il potenziamento linguistico personalizzato per bambini italiani con dislessia evolutiva.

In questo articolo illustriamo un progetto di ricerca che utilizza soluzioni tecnologiche avanzate per migliorare le competenze linguistiche e le abilità di lettura di un gruppo di bambini con diagnosi di dislessia evolutiva (DE) di età compresa tra gli 8 e gli 11 anni. Oltre alle ben note difficoltà nella letto-scrittura, la DE è caratterizzata da marcati deficit a livello fonologico e morfologico, che influenzano negativamente le abilità di lettura con conseguenze dirette anche sulla comprensione del testo. Studi recenti hanno dimostrato l'efficacia di interventi linguistici mirati a potenziare la consapevolezza fonologica e morfologica nei bambini con DE per migliorare le loro abilità di lettura. Tuttavia, l'accesso ai percorsi di potenziamento linguistico non è esente da barriere importanti: in particolare, costi elevati e difficoltà nella personalizzazione dei programmi rendono necessaria l'implementazione di nuovi strumenti tecnologici di marcata accessibilità. In questo contesto, abbiamo sviluppato un programma di training linguistico DEA (Dyslexia Exercises Application), una applicazione web modulare che offre attività linguistiche interattive e altamente personalizzate. Lo studio è attualmente in corso e coinvolge 30 bambini con DE, che partecipano a 10 settimane di training, tre volte a settimana per sessioni di circa 20 minuti. Il programma include esercizi per il potenziamento fonologico e morfologico con difficoltà crescente creati con il supporto dell'AI e permette al terapista di monitorare i progressi del singolo partecipante in modo da fornire un programma di potenziamento personalizzato sulla base delle sue esigenze. DEA consente interventi personalizzati e accessibili per supportare lo sviluppo della letto-scrittura nei bambini con DE, permettendo al terapista di creare percorsi specifici sulla base delle competenze linguistiche e cognitive del partecipante.
Parole chiave: Dislessia Evolutiva; Potenziamento Linguistico; Competenza Morfologica e Fonologica; Tecnologia per la Dislessia; Esercizi Interattivi.

1. INTRODUCTION

Specific Learning Disabilities (SLD) are neurodevelopmental disorders impacting the development of academic skills and affecting 5-15% of children and adolescents globally(DSM-5, First, 2014; 3.5% in Italy, Pido' et al., 2021). Developmental dyslexia (DD) is the most prevalent of these disorders and is characterized by remarkable impairments in the automatization of academic skills such as reading, writing, and arithmetic skills despite adequate intellectual ability, absence of neurological deficits and appropriate literacy exposure (DSM-5). Besides severe reading and spelling difficulties, DD is characterized by a wide range of deficits affecting different linguistic domains. These deficits needs to be seriously considered if we want to have a comprehensive picture of this disorder. Specifically, dyslexics have difficulties in the phonological domain, particularly in activities of phoneme recognition and manipulation (Bishop & Snowling, 2004; Ramus & Szenkovits, 2008; Vellutino et al., 2004; Vender & Melloni, 2021) as well as in the morphological domain. The persistent reading difficulties observed in children with DD may negatively impact on the emotional sphere and social skills of these children and generally lead to an increase rate of school exclusion and dropout (Cremonesi et al., 2017), with important consequences also from a socioeconomic perspective in terms of access to higher education and job opportunities. Developing suitable tools and interventions that support the child in their language development is thus of the utmost importance. Importantly, in fact, both decoding and linguistic abilities can be successfully trained, as shown by a growing number of studies conducted on both children and adults, highlighting the contribution of both phonological and morphological activities in supporting reading development (Galuschka et al., 2014; Bowers et al., 2010; Vender et al., 2022). Phonological awareness, which can be defined as the ability to identify and manipulate sound units (Blachman, 1991) plays indeed a crucial role in literacy acquisition and in developing the child's reading skills, as it drives the automatization of grapheme-phoneme conversion rules (Snowling et al., 2020) and is widely recognized as a strong predictor of reading skills development. More recently, psycholinguistic research has observed how morphological awareness, the ability to identify and manipulate morphemes, also aids reading development, by positively affecting both decoding and text comprehension (Arnbak & Elbro, 2000; Casalis et al., 2004; Kuo & Anderson, 2006; Levesque & Deacon, 2022; Pacheco & Goodwin, 2013). In fact, morphologically complex (i.e., affixed) words are typically read faster and more accurately than simple words matched for length and frequency by both typical and DD readers of any age (Burani et al., 2008; Carlisle & Stone, 2005; Levesque & Deacon, 2022). Despite the attested deficits in both derivational and inflectional domains (Casalis et al., 2004; Vender, 2024; Georgiou et al., 2022), dyslexic and less-skilled readers seem able to resort to their morphological abilities to compensate for their reading difficulties (Burani et al., 2008; Marcolini et al., 2011; Traficante et al., 2011). Intervention studies targeting the development of morphological skills have indeed been reported as particularly effective, leading to improvements in both decoding and comprehension (Bowers et al., 2010; Levesque & Deacon, 2022).

However, there are still significant barriers to accessing intervention, related to factors such as length and expensiveness, and to the difficulty of extending the intervention beyond the clinical context. It has also been observed that to be effective, training should be personalized based on the child's needs, which is often hardly feasible (Catania et al., 2021). To this end, interactive technologies can provide an important asset. In many cases, they have been found to be more effective than traditional interventions in the acquisition of language skills, especially when based on the use of regular devices such as computers, tablets, or smartphones (Crovari et al., 2021). Computerized training programs facilitate the development of activities that have proven particularly effective in automating and speeding up decoding, such as the Reading Acceleration Paradigm, which consists of structured reading exercises in which children are encouraged to read faster than their usual rate (Breznitz, 1997, 2001; Breznitz & Share, 1992). Current technologies also permit the implementation, in an extremely ecological manner, of a range of activities that are traditionally recognized in literature as effective for strengthening both decoding and comprehension. A web-based reading intervention integrating phonological and morphological activities in a technological environment can thus be particularly effected, as shown by a recent study (Vender & Delfitto, 2024) reporting significant gains in reading speed and decoding as well as in phonological skills in adults with dyslexia. Within the Italian intervention framework, a number of digital applications have been developed to support children with SLD. One notable example is EdiTouch[1], a specialized tablet offering several activities for students from primary to secondary school: among others, it includes a text-to-speech eBook reader, a talking calculator, and a talking keyboard that offers real-time spelling feedback

[1] https://www.tabletascuola.net/ (cons. 09.04.2025).

and correction. The Leggimi App[2] automatically applies a dyslexia-friendly font to texts to enhance readability. TurboLettura[3] is a gamified reading app using voice recognition to analyze the child's reading-aloud performance in real time. It provides instant feedback and corrects errors in a game-like progress manner. Finally, RIDInet[4] is a web-based platform that includes customizable apps for reading, writing, math, language, and executive functions. Therapists can monitor the individualized training sessions remotely, integrating activities such as reading short texts with comprehension questions, timed syllable and word reading, and spelling dictation.

2. THE CURRENT STUDY

The aim of the current work is to develop a cutting-edge technological tool providing favorable ecological conditions for enhancing the effectiveness of language and reading training programs for children diagnosed with developmental dyslexia. To do so, we developed a linguistic training program tailored to the linguistic and cognitive profile of each child participant, aimed at specifically addressing their needs. While other platforms, as those mentioned above, already use interactive and multisensory approaches to dyslexia intervention, our tool focuses specifically on improving the child's literacy skills through targeted, tailored linguistic activities aimed at enhancing phonological and morphological competence yielding positive effect of both reading and spelling. For the first time, to our knowledge, it also includes an innovative focus on derivational morphology as in important component for improving decoding. At both levels of linguistic analysis, we offer a range of customized activities designed to strengthen the child's phonological and morphological awareness, which, as shown in the literature, can effectively support the child in the decoding process. The training involves the independent use of a web application (on a computer, tablet or smartphone) for three days a week for a total of ten weeks. Each session lasts approximately 15-20 minutes.

3. METHODOLOGY AND PROPOSED ACTIVITIES

Participants: The training is currently ongoing and involves 30 Italian children diagnosed with DD, aged from 8;1 to 11;10 years (Mean age = 10;1, SD = 1;1, Male participants = 10), recruited through the ULSS Scaligera 9 Verona (Italy). The same number of children with DD will be part of a control group and will therefore not take part in the training program, in order to verify the effectiveness of the intervention. We expect to have the first results of the study comparing the two populations by May 2025.

Design: Before the start of the training, all participants (i.e., both control and experimental groups) are tested in a series of tasks to assess their reading proficiency and linguistic abilities at phonological and morphological levels to gather precise information about their initial reading and linguistic profile. The same tasks will also be administered after the training to verify the presence of improvements related to the intervention.

Reading and linguistic assessment: As for reading skills, we administered the DDE-2 (Sartori et al., 2007) and MorfoLet (Tagliani et al., in preparation), an original test of word and nonword reading to examine the role of morphological analysis in the decoding process. We also tested the children's phonological (Vender et al., 2020; Vender & Delfitto, 2024) and morphological skills. For the latter, we administered a series of novel morphological tests designed for Italian derivational morphology (adapted from Piccinin & Dal Maso, 2023). Participants were also evaluated in their lexical and grammatical competence by means of PPVT (Dunn & Dunn, 1965) and TROG-2 (Bishop, 2009) tasks, as well as in their reading comprehension (Prove MT-3, Pra Baldi et al., 2010).

Training materials: Since both phonological and morphological skills play a crucial role in improving reading and spelling abilities, the training activities we developed aimed at enhancing these two linguistic dimensions. All activities were designed to have an increasing level of difficulty, which can be represented by syllabic complexity, number of syllables or presence of complex orthographic groups in phonological exercises, or by semantic and orthographic transparency of derived words and their frequency in the morphological ones. Participants can proceed to the next level once they achieve 80% accuracy in the task. Otherwise, the level must be repeated. Immediate feedback is provided in all exercises, and if an error is made, the correct solution is shown.

[2] https://leggimi.app/ (cons. 09/04/2025).

[3] https://onehealthvision.com/app/turbolettura-trattare-dislessia-difficolta-lettura/ (cons. 09/04/2025)

[4] https://www.anastasis.it/ridinet/ (cons. 09/04/2025)

The training starts with the phonological activities until about halfway through the program. The first activity is *Rapid reading*, which involves the modulation of on-screen persistence of given stimuli ranging from 200 to 2000ms. Children are shown a stimulus (syllable, word or nonword) displayed for a limited time and once it disappears, they are asked to rewrite it. The on-screen persistence is tailored to each participant's reading proficiency, as assessed during the preliminary testing phase, and gradually decreases through the levels. In *Pick the correct word*, children are presented with two words and must choose the one spelled correctly. This exercise is also offered in a version with auditory aid. In *Dictation*, participants listen to a stimulus (syllable, word or nonword) and are asked to write it correctly. In *Listen and choose the right match*, children listen to an auditory stimulus (syllable, word or nonword) and must select the correct writing among four options, including phonologically or visually similar competitors. Similarly, in the *Listen, look and choose* task, participants must select the correct writing form upon hearing the stimulus, but they are also aided by the presence of a picture representing the target word (see Figure 1-left panel).

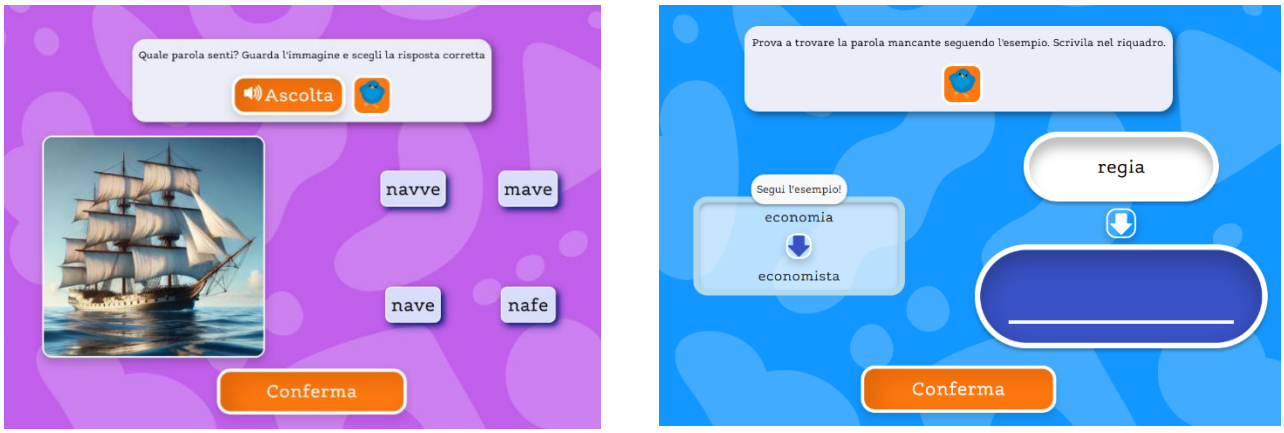


Figure 1. Display screens of two activities: the *Listen and choose the right match* task (left panel) and the *Morphological pattern* task (right panel).

Morphological training is introduced by the 15th session. We decided to focus on 12 high-frequency Italian suffixes (Gaeta & Ricca, 2003): *-ista*, *-tore*, *-mento*, *-zione*, *-aio*, *-eria*, *-aggio*, *-ismo*, *-ità*, *-oso*, *-bile*, and *-ale*. Suffixes are introduced gradually and each time new suffixes are added, they are presented in a short video featuring the character *Ispettore Morfolix*, who explains the meaning of the specific suffix. The first morphological activity resembles the *Rapid reading* task described above. The only difference is in the type of items presented, which in this case are derived words of 4 or 5 syllables. In *Word sort*, children must determine whether a word is simple or derived (e.g., *chitarra* 'guitar' vs. *chitarrista* 'guitarist') and place it into the correct basket. In *Word Building*, children are presented with a series of bases and suffixes and are asked to combine them to form the correct derived word (e.g., *dentista*, 'dentist', composed by the root *dent-* 'tooth' and the agentive suffix '-ista'). In *Word puzzle*, they are asked to identify both simple and derived words and pair them appropriately (e.g., *chitarra-chitarrista*,'guitar – guitarist'; *costruire-costruzione*, 'construct-construction'). In *Morphological pattern*, participants are shown two words belonging to the same semantic family in a base/derived word relationship (e.g., auto-autista, 'car-driver'). Based on the morphological relationship established between these two words, children must complete another two-word set for which only one element is given to form a similar morphological pair (e.g., musica-…, 'music'-…, missing word: musicista 'musician') – see Figure 1-right panel. Finally, in the *Find the Word in the Text* activity, participants are presented with short sentences and asked to identify and write down the derived word found in the text.

4. APPARATUS

To facilitate language training and supply exercises in an engaging way, we have developed DEA (*Dyslexia Exercises Application*). DEA is a web-based application built on a modular architecture, designed to manage and deliver interactive exercises. The platform allows therapists to create and assign training sessions to participants, monitor their progress and support tailored interventions. Participants, instead, can engage in their assigned training sessions through the application.

Training session: All the training sessions are designed to be conducted with the aid of interactive exercises provided through activities. Activities are composed of at least one exercise, and each exercise is divided into several tasks. To create a session, therapists (experimenters, in our study) must name the session, select the exercises that must be executed during the session and specify the corresponding user (see Figure 2). Users can be also added after the creation of the activity. Therefore, the training session is a sequence of exercises, and at the same time exercises are a sequence of tasks.

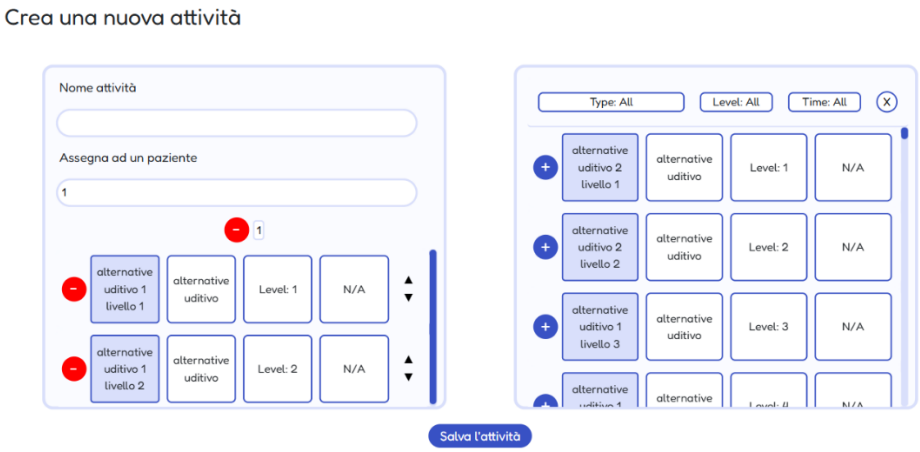


Figure 2: Display screen of the activity creation, here therapists set the activity name, exercises and the assigned users.

User flow: Users log into the application by using their username and password. On the first screen after the login, a page displays all the assigned uncompleted sessions, along with the corresponding completion rate (see Figure 3). Users can start and complete the sessions in any order and can also pause or resume them as needed. Once the user selects a session, the training experience will start. While performing tasks, users can at any time ask for help from the system by clicking on the bird icon, answer the task or quit. If a user asks for help, the system will play an audio that contains information about the task and how to interact with the system to complete it. The session progress is saved at the end of each single task: in this way when a user resumes their session they'll continue from the first uncompleted task. This allows users to break their activities when overwhelmed and rest until they're motivated to resume the session. For each task the system displays the instruction text, its position within the exercise and the exercise's difficulty level (see Figure 4 – right panel).

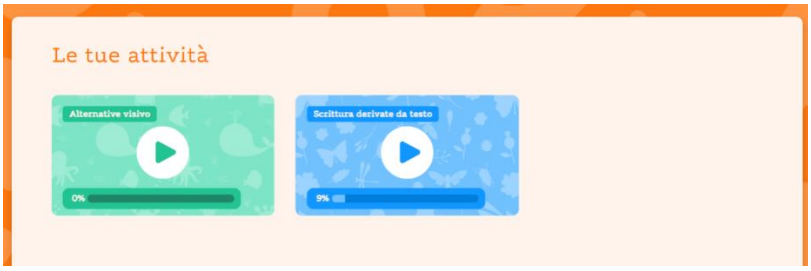


Figure 3: Display screen of the uncompleted activities of the logged user.

To complete an activity, users must complete all the exercises in it, and to complete an exercise the users must answer all the tasks that compose it. When the user answers correctly to a task, the system displays positive visual feedback (see Figure 4 – top left panel) and then moves to the next step. This could be the next task within the same exercise, the next exercise if the current one was completed, or the home screen if all the scheduled exercises have been completed. If the users provide a wrong answer, the system will show negative visual feedback (see Figure 4 – bottom left panel); in some cases, users can repeat a task for a limited number of times (e.g., in the *Word Puzzle* activity), while in others the tasks will not be repeated, and the system will move forward to the next step.

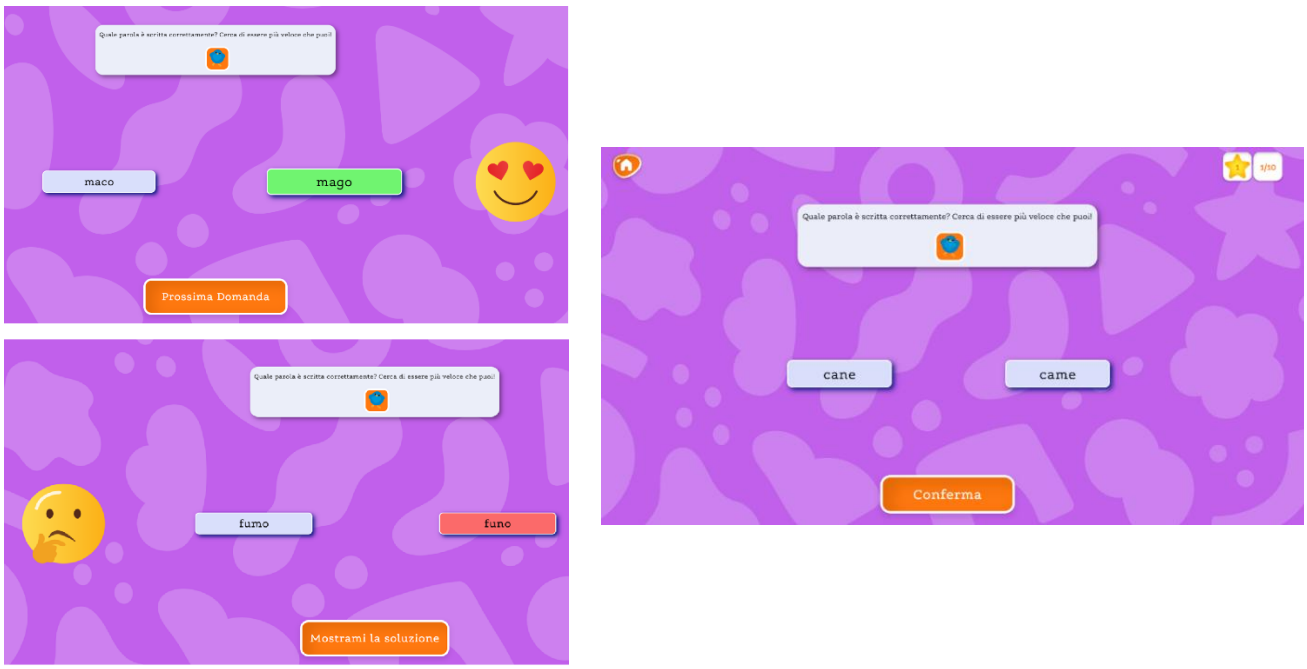


Figure 4: The right panel shows the display screen of a task. The task position and exercise difficulty level are shown at the top-right corner of the page. The help button (Blue Bird) is placed immediately below the instruction text. The left panel shows display screens of visual feedback. The top screen shows negative feedback, the bottom screen shows positive one.

5. CONCLUSION

We presented the design of an experimental study aiming to create an advanced technological tool for the linguistic training of children with DD. Our research has a very high potential social impact as it makes it possible to create technological tools for the improvement of reading difficulties through a web-designed application that is highly ecological for both the user and the therapist, with the scheduling of exercises tailored to the participant's linguistic and cognitive profile. The modular nature of the web application also paves the way for future implementations to other areas of intervention relevant to SLD learners (e.g., dyscalculia) or for adapting exercises to different school grade levels. We are currently collecting data, and the next step will involve analyzing the effectiveness of the training by comparing children's performance on the pre- and post-tests, as well as administering usability surveys of the app to both parents and children.

ACKNOWLEDGMENTS

This work was supported by the Italian National Recovery and Resilience Plan (PNRR), Mission 4, Component 2, Investment 1.1, funded by the European Union – NextGenerationEU, under the PRIN 2022 project "A Technology-driven linguistic intervention for children with language and literacy weaknesses", CUP: B53D23014380006.

This work would not have been possible without the collaboration and support of ULSS9 Scaligera Verona (Verona and Tregnago), the teachers of the *Istituto Comprensivo "Golosine"* in Verona, and the *Istituto San Girolamo Emiliani* in Corbetta (MI) as well as the participation of all the children and parents involved in the project.

REFERENCES

Arnbak, E., & Elbro, C. (2000). The Effects of Morphological Awareness Training on the Reading and Spelling Skills of Young Dyslexics. *Scandinavian Journal of Educational Research*, *44*(3), 229–251. https://doi.org/10.1080/00313830050154485

Bishop, D. V. (2009). TROG 2: test for reception of grammar-version 2. *Edizioni Giunti OS, Firenze*.

Bishop, D. V. M., & Snowling, M. J. (2004). Developmental Dyslexia and Specific Language Impairment: Same or Different? *Psychological Bulletin, 130*(6), 858–886. https://doi.org/10.1037/0033-2909.130.6.858

Blachman, B. A. (1991). Early intervention for children's reading problems: Clinical applications of the research in phonological awareness. *Topics in language disorders*, *12*(1), 51-65.

Bowers, P. N., Kirby, J. R., & Deacon, S. H. (2010). The Effects of Morphological Instruction on Literacy Skills: A Systematic Review of the Literature. *Review of Educational Research, 80*(2), 144–179. https://doi.org/10.3102/0034654309359353

Breznitz, Z. (1997). Enhancing the reading of dyslexic children by reading acceleration and auditory masking. *Journal of Educational Psychology, 89*(1), 103–113. https://doi.org/10.1037/0022-0663.89.1.103

Breznitz, Z. (2001). The determinants of reading fluency: A comparison of dyslexic and average. In M. Wolf (Ed.), *Time, fluency, and dyslexia* (pp. 245–277). York Press.

Breznitz, Z., & Share, D. L. (1992). Effects of accelerated reading rate on memory for text. *Journal of Educational Psychology, 84*(2), 193–199. https://doi.org/10.1037/0022-0663.84.2.193

Burani, C., Marcolini, S., De Luca, M., & Zoccolotti, P. (2008). Morpheme-based reading aloud: Evidence from dyslexic and skilled Italian readers. *Cognition, 108*(1), 243–262. https://doi.org/10.1016/j.Cognition.2007.12.010

Carlisle, J. F. (2000). Awareness of the structure and meaning of morphologically complex words: Impact on reading. *Reading and Writing*, *12*(3), 169–190. https://doi.org/10.1023/A:1008131926604

Casalis, S., Cole, P., & Sopo, D. (2004). Morphological awareness in developmental dyslexia. *Annals of Dyslexia, 54*(1), 114–138. https://doi.org/10.1007/s11881-004-0006-z

Catania, F., Spitale, M., Cosentino, G., & Garzotto, F. (2021). Conversational Agents to Promote Children's Verbal Communication Skills. In: *CONVERSATIONS 2020: Chatbot Research and Design. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE,* vol. 12604, p. 158-172, Amsterdam, https://doi:10.1007/978-3-030-68288-0_11

Cremonesi, P., Elahi, M., & Garzotto, F. (2017). User interface patterns in recommendation-empowered content intensive multimedia applications. *MULTIMEDIA TOOLS AND APPLICATIONS*, vol. 76, p. 5275-5309, ISSN: 1380-7501, https://doi:10.1007/s11042-016-3946-5

Crovari, P., Pidò, S., Pinoli, P., Bernasconi, A., Canakoglu, A., Garzotto, F., Ceri, S. (2022). GeCoAgent: A Conversational Agent for Empowering Genomic Data Extraction and Analysis. *ACM TRANSACTIONS ON COMPUTING FOR HEALTHCARE*, vol. 3, p. 1-29, ISSN: 2691-1957, https://doi:10.1145/3464383

Deacon, S. H., Tong, X., & Francis, K. (2017). The relationship of morphological analysis and morphological decoding to reading comprehension. *Journal of Research in Reading*, *40*(1), 1–16. https://doi.org/10.1111/1467-9817.12056

Dunn, L. M., & Dunn, L. M. (1965). *Peabody picture vocabulary test.*

First, M. B. (2014). *DSM-5 handbook of differential diagnosis*. American psychiatric publ.

Galuschka, K., Ise, E., Krick, K., & Schulte-Korne, G. (2014). Effectiveness of Treatment Approaches for Children and Adolescents with Reading Disabilities: A Meta-Analysis of Randomized Controlled Trials. *PLoS ONE, 9*(2), https://doi.org/10.1371/journal.pone.0105843

Georgiou, G., Vieira, A., Rothou, K., Kirby, J., Antoniuk, A., Martinez, D., & Guo, K. (2022). A Meta-analysis of Morphological Awareness Deficits in Developmental Dyslexia. *Scientific Studies of Reading*, *27*. https://doi.org/10.1080/10888438.2022.2155524

Kuo, L., & Anderson, R. C. (2006). Morphological Awareness and Learning to Read: A Cross-Language Perspective. *Educational Psychologist*, *41*(3), 161–180. https://doi.org/10.1207/s15326985ep4103_3

Levesque, K. C., & Deacon, S. H. (2022). Clarifying links to literacy: How does morphological awareness support children's word reading development? *Applied Psycholinguistics, 43*(4), 921–943. https://doi.org/10.1017/S0142716422000194

Marcolini, S., Traficante, D., Zoccolotti, P., & Burani, C. (2011). Word frequency modulates morpheme-based reading in poor and skilled Italian readers. *Applied Psycholinguistics*, *32*(3), 513–532. https://doi.org/10.1017/S0142716411000191

Pacheco, M. B., & Goodwin, A. P. (2013). Putting Two and Two Together: Middle School Students' Morphological Problem-Solving Strategies For Unknown Words. *Journal of Adolescent & Adult Literacy*, *56*(7).

Piccinin, S., & Dal Maso, S. (2023). Investigating factors affecting reading comprehension in second-generation pupils: The role of morphological skills. *Lingue e Linguaggio*, *22*(1), 117–142.

Pido' S., Crovari P., & Garzotto F. (2021). Modelling the bioinformatics tertiary analysis research process. *BMC BIOINFORMATICS*, vol. 22, p. 452-478, ISSN: 1471-2105, https://doi:10.1186/s12859-021-04310-5

Pra Baldi, A., Carretti, B., Cornoldi, C., & Giofrè, D. (2010). *Prove MT-3 Clinica: Nuove prove di lettura MT per la scuola primaria e secondaria di I grado*. Giunti OS.

Ramus, F., & Szenkovits, G. (2008). What Phonological Deficit? *Quarterly Journal of Experimental Psychology, 61*(1), 129–141. https://doi.org/10.1080/17470210701508822

Sartori, G., Job, R., & Tressoldi, P. (2007). *DDE-2,Batteria per la valutazione della dislessia e della disortografia evolutiva-2.* https://www.research.unipd.it/handle/11577/167982

Snowling, M. J., Hulme, C., & Nation, K. (2020). Defining and understanding dyslexia: Past, present and future. *Oxford Review of Education,* 46(4), 501–513. https://doi.org/10.1080/03054985.2020.1765756

Traficante, D., Marcolini, S., Luci, A., Zoccolotti, P., & Burani, C. (2011). How do roots and suffixes influence reading of pseudowords: A study of young Italian readers with and without dyslexia. *Language and Cognitive Processes*, *26*(4–6), 777–793. https://doi.org/10.1080/01690965.2010.496553

Vellutino, F. R., Fletcher, J. M., Snowling, M. J., & Scanlon, D. M. (2004). Specific reading disability (dyslexia): What have we learned in the past four decades? *Journal of Child Psychology and Psychiatry, and Allied Disciplines, 45*(1), 2–40. https://doi.org/10.1046/j.0021-9630.2003.00305.x

Vender, M. (2024). Potenziare la consapevolezza morfologica nella lettura e nella comprensione del testo in italiano in contesti di fragilità linguistica. *Italiano LinguaDue*, *16*(1), Article 1. https://doi.org/10.54103/2037-3597/23860

Vender, M., & Delfitto, D. (2024). Bridging the Gap in Adult Dyslexia Research: Assessing the Efficacy of a Linguistic Intervention on Literacy Skills. *Annals of Dyslexia*. https://doi.org/10.1007/s11881-024-00314-x

Vender, M., Melloni, C., & Delfitto, D. (2022). The effectiveness of reading intervention in adults with developmental dyslexia: A systematic review. *Italian Journal of Linguistics, 34*(1), 189–232

Vender, M., & Melloni, C. (2021). Phonological Awareness across Child Populations: How Bilingualism and Dyslexia Interact. *Languages, 6*(1), 1. https://doi.org/10.3390/languages6010039

Vender, M., Delfitto, D., & Melloni, C. (2020). How do bilingual dyslexic and typically developing children perform in nonword repetition? Evidence from a study on Italian L2 children. *Bilingualism: Language and Cognition*, *23*(4), 884–896. https://doi.org/10.1017/S1366728919000828

Il ruolo delle Infrastrutture nella costruzione di un ambiente di ricerca inclusivo. Un modello di buone pratiche

Marta Caradonna[1], Nicola Giampietro[1], Roberta Bianca Luzietti[2], Monica Monachini[2], Valeria Quochi[2], Emiliano Degl'Innocenti[3], Alessia Spadi[3], Alessandra Caravale[4], Antonio D'Eredità[4], Paola Moscati[4], Giacomo Mancuso[4]

[1] CNR – ILIESI, Italia, marta.caradonna@cnr.it, nicola.giampietro@cnr.it
[2] CNR – ILC, Italia, robertabianca.luzietti@cnr.it, monica.monachini@cnr.it, valeria.quochi@cnr.it
[3] CNR – OVI, Italia, emiliano.deglinnocenti@cnr.it, alessia.spadi@cnr.it
[4] CNR – ISPC, Italia, alessandra.caravale@cnr.it, antonio.deredita@cnr.it, giacomo.mancuso@cnr.it, paola.moscati@cnr.it

ABSTRACT (ITALIANO)

Il contributo presenta, attraverso la lente dell'inclusività, lo stato di avanzamento delle attività di mappatura e community-building del progetto infrastrutturale H2IOSC, che, federando i nodi italiani di quattro Infrastrutture di Ricerca ESFRI (CLARIN, DARIAH, E-RIHS, OPERAS), promuove un ecosistema digitale collaborativo e aperto. Fondato sui principi della Scienza Aperta e della Raccomandazione UNESCO del 2021, il progetto adotta metodologie partecipative e trasversali, combinando approcci quantitativi e qualitativi per analizzare risorse, servizi e comunità. Particolare rilievo è dato a concetti quali la qualità e l'integrità della ricerca, il beneficio collettivo, l'equità e la correttezza, nonché la diversità e l'inclusività. Il lavoro contribuisce così a superare le barriere disciplinari e propone un modello replicabile per ambienti di ricerca equi, accessibili e interoperabili.
Parole chiave: Infrastrutture di Ricerca; Scienza Aperta; Inclusione; Buone pratiche

ABSTRACT (ENGLISH)

The role of infrastructures in building an inclusive research environment. A model of good practices.
This contribution presents, through the lens of inclusiveness, the progress of the mapping and community-building activities within the H2IOSC project, which federates the Italian nodes of four ESFRI Research Infrastructures (CLARIN, DARIAH, E-RIHS, OPERAS) to promote a collaborative and open digital ecosystem. Grounded in the principles of Open Science and the 2021 UNESCO Recommendation, the project adopts participatory and cross-disciplinary methodologies, combining quantitative and qualitative approaches to analyse data resources, tools, software services, and research communities. Particular emphasis is placed on values such as research quality and integrity, collective benefit, equity and fairness, as well as diversity and inclusion. The work thus contributes to overcoming disciplinary boundaries and proposes a replicable model for building equitable, accessible, and interoperable research environments.
Keywords: Research Infrastructures; Open Science; Inclusion; Best Practices

1. INTRODUZIONE

Il progetto Humanities and Heritage Italian Open Science Cloud (H2IOSC) federa quattro nodi di infrastrutture di ricerca (IR) italiane: CLARIN, DARIAH, E-RIHS e OPERAS. Nell'ambito di questo progetto, il secondo Work Package WP2 si è concentrato sulla definizione, la mappatura e il monitoraggio delle strategie per valutare i contesti delle quattro IR, le caratteristiche delle loro comunità di utenti e le risorse, i servizi e gli strumenti specifici utilizzati, di nuova creazione o auspicati.
Il lavoro svolto porterà alla creazione di un **Osservatorio** nazionale che avrà l'obiettivo di fornire informazioni indispensabili per pianificare e implementare i sevizi del Marketplace di H2IOSC. H2IOSC è un'iniziativa di ampio respiro che si colloca all'intersezione tra ricerca scientifica e sviluppo tecnologico, con l'obiettivo di rispondere alle esigenze di una comunità scientifica sempre più orientata verso approcci basati sui dati collaborativi e interdisciplinari.
In questo contesto il WP2, dedicato alla mappatura delle risorse e delle esigenze delle comunità di riferimento, ha un ruolo centrale, poiché contribuisce a coordinare **realtà molto eterogenee** e ad individuare lacune e opportunità, tracciando una **"buona pratica"** per lo sviluppo di uno **spazio di ricerca inclusivo**. Le iniziative portate avanti contribuiscono al superamento di barriere disciplinari, identificando obiettivi comuni e creando un **ecosistema collaborativo**. Includendo nel workflow l'applicazione dei principi della **Scienza Aperta** il progetto intende favorire la ricerca, l'innovazione e la valorizzazione delle risorse scientifiche e tecnologiche in un ambiente interdisciplinare pensato dai ricercatori per i ricercatori. Il contributo che le quattro Infrastrutture federate stanno fornendo affinché si crei un ambiente di ricerca equo e democratico emerge dall'impianto stesso del progetto, il quale verte

sull'applicazione dei **principi FAIR** che prevedono, tra le altre cose, la standardizzazione della metadatazione dei dati della ricerca, proprio con l'obiettivo di garantire a tutte le ricercatrici e a tutti i ricercatori la comprensione e l'utilizzo dei dati e l'uso delle risorse ad essi correlati.
Vengono qui forniti i temi principali che guidano la progettazione e la gestione della federazione H2IOSC, ma la complessità dei risultati raggiunti e delle singole attività svolte rende impossibile la restituzione completa in una breve presentazione. In una successiva versione estesa ciascuno di questi aspetti potrà essere descritto in ogni suo dettaglio, quantitativo e qualitativo.

2. RICERCA INCLUSIVA: STRATEGIE CONDIVISE E METODOLOGIE TRASVERSALI

Per quanto concerne lo stato dell'arte dell'argomento qui trattato, il mondo della ricerca europea già da diversi anni integra i concetti di inclusione nei pilastri fondativi per la costruzione di un più equo spazio per la ricerca (Borrell-Damián, Braem and Famprikezi 2024; Horizon Europe Work Programme 2023-2025; EOSC Executive Board, 2020). La centralità di questi temi ha guidato fin dall'inizio la progettazione dell'Infrastruttura di Ricerca in oggetto, nella quale ogni Work Package lavora con il fine di contribuire non solo alla diffusione della conoscenza ma anche alla costruzione di una partecipazione ampia e diversificata. Questa impostazione inclusiva ha richiesto un ripensamento delle modalità di coinvolgimento delle comunità scientifiche, sviluppando strumenti capaci di catturare la complessità delle loro esperienze e dei loro bisogni. Discutere le sfide emerse è stato fondamentale per costruire un set di buone pratiche, replicabili in iniziative analoghe.
Le attività del WP2 *Landscaping and Building Communities* si sono basate su un mix di metodi quantitativi e qualitativi, utilizzando strumenti innovativi per raccogliere dati rappresentativi e significativi. La somma delle attività svolte all'interno del WP2 secondo le metodologie indicate dà vita al cosiddetto **Osservatorio**, che costituisce il punto di raccolta delle informazioni acquisite e processate dal gruppo di lavoro (cfr. Luzietti et al. 2024). Nel quadro di tali attività si conduce un lavoro sistematico di mappatura delle risorse digitali rilevanti per le infrastrutture di ricerca coinvolte nella Federazione.
Una matrice condivisa per la catalogazione consente di organizzare in modo strutturato dati provenienti da progetti, riviste specializzate, atti di convegni e altre pubblicazioni. La matrice raccoglie e descrive informazioni su dataset e strumenti digitali, includendo denominazione, acronimo, tipologia (es. corpus, archivio, risorsa lessicale), descrizione sintetica, strumenti di fruizione, link diretti, ambiti disciplinari, infrastrutture di riferimento, centri responsabili, formati e standard adottati (es. XML, EAD, EAC-CPF). Le comunità di ricerca partecipano attivamente all'identificazione delle risorse e alla definizione di modelli di rappresentazione aderenti ai principi FAIR. Particolare attenzione è rivolta alla rilevazione delle interrelazioni con altri progetti in corso presso le quattro infrastrutture del cluster e alla tracciabilità dei referenti scientifici e istituzionali, al fine di favorire interoperabilità, riuso e accesso ai dati. La mappatura si configura come passaggio strategico per la costruzione di un ecosistema informativo coerente e interoperabile, capace di superare la frammentazione disciplinare e promuovere la condivisione di pratiche. Su questa base si sviluppano i progetti pilota previsti dal WP2, che sperimentano workflow condivisi attraverso casi studio interdisciplinari, supportati dal feedback delle comunità. La definizione dell'Osservatorio tiene conto della natura collaborativa del progetto e delle interconnessioni multilivello tra gruppi di lavoro e ricercatori.
Le attività del gruppo di lavoro del WP2, *Landscaping and building communities*, contribuiscono anche allo sviluppo di progetti pilota, basati sui workflow elaborati da ciascuna infrastruttura, che dimostrino attraverso casi studio l'applicazione delle buone pratiche e l'interdisciplinarità delle soluzioni adottate.
Il WP2 supporta la costruzione dei casi studio fornendo non solo la mappatura multidisciplinare delle risorse del progetto ma anche il *feedback* raccolto presso le comunità di ricerca.
La progettazione di un **questionario** esplorativo ha rappresentato una delle sfide più complesse del progetto. Nonostante gli sforzi iniziali per sviluppare uno strumento completo e rappresentativo, il *field trial* ha evidenziato limiti significativi in termini di lunghezza e complessità delle domande. Il gruppo di lavoro ha quindi intrapreso un processo di semplificazione e riorganizzazione delle domande, suddividendo il questionario in due parti distinte per garantire un maggiore tasso di completamento. Questa strategia si è rivelata efficace, consentendo di raccogliere dati più rappresentativi e di migliorare l'esperienza degli utenti. Inoltre, i feedback raccolti hanno contribuito a raffinare ulteriormente gli strumenti d'indagine, rendendoli più accessibili e intuitivi.
Per approfondire alcune questioni che nel questionario rimangono solo parzialmente esplorate, sono stati introdotti i *focus group* come strumento qualitativo complementare. Progettati per raccogliere dati

dettagliati sulle esperienze e le percezioni dei ricercatori, essi offrono una prospettiva più approfondita sulle sfide e le opportunità legate all'uso delle tecnologie digitali.

2.1 FOCUS GROUP

La **multidisciplinarietà** e l'apertura a prospettive diverse rappresentano un punto di forza del progetto, garantendo una maggiore rappresentatività e rispondendo alle richieste della Commissione Europea di promuovere una ricerca scientifica equa e inclusiva2. Questo approccio è ulteriormente rafforzato dalla partecipazione attiva di ricercatori junior e senior provenienti da settori differenti, che contribuiscono a rendere il progetto un esempio di innovazione e collaborazione intersettoriale.
Tra luglio 2024 e marzo 2025 si sono svolti quattro incontri di gruppo, finalizzati a testare l'ipotesi secondo cui i ricercatori e i professori senior, afferenti a quattro diverse infrastrutture di ricerca, fossero più orientati verso il futuro rispetto agli studenti e ai giovani ricercatori.
Per analizzare tale orientamento, i gruppi sono stati costituiti bilanciando tre variabili: sesso biologico, fascia d'età e area disciplinare di appartenenza, in rappresentanza delle quattro infrastrutture che andranno a confluire nella federazione H2IOSC. Hanno partecipato complessivamente 24 persone, equamente suddivise tra 12 ricercatrici e 12 ricercatori, con un'età media complessiva di 38 anni. Le età medie per ciascun gruppo e la distribuzione dei partecipanti per disciplina sono riportate in Tabella 1.

Gruppo	Età media	Disciplina	Partecipanti
Senior 1	42	Archeologia	5
Senior 3	51	Filologia	4
Junior 2	30	Linguistica	8
Junior 4	31	Filosofia	7
		Totale	24

Tabella 1: Età media per ciascun gruppo e per disciplina

Ipotesi di ricerca

Sebbene tutte le comunità coinvolte nel progetto H2IOSC appartengano al dominio delle Scienze Umane e Sociali (SSH), esse si caratterizzano per differenze significative, che hanno reso necessario un lavoro di community building sin dall'inizio del progetto. Questo processo è facilitato dallo sviluppo congiunto di strumenti di ricerca e dalla conoscenza reciproca dei diversi contesti accademici. Di particolare utilità si è rivelata l'introduzione di metodologie qualitative a supporto delle linee di ricerca quantitative, nell'analisi delle esperienze e dei bisogni relativi alle risorse, agli strumenti e ai servizi da offrire nel Marketplace (MP).
L'ipotesi di partenza assume che le ricercatrici e i ricercatori con maggiore esperienza abbiano aspettative più articolate e richieste più elevate riguardo alle risorse e ai servizi digitali. Di conseguenza, si riteneva che potessero fornire indicazioni strategiche sulle esigenze infrastrutturali di breve e medio termine. Al contrario, si supponeva che le giovani ricercatrici e i giovani ricercatori, pur più vicini al mondo digitale, fossero meno esigenti a causa della loro minore esperienza di ricerca.
La Figura 1 rappresenta graficamente questa ipotesi, indicando con Δt il differenziale atteso di "proiezione verso il futuro", associato all'esperienza accademica e alla maturità scientifica.

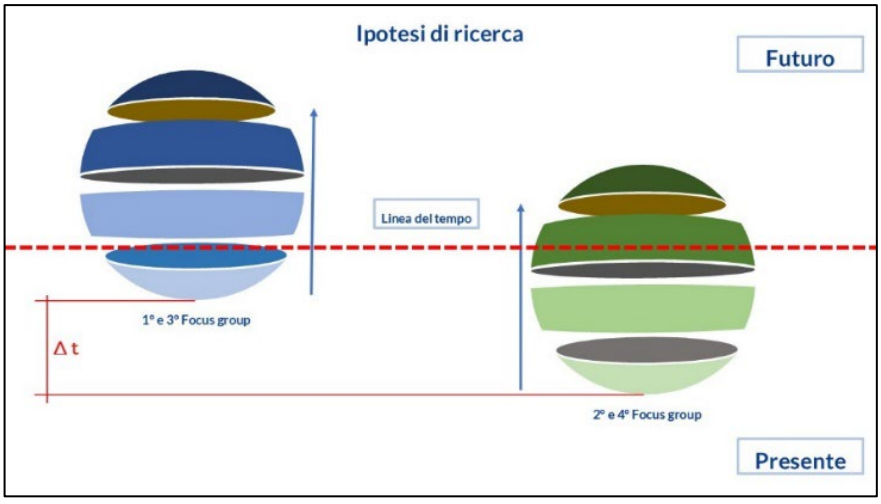


Figura 1: Ipotesi di ricerca

Primi risultati

La discussione nei gruppi si è articolata attorno a tre assi principali:

1. Esperienze di ricerca passate e attuali, con particolare riferimento all'uso di strumenti digitali;
2. Conoscenza e applicazione dei principi della Scienza Aperta e dei FAIR Principles;
3. Visione del futuro della propria area disciplinare in relazione a risorse, servizi e modalità di accesso a un'infrastruttura di ricerca.

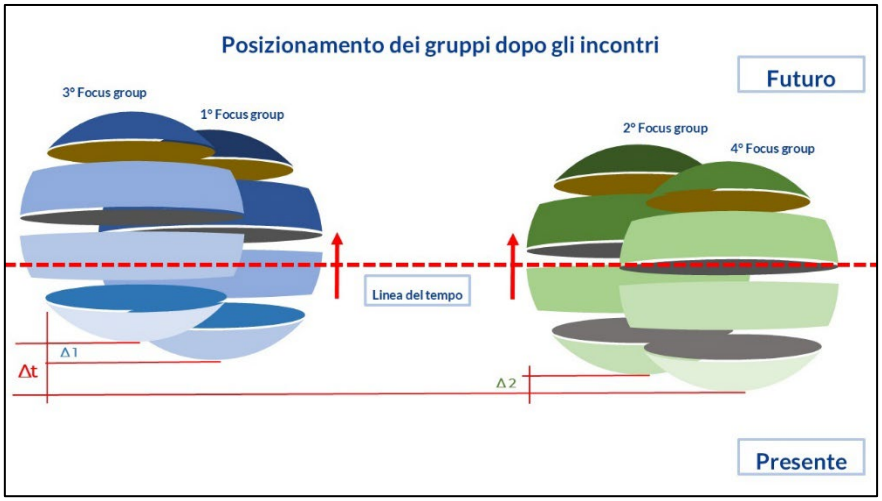


Figura 2: Posizionamento dei gruppi dopo gli incontri

Il Δt tra i gruppi junior (dottorandi e assegnisti) e quelli senior è stato sostanzialmente confermato, ma la differenza è risultata meno marcata del previsto. Interessanti sono emerse anche le differenze intragruppo. Il primo gruppo senior ha mostrato un buon orientamento al futuro, ma meno "visionario" rispetto al secondo gruppo senior (incontro 3), come evidenziato dalla differenza Δ1 in Figura 2. Analogamente, tra i due gruppi junior, il primo (secondo incontro) ha mostrato maggiore proiezione futura rispetto al secondo (quarto incontro), determinando una differenza Δ2.

Approccio ai temi della Scienza Aperta e dei FAIR Principles

Tutti i partecipanti hanno dichiarato di essere utenti di risorse digitali; solo alcuni, in particolare nei settori di archeologia e linguistica, si sono identificati anche come produttori di tali risorse. Per quanto riguarda la Scienza Aperta, i ricercatori senior hanno dimostrato una conoscenza più approfondita, mentre i giovani, pur manifestando entusiasmo per l'argomento, ne hanno offerto una visione più idealizzata e meno pratica. Simili dinamiche si sono osservate relativamente ai principi FAIR: i senior ne conoscono i limiti e le potenzialità, mentre i giovani, pur avendo scarsa esperienza diretta, si sono dichiarati aperti all'adozione futura di tali pratiche. Per i temi dell'inclusività, estremamente interessante sono state le dichiarazioni di uno dei giovani ricercatori.

Metafore infrastrutturali: la "casa digitale" della ricerca

Un momento particolarmente stimolante è emerso nel secondo incontro (gruppo junior), dove i giovani ricercatori hanno proposto una metafora dell'infrastruttura come "casa (digitale) della ricerca". Nei gruppi successivi, questa metafora è stata discussa criticamente. I ricercatori senior hanno suggerito alternative come "albergo" o "condominio", mentre alcuni giovani hanno espresso perplessità: "L'idea di casa mi incute timore (…) potrebbe diventare uno spazio escludente se i criteri di accesso non sono condivisi."

3. ASPETTATIVE PER IL FUTURO E CONCLUSIONI

Tutti i partecipanti, sia junior che senior, si aspettano un futuro della ricerca sempre più digitalizzato. Le opinioni sono risultate sostanzialmente allineate anche per quanto riguarda le modalità di accesso: l'accesso al MP deve essere gratuito per i singoli ricercatori, con eventuali costi sostenuti dagli enti di affiliazione. Nel terzo incontro (gruppo senior), inoltre, è stata posta attenzione alla necessità di un supporto umano nell'utilizzo del Marketplace.

Per affinare i dati emersi dalle risposte del questionario e per approfondire le tematiche relative al bisogno e all'utilizzo delle risorse digitali negli specifici settori del Cultural Heritage e dell'Heritage Science, sono state condotte una serie di interviste con ricercatori e docenti attivi in questo ambito di studi. In linea con i principi di inclusività del progetto, si è volutamente coinvolto un campione eterogeneo per sesso, età e posizione di carriera, in modo da dare voce ad interlocutori provenienti da background differenti, evitando parallelamente che le opinioni raccolte risultassero influenzate da pregiudizi legati a tali variabili.
Questi esperti hanno sottolineato l'importanza crescente di disporre di dati aperti, centralizzati e facilmente consultabili, evidenziando come tale disponibilità possa facilitare la ricerca e l'innovazione nel settore. In linea con queste esigenze e in accordo con le attività del WP7, è stata così sviluppata una piattaforma finalizzata all'aggregazione e alla condivisione di risorse e strumenti per l'archeologia digitale, inserita nel più ampio contesto delle scienze del patrimonio culturale, come parte integrante di una delle quattro infrastrutture. L'*Open Digital Archaeology Hub*[1] si configura come un punto di riferimento per la raccolta, l'aggregazione e la diffusione di risorse eterogenee attraverso l'esposizione dei loro metadati. Al suo interno, si trovano l'archivio *DHeLO* (*Digital Heritage Landscaping PlatfOrm*)[2], sviluppato nel WP2 come strumento di *landscaping* delle risorse digitali nel Cultural Heritage/Heritage Science , e i *repository* di tipo testuale, bibliografico (*BiDiar*) e visivo (*IADI*[3]) legati alla rivista *Diamond Open Access* Archeologia e Calcolatori[4]. Le risorse sono organizzate all'interno di un *framework* di *Linked Open Data*, che facilita la condivisione, l'accesso e l'interoperabilità dei dati, con un accesso privilegiato di tipo geografico, atto a valorizzare alcune delle aree e dei siti di maggiore interesse del nostro Paese. La piattaforma è progettata per essere scalabile, consentendo l'integrazione di dati provenienti da altri archivi, *repository* o progetti di ricerca, migliorandone la rintracciabilità e l'accessibilità e favorendone una diffusione inclusiva ed è a sua volta integrabile con infrastrutture di più ampio respiro, come ad esempio il Geoportale Nazionale per l'Archeologia[5] in una prospettiva tesa a valorizzare una sempre maggiore condivisione delle risorse scientifiche secondo i principi dell'Open Science.
H2IOSC si inserisce nel contesto delle discussioni emerse dal secondo vertice globale UNESCO sul Diamond Open Access, tenutosi a Città del Capo nel dicembre 2024[6]. L'adozione dei principi del *Diamond Open Access* è stata considerata cruciale per garantire l'accessibilità gratuita e la condivisione equa delle risorse scientifiche. Questa prospettiva ha rafforzato l'impegno del progetto verso la creazione di un'infrastruttura inclusiva, capace di rispondere alle esigenze di una comunità scientifica diversificata. La combinazione di strategie quantitative e qualitative, unita all'attenzione costante ai principi FAIR e agli obiettivi di inclusione, ha posto le basi per la creazione di un **Osservatorio permanente**. Questo strumento, progettato per fornire aggiornamenti periodici sullo stato delle Infrastrutture di ricerca, rappresenta un passo fondamentale verso una maggiore **trasparenza** e un dialogo continuo con le comunità scientifiche. La combinazione di analisi contestuale, coinvolgimento delle comunità e adozione di valori condivisi rappresenta un approccio strategico per rafforzare l'impatto delle Infrastrutture di ricerca nel panorama scientifico nazionale e internazionale.
In conclusione, il progetto descritto in estrema sintesi rappresenta un esempio significativo di come un approccio inclusivo e interdisciplinare possa contribuire alla realizzazione di un'Infrastruttura di Ricerca moderna e rappresentativa. L'attenzione ai temi dell'inclusione, combinata con un solido impianto metodologico, ha permesso di affrontare sfide complesse, fornendo al contempo un modello replicabile per iniziative future. Ulteriori sviluppi del progetto si concentreranno sull'implementazione di tecnologie avanzate per la gestione e l'analisi dei dati, nonché sull'estensione delle collaborazioni a livello internazionale, rafforzando così la posizione dell'Italia nel panorama delle Infrastrutture di Ricerca europee. Sarà essenziale continuare a promuovere il dialogo interdisciplinare e a favorire la partecipazione attiva delle comunità scientifiche, per garantire che i benefici di questo progetto possano essere condivisi su scala globale.

[1] https://open-archaeohub.cnr.it/
[2] https://dhelo.cnr.it/
[3] https://iadi.archcalc.cnr.it/
[4] https://www.archcalc.cnr.it/
[5] https://gna.cultura.gov.it/
[6] https://www.unesco.org/en/articles/diamond-open-access-advancing-knowledge-public-good, (cons. 20/01/2025).

RINGRAZIAMENTI

Questo lavoro è supportato dal Progetto H2IOSC - Humanities and Cultural Heritage Italian Open Science Cloud, finanziato dall'Unione Europea – NextGenerationEU – PNRR M4C2 - Codice progetto IR0000029 - https://www.h2iosc.cnr.it/

L'articolo riflette il lavoro di collaborazione di tutti gli autori. I nomi sono elencati in ordine alfabetico, raggruppati per infrastruttura di afferenza — OPERAS (che ha coordinato le attività legate ai focus group), CLARIN, DARIAH ed E-RIHS.

BIBLIOGRAFIA

Brunet, P., De Luca, L., Hyvönen, E., Joffres, A. et al., (2022) *European Commission: Directorate-General for Research and Innovation, Report on a European collaborative cloud for cultural heritage – Ex – ante impact assessment*, Publications Office of the European Union. https://data.europa.eu/doi/10.2777/64014

Borrell-Damián L., Braem, A. & Famprikezi, T., (2024), *Briefing on the ERA Policy Agenda 2022–2024*. Science Europe. 10.5281/zenodo.13862276, https://scienceeurope.org/media/wixlkytt/2024-briefing-era.pdf

EOSC Executive Board (2020) *Strategic Research and Innovation Agenda (SRIA) of the European Open Science Cloud (EOSC)*, Version 0.9 16 November 2020. European Science Cloud. https://www.eosc.eu/sites/default/files/EOSC-SRIA-V09.pdf

European Commission (2024) *Horizon Europe Work Programme 2023-2025, 11. Widening participation and strengthening the European Research Area.* European Commission Decision C 2371 of 17 April 2024. https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/wp-call/2023-2024/wp-11-widening-participation-and-strengthening-the-european-research-area_horizon-2023-2024_en.pdf

Märta Karlén (2023) *MINDtheGEPs Open Forum on Large scale international research infrastructures.* MINDthe GEPs gender equality in research, 21 marzo 2023. https://www.mindthegeps.eu/w/home/news/archive/2023-04-28-mindthegeps-open-forum-on-large-scale-international-research-infrastructures

Luzietti, R. B., Caradonna, M., Caravale, A., D'Eredita, A., Giampietro, N., Mancuso, G., Moscati, P., Quochi, V., Spadi, A., Monachini, M., & Degl'Innocenti, E. (2024). Digital Humanities and Heritage Science: Moving from landscaping to a dynamic research observatory in an Open Science Cloud. In: *Me.Te. Digitali. Mediterraneo in Rete Tra Testi e Contesti, Proceedings Del XIII Convegno Annuale AIUCD2024*. https://doi.org/10.6092/unibo/amsacta/7927

United Nations Educational Scientific and Cultural Organization (2021), *UNESCO Recommendation on Open Science*, UNESCO, Paris. https://doi.org/10.54677/MNMH8546

Predicting Grammatical Cases in Slovenian Varieties in Italy: A Use Case from the LORIS 1.1 Language Assistant

David Bordon[1]
[1] University of Ljubljana, Slovenia – david.bordon@ff.uni-lj.si

ABSTRACT (ENGLISH)

This work showcases the LORIS 1.1[1] language assistant – a web-based application developed to support the Slovenian speaking minority living in Italy, with a specific focus on the linguistic challenges arising from contact with Italian. Alongside integration of different language technology tools for Slovenian, LORIS includes a database of language-contact phenomena specifically created to target the language minority. Our main focus is on the technical challenges of operationalizing this dataset—namely, predicting all possible grammatical case forms in Slovenian and ensuring that the system avoids false positives. We discuss our approach to handling non-standard forms and highlight how this technology can promote linguistic inclusivity and preservation for minority languages.

Keywords: Slovenian Language; Language Technology; Minority Languages; Morphology Prediction; Digital Humanities

ABSTRACT (ITALIANO)

Previsione dei casi grammaticali nelle varietà di sloveno in Italia: Esempio pratico dell'assistente linguistico LORIS 1.1

Questo contributo presenta l'assistente linguistico LORIS 1.1, un'applicazione web-based sviluppata per supportare la minoranza di lingua slovena che risiede sul territorio italiano, con un interesse specifico per i fenomeni linguistici derivanti dal contatto con l'italiano. Oltre all'integrazione di diversi strumenti di tecnologia linguistica per lo sloveno, LORIS include un database di fenomeni di contatto linguistico creato appositamente per la minoranza slovena. Il presente lavoro si concentra principalmente sulle sfide tecniche legate all'operatività di questo set di dati, ovvero la previsione di tutte le possibili forme di casi grammaticali in sloveno, garantendo che il sistema eviti i falsi positivi. Discutiamo il nostro approccio alla gestione delle forme non standard e sottolineiamo come questa tecnologia possa promuovere l'inclusività linguistica e la conservazione delle lingue minoritarie.

Parole chiave: Lingua Slovena; Tecnologia Linguistica; Lingue Minoritarie; Previsione Morfologica; Digital Humanities

1. INTRODUCTION

Slovenia and Italy share a 232-kilometer border that stretches across much of Slovenia's western region. Because of a complex historical background and socio-economic factors, each country hosts a minority of the other's population – Slovenians in Italy and Italians in Slovenia – leading to mutual linguistic influence over time.

The Slovenian community in Italy has thus developed specific linguistic traits, which have been the subject of extensive research over the past two decades (Grgič, 2017; Jagodic et al., 2017). Data indicate that the language code in areas where the Slovenian minority in Italy is present is gradually diverging from the central standard of Slovenian to the point that "local usage sometimes significantly distances itself from the Slovenian linguistic continuum" (Grgič, 2017).

A study (Grgič & Popič, 2023) employing modern corpus linguistics methods analyzed approximately 20 million words from two annual volumes (2020–2021) of *Primorski dnevnik*, a Slovenian-language newspaper published in Italy focused on COVID-19-related terminology – an area where terminological consistency is to be expected due to its technical nature and international relevance. However, the analysis revealed a significant degree of lexical variation between the standard Slovenian used in Slovenia and the variety found in cross-border public mediatic discourse which demonstrates that the divergence is not limited to informal registers or locally contextualized reference but extends to structured, formal texts and standardized terminology.

[1] https://www.jeziknaklik.it/loris/

Minority, marginalized, or indigenous languages face the threat of decline or even (digital) extinction, while at the same time, language technologies offer them new opportunities (Grenoble & Whaley, 2006). The LORIS 1.1 language assistant, a browser-based application, is designed to provide linguistic advice to the Slovene minority living in Italy, navigating between two formal languages and systems.
While in recent times there have been many clones of the Grammarly app, and after the surge of generative AI tools there has been a great influx of AI grammar checkers, tools targeted at low resource languages, especially minority ones, are not as frequent. For example, initiatives such as GaelSpell[2] or Divvun[3], which are applications focused on minority languages, mostly act as spellcheckers. In this respect, Loris with its focus on language contact linguistic phenomena and by providing definitions and explanations of these phenomena, is among the more unique and specialized ones.

2. LORIS 1.1 Overview

LORIS 1.1 was designed to suggest standardized forms and explanations to users based on a database of common linguistic errors and interferences (e.g., false friends ("Avtist" (SI) *person with authism* vs. "Autista" (IT) *chauffeur*), incorrect case endings), as well as normative sources. The program is highly modular in its design, combining manually compiled and exported lexical lists of cases of language-contact, alongside more computationally heavy services integrated as "plugins".
The current version includes the following key resources:

- Data related to language contact between Slovenian and Italian (examples of paronyms, false friends, calques e.g. *ne vidim ure* (I cannot see the clock)/*non vedo l'ora* (I cannot wait))
- An internal database of toponyms in both, Slovenian and Italian (Slovene names of towns in Italy, e.g. *Devin/Duino* or *Nabrežina/Aurisina*)
- Standardized vocabulary from the *Slovenian Orthographic Dictionary 2001* (including words labeled as *incorrect*, *prohibited*, or *discouraged* – mostly tied to declination of proper names)
- Standardized vocabulary from the Slovenian morphological lexicon *Sloleks*[4]
- Synonyms for some of the most common lexemes, sourced from the *Synonyms 1.0* dictionary[5]

Except for internal SLORI lists, all resources were obtained from open-access and freely available databases. The custom lists were created by domain experts, aggregating materials from dictionaries, editorial and advisory services, and user suggestions to name a few.
In addition, LORIS incorporates various applications based on machine-learned models, all of which are based on the neural model BERT or sloBERTa. The tools in question are the automatic comma insertion tool *Vejice 1.0* (Božič et al., 2020) with an accuracy of 94 %[6], a neural spellchecker (Klemen et al., 2024) with $F_{0,5}$ scores on evaluation tasks exceeding 0.9, and the *EssayHelper* (Petrič, 2021) system for automatic correction of case forms and verb numbers (dual vs. plural).
The application works only on written text – there is no speech component due to the fact that open-source solutions for a low resource language such as Slovene are not readily available, while dialects and non-standardized vocabulary are not supported in current applications.
From launch in February 2023 until summer 2024 (last analysis), the tool was used 4970 times.
The full program code is available upon request on a GitHub repository curated by the authors[7]. Due to contractual obligations and third-party API integration, it is not possible to make it fully open source under a Creative Commons license.

3. Use Case – Predicting Grammatical Cases

An existing challenge during the development of LORIS was that none of the existing systems could predict non-standard cases. This issue is amplified by Slovenian's inherent morphological complexity – six cases, three grammatical genders and three grammatical numbers. To address these gaps, we developed a custom solution capable of predicting *all* possible variants for a given lemma, including unusual or non-standard forms.

[2] https://cadhan.com/gaelspell/index-en.html
[3] https://divvun.no/en/korrektur/speller-demo.html
[4] https://viri.cjvt.si/sloleks/slv/
[5] https://viri.cjvt.si/sopomenke/slv/
[6] https://orodja.cjvt.si/vejice/about
[7] To obtain access please send an e-mail specifying how you'd wish to use the data.

Unlike regex-based approaches, which struggle to differentiate between forms like *ravnatelj* (male) and *ravnateljica* (female), LORIS employs a manually curated list of lemma-specific cases. This approach ensures that every lemma and its potential non-standard cases are accounted for individually. The same principle applies to non-standard forms, where incorrect case endings must be manually listed to guarantee reliable detection.

The pipeline for implementing case detection and correction in LORIS follows a structured approach and is carried out manually using open-source tools for the Slovenian language.

First, all cases for a specific lemma are extracted from Sloleks. For lemmas not available in the database, cases are added following the declension rules for the Slovene language. For multi-word expression, every word in an expression is processed individually. The extracted cases are compiled into a structured dataset for manual cleaning. For multi-word expressions, every possible combination is tested and the ones not grammatically adequate are removed. Next, all duplicate entries are removed, and non-standard cases are manually added. The final step is to structure the data into strings and to add syntax coding for implementation in the build.

By curating every potential variant, from conventional to non-standard, this pipeline reduces false positives and enhances detection accuracy.

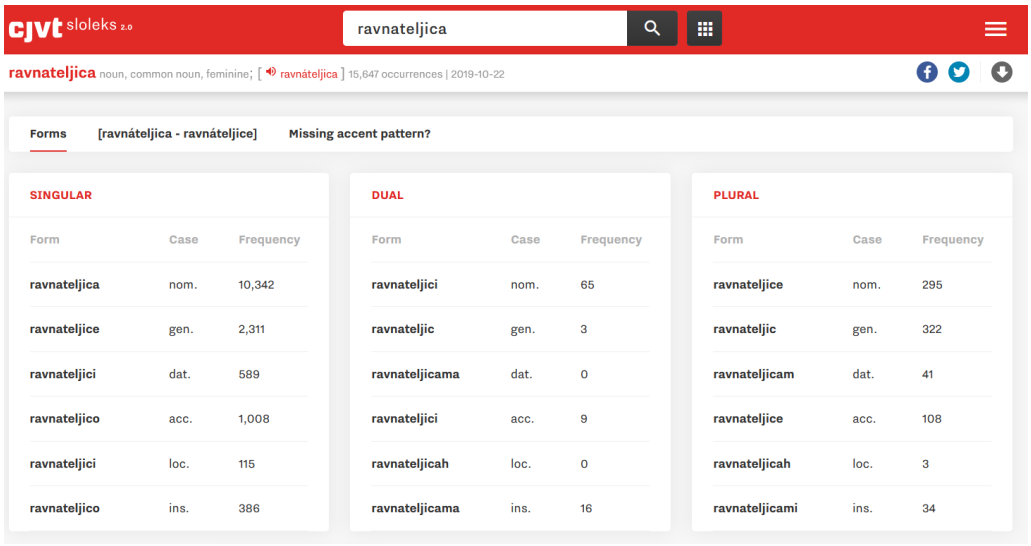


Figure 1. Lemmas for "ravnateljica" as viewed in Sloleks

ravnateljic
ravnateljica
ravnateljicah
ravnateljicam
ravnateljicama
ravnateljicami
ravnateljice
ravnateljici
ravnateljico

Figure 2. All cases for the lemma "ravnateljica"

```
/(?<!\S)ravnatelja(?![^\s.!?,:;])|(?<!\S)ravnatelju(?![^\s.!?,:;])|(?<!\S)
ravnateljem(?![^\s.!?,:;])|(?<!\S)ravnateljev(?![^\s.!?,:;])|(?<!\S)
ravnateljema(?![^\s.!?,:;])|(?<!\S)ravnateljih(?![^\s.!?,:;])|(?<!\S)
ravnatelji(?![^\s.!?,:;])|(?<!\S)ravnatelje(?![^\s.!?,:;])|(?<!\S)
ravnateljevega(?![^\s.!?,:;])|(?<!\S)ravnateljevemu(?![^\s.!?,:;])|(?<!\S)
ravnateljevem(?![^\s.!?,:;])|(?<!\S)ravnateljevim(?![^\s.!?,:;])|(?<!\S)
ravnateljeva(?![^\s.!?,:;])|(?<!\S)ravnateljevih(?![^\s.!?,:;])|(?<!\S)
ravnateljevima(?![^\s.!?,:;])|(?<!\S)ravnateljevi(?![^\s.!?,:;])|(?<!\S)
ravnateljeve(?![^\s.!?,:;])|(?<!\S)ravnateljevimi(?![^\s.!?,:;])|(?<!\S)
ravnateljevo(?![^\s.!?,:;])|(?<!\S)ravnatelj(?![^\s.!?,:;])/gmi,


/(?<!\S)ravnateljice(?![^\s.!?,:;])|(?<!\S)ravnateljici(?![^\s.!?,:;])|(?<!
\S)ravnateljico(?![^\s.!?,:;])|(?<!\S)ravnateljic(?![^\s.!?,:;])|(?<!\S)
ravnateljicama(?![^\s.!?,:;])|(?<!\S)ravnateljicah(?![^\s.!?,:;])|(?<!\S)
ravnateljicam(?![^\s.!?,:;])|(?<!\S)ravnateljicami(?![^\s.!?,:;])|(?<!\S)
ravnateljica(?![^\s.!?,:;])/gmi,
```

Figure 3. Final code output for "ravnatelj" and "ravnateljica"



Figure 4. The case detected in the program

4. Conclusion

LORIS 1.1 represents a useful tool for supporting Slovenian language use for the minority living in Italy and showcases how different open-source tools (if available) can be integrated into more focused applications to support a specific minority language community. Digital support for minority languages is crucial for their preservation and development – LORIS exemplifies how modern technologies can contribute to linguistic inclusion and awareness.

The authors hope this approach inspires comparable initiatives and strongly encourage the Italian speaking community in Slovenia to develop digital solutions to support their needs. We also envision cross-border collaboration that would allow other linguistic minorities to adopt similar strategies, fostering inclusivity and linguistic equity. In this way, LORIS 1.1 stands as both a practical resource for users and a model for developing digital tools that support minority language vitality and cultural identity.

ACKNOWLEDGEMENTS

The project was made possible with the support of the Slovene Research Institute SLORI and Central Office for the Slovenian Language of the Autonomous Region of Friuli–Venezia Giulia.

REFERENCES

Bird, S. (2020). *Decolonising speech and language technology*. In D. Scott, N. Bel, & C. Zong (Eds.), *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 3504–3519). International Committee on Computational Linguistics.

Fabjan Bajc, D. (1994). *Lažni prijatelji: Slovensko-italijanski slovar paronimov / I falsi amici: Vocabolario italiano-sloveni dei paronimi*. Mladika.

Fabjan Bajc, D. (1995). *Dve muhi na en mah: Slovensko-italijanski frazeološki slovar / Due piccioni con una fava: Vocabolario fraseologico sloveno-italiano*. Zadruga Goriška Mohorjeva.

Grenoble, L. A., & Whaley, L. J. (2006). *Saving languages: An introduction to language revitalization*. Cambridge University Press.

Grgič, M. (2017). *Italijansko-slovenski jezikovni stik med ideologijo in pragmatiko*. *Jezik in slovstvo, 62*(1), 89–98.

Grgič, M., & Popič, D. (2023). *Procesi jezikovnega separatizma pri čezmejnih jezikovnih manjšinah: Series Historia et Sociologia, 33*(1), 151–166. http://www.dlib.si/details/URN:NBN:SI:doc-QIYGXKVD

Jagodic, D., Kaučič Baša, M., & Dapit, R. (2017). *Jezikovni položaj Slovencev v Italiji*. In N. Bogatec & Z. Vidau (Eds.), *Skupnost v središču Evrope: Slovenci v Italiji od padca Berlinskega zidu do izzivov tretjega tisočletja* (pp. 66–88). ZTT-SLORI.

Klemen, M., Božič, M., Holdt, Š. A., & Robnik-Šikonja, M. (2024). *Neural spell-checker: Beyond words with synthetic data generation*. In E. Nöth, A. Horák, & P. Sojka (Eds.), *Text, speech, and dialogue. 27th International Conference, TSD 2024* (Lecture Notes in Computer Science, Vol. 15048). Springer.

Petrič, T. (2022). *Predlogi jezikovnih popravkov v slovenščini z modelom SloBERTa* (Bachelor's thesis, University of Ljubljana). University of Ljubljana Repository.

Verso un futuro senza barriere: l'accessibilità dei documenti elettronici nell'European Accessibility Act

Stefano Allegrezza[1]
[1] Università degli Studi di Macerata, Italy, stefano.allegrezza@unimc.it

ABSTRACT (ITALIANO)
L'accessibilità digitale, che ha l'obiettivo di garantire la piena inclusione delle persone con disabilità nell'ambiente digitale, rappresenta una delle sfide più importanti della trasformazione tecnologica. La Direttiva (UE) 2019/882, nota come European Accessibility Act (EAA), approvata nel 2019 e destinato a diventare pienamente operativo entro il 28 giugno 2025, rappresenta un passo avanti fondamentale in questa direzione. Tale direttiva non riguarda soltanto l'accessibilità dei siti web, dei prodotti o dei servizi, come si potrebbe concludere da una prima lettura, ma anche quella dei documenti elettronici, compresi quelli che vengono pubblicati sui siti web, e ciò ha delle conseguenze molto importanti sulle modalità di produzione dei documenti medesimi. Questo contributo analizza il contesto normativo e le novità introdotte dal regolamento, con particolare attenzione all'impatto sulla produzione di documenti informatici. Vengono esplorate le sfide e le opportunità per i soggetti produttori di contenuti, fornendo indicazioni sulle tecnologie e sulle buone pratiche per garantire la conformità, in particolare focalizzando l'attenzione sul formato PDF/UA.

Parole chiave: accessibilità; European Accessibility Act; documenti; archivi; inclusione.

ABSTRACT (ENGLISH)
Paper Title for AIUCD2025. Towards a barrier-free future: accessibility of electronic documents in the European Accessibility Act

Digital accessibility, which aims to ensure the full inclusion of people with disabilities in the digital environment, is one of the most important challenges of the technological transformation. Directive (EU) 2019/882, known as the European Accessibility Act (EAA), passed in 2019 and set to become fully operational by 28 June 2025, is a key step in this direction. This directive does not only concern the accessibility of websites, products or services, as one might conclude from a first reading, but also the accessibility of electronic documents, including those that are published on websites, and this has very important consequences on the way documents are produced. This contribution analyses the regulatory context and the novelties introduced by the regulation, with a particular focus on the impact on the production of electronic documents. Challenges and opportunities for content producers are explored, providing guidance on technologies and best practices for ensuring compliance, with a particular focus on the PDF/UA format.

Keywords: accessibility; European Accessibility Act; documents; archives; inclusion.

1. INTRODUZIONE
L'accessibilità digitale è un diritto fondamentale sancito dalla Convenzione delle Nazioni Unite sui diritti delle persone con disabilità, ratificata dall'Unione Europea nel 2010. In un contesto in cui sempre più servizi, documenti e interazioni avvengono in forma digitale, garantire l'accessibilità universale è cruciale per evitare nuove forme di esclusione sociale. Infatti, secondo l'Organizzazione mondiale della sanità (World Health Organization) oltre 1,3 miliardi di persone soffrono di disabilità significative. Ciò rappresenta il 16% della popolazione mondiale, ovvero 1 persona su 6. Questo numero è in crescita a causa dell'aumento delle malattie non trasmissibili e dell'allungamento della vita (World Health Organization, 2023).
Questo significa che l'inclusione della disabilità è un obiettivo che non può più essere trascurato ed è fondamentale per raggiungere gli Obiettivi di Sviluppo Sostenibile (Sustainable Development Goals).
Allo scopo di contribuire al soddisfacimento dei requisiti di accessibilità dei prodotti e dei servizi a favore delle persone con disabilità, il 17 aprile 2019 è stata emanata la "Direttiva (UE) 2019/882 del Parlamento europeo e del Consiglio del 17 aprile 2019 sui requisiti di accessibilità dei prodotti e dei servizi", meglio

nota come European Accessibility Act (EAA). La direttiva è stata recepita dall'Italia nel mese di maggio 2022 ed entrerà pienamente in vigore il prossimo 28 giugno 2025.
L'Accessibility Act mira a rimuovere le barriere digitali e a promuovere la creazione di un mercato unico in cui i requisiti di accessibilità siano uniformi. La sua applicazione è particolarmente rilevante per la produzione di documenti informatici, un ambito che richiede una combinazione di conoscenze tecniche, normative e gestionali per garantire conformità e inclusività.

2. IL CONTESTO NORMATIVO: DALLA CONVENZIONE DELLE NAZIONI UNITE SUI DIRITTI DELLE PERSONE CON DISABILITÀ ALL'EUROPEAN ACCESSIBILITY ACT

La Convenzione delle Nazioni Unite sui diritti delle persone con disabilità (Convention on the Rights of Persons with Disabilities, CRPD), approvata dall'Assemblea delle Nazioni Unite nel dicembre 2006 (United Nations, 2006), rappresenta il fondamento normativo internazionale per l'accessibilità. Il 24 febbraio 2009 il Parlamento della Repubblica Italiana ha ratificato la Convenzione, che è diventata legge dello Stato; il 23 dicembre 2010 la Convenzione è stata ratificata dall'Unione europea. L'articolo 9 di tale Convenzione stabilisce che gli Stati firmatari devono garantire che le persone con disabilità abbiano accesso, su base di uguaglianza, all'ambiente fisico, ai trasporti, alle tecnologie dell'informazione e della comunicazione (ICT) e ad altri servizi aperti o forniti al pubblico. La CRPD ha avuto un impatto significativo sulle politiche globali, fungendo da catalizzatore per l'adozione di legislazioni nazionali e regionali. L'Unione Europea, in quanto organizzazione sovranazionale parte della Convenzione, ha utilizzato questo quadro come punto di riferimento per sviluppare un corpus normativo incentrato sull'accessibilità digitale, adottando un approccio graduale. Un primo passaggio cruciale è stata l'emanazione della "Direttiva (UE) 2016/2102 del Parlamento europeo e del Consiglio del 26 ottobre 2016 relativa all'accessibilità dei siti web e delle applicazioni mobili degli enti pubblici", che ha imposto agli enti pubblici di rendere accessibili i propri contenuti digitali conformemente agli standard WCAG 2.1 di livello AA (Chisholm et al., 2021). Le disposizioni si applicano a siti web, documenti digitali scaricabili e applicazioni mobili, rendendo obbligatoria la pubblicazione di una dichiarazione di accessibilità per ciascun servizio digitale offerto (European Parliament and Council, 2016).
Questa direttiva è stata recepita in Italia con il Decreto Legislativo del 10 Agosto 2018, n.106 "Riforma dell'attuazione della direttiva (UE) 2016/2102 relativa all'accessibilità dei siti web e delle applicazioni mobili degli enti pubblici" che ha integrato la Legge n.4 del 9 Gennaio 2004 "Disposizioni per favorire e semplificare l'accesso degli utenti e, in particolare, delle persone con disabilità agli strumenti informatici" (Comunemente nota come "Legge Stanca", in quanto venne proposta da Lucio Stanca, Ministro per l'innovazione e le tecnologie). La "Legge Stanca" è stata anche il riferimento normativo da cui poi sono nate le Linee guida sull'accessibilità degli strumenti ICT dell'Agenzia per l'Italia Digitale (AgID), che mantiene una sezione del proprio sito dedicata a questo tema (AgiD, 2023).
Tuttavia, la Direttiva (UE) 2016/2102 si concentra esclusivamente sul settore pubblico, lasciando il settore privato al di fuori del suo campo di applicazione. Ciò ha portato a un panorama normativo frammentato, con alcune aziende private che hanno adottato l'accessibilità su base volontaria mentre altre ignoravano del tutto la questione. Anche a livello di pubbliche amministrazioni, ha regnato una grande confusione, e molte di esse (per non dire la maggior parte) ha continuato a produrre documenti elettronici in formati non accessibili, compresi quelli pubblicati all'albo pretorio on-line del proprio sito web. Si sono verificati, ad esempio, casi di pubbliche amministrazioni che pubblicavano regolarmente sul proprio sito documenti in formato PDF o addirittura anche PDF/A-1a ottenuti da scansioni di documenti nativi analogici (cartacei), non sottoposti ad OCR e quindi del tutto inaccessibili, ad esempio, agli screen reader utilizzati da soggetti con disabilità visive.
L'European Accessibility Act, emanato dal Parlamento europeo e dal Consiglio nel 2019 e recepito in Italia con il Decreto Legislativo 27 maggio 2022, n. 82 "Attuazione della direttiva (UE) 2019/882 del Parlamento europeo e del Consiglio, del 17 aprile 2019, sui requisiti di accessibilità dei prodotti e dei servizi", rappresenta una risposta a questa lacuna, ampliando il campo di applicazione della normativa europea sull'accessibilità. A differenza della Direttiva 2016/2102, che riguardava i siti web e le applicazioni *mobile*, l'European Accessibility Act si applica a una vasta gamma di prodotti e servizi, tra cui le comunicazioni elettroniche, i terminali self-service, i servizi bancari, e i documenti digitali elettronici associati a tali servizi.
L'European Accessibility Act si colloca nel contesto della *Strategia per i diritti delle persone con disabilità 2021-2030,* elaborata dalla Commissione europea per ampliare gli obiettivi definiti nella precedente *Strategia 2010-2020,* allo scopo di migliorare la qualità della vita delle persone con disabilità, costruendo

un ambiente inclusivo e libero da barriere o discriminazioni. L'obiettivo è duplice: da una parte, quello di ridurre i fattori che alimentano l'esclusione sociale e la povertà, garantendo parità di accesso, partecipazione e opportunità, in linea con i principi del Pilastro europeo dei diritti sociali e promuovendo l'inclusione sociale delle persone con disabilità; dall'altra, creare un mercato unico per prodotti e servizi accessibili, eliminando barriere normative tra gli Stati membri (European Commission and Council, 2019). La scadenza fissata per il giugno 2025 rappresenta un punto di svolta, dato che da quel momento in poi i prodotti e i servizi non conformi potrebbero essere esclusi dal mercato. Tale deadline richiede un'azione immediata da parte delle organizzazioni per aggiornare i propri processi di produzione documentale (European Commission and Council, 2019).

3. GLI STANDARD TECNICI SULL'ACCESSIBILITÀ

L'implementazione dell'European Accessibility Act è strettamente collegata a una serie di standard tecnici e normative complementari. Tra questi, i più rilevanti sono:
- le Web Content Accessibility Guidelines (WCAG 2.1), sviluppate dal World Wide Web Consortium (W3C), che forniscono le linee guida principali per garantire l'accessibilità di contenuti web e documenti digitali (World Wide Web Consortium 2018);
- lo standard "ISO 14289-1:2014 -- Document management applications — Electronic document file format enhancement for accessibility -- Part 1: Use of ISO 32000-1 (PDF/UA-1)", che definisce un profilo del formato PDF, denominato PDF/UA-1, specificatamente pensato per la produzione di documenti elettronici accessibili in PDF (ISO, 2014);
- lo standard "ISO 14289-2:2024 -- Document management applications — Electronic document file format enhancement for accessibility -- Part 2: Use of ISO 32000-2 (PDF/UA-2)" che definisce un'estensione del formato PDF/UA-1 denominata PDF/UA-2 (ISO, 2024);
- lo standard europeo EN 301549 per l'accessibilità delle ICT, che riguarda prodotti hardware, software e servizi digitali ed è stata recepita dall'UNI con la norma "UNI EN 301549 – Requisiti di accessibilità per prodotti e servizi ICT" il 17 dicembre 2020 (UNI, 2020).
Questi standard fungono da base tecnica per valutare la conformità dei prodotti e servizi ai requisiti di accessibilità previsti dalla normativa europea.
Si noti che una delle principali innovazioni introdotte dall'European Accessibility Act è la responsabilità estesa non solo ai produttori di tecnologie, ma anche ai fornitori di servizi, comprese le aziende private. Questo implica che qualsiasi documento informatico che accompagna un prodotto o servizio — manuali d'uso, contratti, comunicazioni ufficiali, comprese le norme e gli standard — deve essere reso accessibile a tutti, indipendentemente dalla disabilità dell'utente.

4. L'ACCESSIBILITÀ DEI DOCUMENTI ELETTRONICI

L'accessibilità dei documenti elettronici è un requisito essenziale per assicurare che i contenuti siano fruibili da chi utilizza tecnologie assistive e per coloro che richiedono configurazioni specifiche per poter interagire con il materiale digitale. Non solo: questo requisito è fondamentale anche per assicurare la creazione di archivi digitali che siano accessibili non solo oggi ma anche in futuro, da ogni categoria di persona.
A questo proposito è bene ricordare che quando si parla di accessibilità di documenti elettronici di solito si fa riferimento ad una ristretta categoria di utenti, rappresentata tipicamente da soggetti senza vista o ipovedenti, che utilizzano gli screen-reader per poter "leggere" quanto compare a schermo; tuttavia, è necessario cominciare ad allargare il campo d'azione ad altre categorie di utenti con disabilità, come le persone senza percezione del colore, senza udito o con udito limitato, con disabilità motoria, con neurodivergenza, con disabilità cognitiva, con disabilità psichica, con capacità di manipolazione o forza limitata, ecc.
A questo proposito, l'Allegato I dello European Accessibility Act ricorda che nella fornitura di informazioni occorre rispettare tutta una serie di requisiti affinché tali informazioni siano:

- rese disponibili attraverso più di un canale sensoriale;
- presentate in modo comprensibile;
- presentate agli utenti in modalità percepibili;
- presentate in caratteri di dimensioni e forme idonee, tenendo conto delle condizioni d'uso prevedibili e usando un contrasto sufficiente nonché una spaziatura regolabile tra lettere, righe e paragrafi;
- rese disponibili, con riferimento al contenuto, in formati testuali utilizzabili per la produzione di formati assistivi alternativi fruibili in modi diversi e attraverso più di un canale sensoriale;

- accompagnate da una presentazione alternativa di eventuale contenuto non testuale;
ecc.

In questi ultimi anni c'è stato uno sforzo da parte di alcune pubbliche amministrazioni ed aziende finalizzato alla redazione di raccomandazioni e linee guida che potessero fornire consigli, soprattutto di taglio operativo, sulla produzione di documenti elettronici accessibili, allo scopo di assicurare la conformità sia ai citati Regolamenti Europei che alla normativa nazionale. Ad esempio, l'Agenzia per l'Italia Digitale ha reso disponibile una "Guida pratica per la creazione di un documento accessibile" che contiene diversi suggerimenti interessanti (AgID 2017); la multinazionale Adobe ha reso disponibili delle pagine di help on line sulla creazione di documenti in formato PDF accessibile (Adobe 2023), ecc. Queste raccomandazioni si concentrano in genere sugli aspetti tecnici del documento – come la presenza di una struttura, l'utilizzo di testi alternativi per le immagini, l'utilizzo delle tabelle e di collegamenti ipertestuali, ecc. – e sulla modalità di produzione del documento, ad esempio a partire da un formato di office automation (tipicamente DOCX o ODT) per arrivare ad un formato di rappresentazione (come il PDF o il PDF/A). Tuttavia, né il PDF né il PDF/A sono formati specificamente pensati per assicurare l'accessibilità universale; come tali non assicurano la fruibilità da parte di un soggetto con disabilità.

5. IL FORMATO PDF/UA

Sul versante dei formati elettronici in grado di assicurare i requisiti di accessibilità, una delle proposte più interessanti è il PDF/UA (PDF/Universal Accessibility), un profilo del formato PDF che risponde alle raccomandazioni di accessibilità ai contenuti da parte delle persone con disabilità, in conformità con le normative attuali. Le specifiche del formato stabiliscono quali caratteristiche di un documento elettronico in formato PDF devono obbligatoriamente essere presenti e quali, invece, devono essere assenti affinché quel documento risulti accessibile alle persone con disabilità, garantendo il diritto costituzionale di fruizione dei servizi informatici e telematici della Pubblica Amministrazione, ai servizi di pubblica utilità, e alle informazioni diffuse in formati digitali.
Tale formato è stato definito per la prima volta nel 2012, con la pubblicazione della norma ISO 14289-1:2012 (ora ritirata) che ha introdotto il PDF/UA-1; successivamente nel 2014 è stata pubblicata la norma ISO 14289-1:2014, che conteneva una serie di correzioni rispetto alla prima versione. Tuttavia, l'interesse verso il formato è cresciuto notevolmente solo negli ultimi anni, da una parte per il fatto che recentemente è stata pubblicata la norma ISO 14289-2:2024 che ha definito la seconda versione del formato (PDF/UA-2); dall'altra per il fatto che l'imminente scadenza del 28 giugno 2025 ha spinto le organizzazioni alla ricerca di un formato elettronico in grado di soddisfare pienamente i requisiti previsti dall'European Accessibility Act.
Affinché un documento in formato PDF possa dirsi conforme alle specifiche del formato PDF/UA deve possedere tutta una serie di caratteristiche. Senza addentrarci in dettagli tecnici eccessivi, possiamo citare, a titolo di esempio, il fatto che i tag devono rappresentare correttamente le strutture semantiche del documento (titoli, elenchi, tabelle, ecc.); sono vietati i contenuti problematici, compresi i titoli illogici, l'uso di colori/contrasti per trasmettere informazioni, JavaScript inaccessibile e altro ancora; le immagini devono includere descrizioni di testo alternative; le impostazioni di sicurezza devono consentire alle tecnologie assistive di accedere ai contenuti; i font devono essere incorporati e il testo deve essere mappato in Unicode; ecc. (ISO, 2024).
Al convegno AIUCD 2025 verrà presentata un'analisi approfondita del formato PDF/UA al fine di valutare la sua corrispondenza ai requisiti dell'European Accessibility Act e la sua scelta come formato di elezione per la produzione di documenti elettronici accessibili.

6. CONCLUSIONI

L'imminente scadenza del 28 giugno 2025 per la piena entrata in vigore dell'European Accessibility Act costituisce un forte stimolo all'approfondimento dei temi legati all'accessibilità, allo scopo non soltanto di "mettersi in regola" con gli obblighi normativi ma anche di avviare concretamente delle politiche inclusive. L'evoluzione normativa dell'accessibilità digitale dimostra un impegno crescente a livello internazionale ed europeo per promuovere l'inclusione e la parità di accesso. Tuttavia, l'efficacia di queste norme dipenderà dalla capacità degli Stati membri di implementare misure adeguate e delle organizzazioni di adottare tecnologie e pratiche che garantiscano la conformità. Alcuni Paesi, come la Germania e la Svezia, hanno

già sviluppato percorsi nazionali avanzati sul tema dell'accessibilità digitale, che possono fungere da modello per altri Stati membri, ed è auspicabile che anche in Italia vengano avviati percorsi analoghi. Il prossimo grande banco di prova sarà rappresentato proprio dalla piena entrata in vigore dell'European Accessibility Act il 28 giugno 2025, che determinerà il livello di impegno dell'Unione Europea verso un ecosistema digitale realmente accessibile per tutti.

BIBLIOGRAFIA

Agenzia per l'Italia Digitale (AgID) (2023). Linee guida accessibilità – PA, https://www.agid.gov.it/it/design-servizi/accessibilita/linee-guida-accessibilita-pa

Agenzia per l'Italia Digitale (AgiD) (2017). Guida pratica per la creazione di un documento accessibile. Versione 27 marzo 2017 - Aggiornamento del documento del 18 luglio 2016. Disponibile sul sito dell'Università di Macerata, https://www.unimc.it/it/amministrazionedigitale/accessibilita/guida-pratica-per-la-creazione-di-un-documento-accessibile/guida-pratica-per-la-creazione-di-un-documento-accessibile-1.pdf.

Adobe (2023). Creazione di PDF accessibili. https://helpx.adobe.com/it/indesign/using/creating-accessible-pdfs.html.

Chisholm, W., Vanderheiden, G., & Jacobs, I. (2021). *Web Content Accessibility Guidelines (WCAG) 2.1*. World Wide Web Consortium. https://www.w3.org/TR/WCAG21.

European Parliament and Council (2016). Directive (EU) 2016/2102 on the accessibility of the websites and mobile applications of public sector bodies. Official Journal of the European Union. https://eur-lex.europa.eu/legal-content/IT/AUTO/?uri=celex:32016L2102.

European Parliament and Council (2019). Directive (EU) 2019/882 on the accessibility requirements for products and services (Accessibility Act). Official Journal of the European Union. https://eur-lex.europa.eu/eli/dir/2019/882/oj.

International Standardization Organization (ISO) (2014). ISO 14289-1:2014 -- Document management applications — Electronic document file format enhancement for accessibility -- Part 1: Use of ISO 32000-1 (PDF/UA-1). https://www.iso.org/standard/64599.html.

International Standardization Organization (ISO) (2024). ISO 14289-2:2024 -- Document management applications — Electronic document file format enhancement for accessibility -- Part 2: Use of ISO 32000-2 (PDF/UA-2). https://www.iso.org/standard/82278.html

UNI (Ente Nazionale di Unificazione) (2020). EN 301549 – Requisiti di accessibilità per prodotti e servizi ICT. https://www.uni.com/images/stories/uni/allegati_norme/UNIEN301549/UNIEN301549_accessibile.pdf.

United Nations (2006). Convention on the Rights of Persons with Disabilities (CRPD). https://www.ohchr.org/en/instruments-mechanisms/instruments/convention-rights-persons-disabilities.

World Health Organization (2023). Disability. https://www.who.int/en/news-room/fact-sheets/detail/disability-and-health

World Wide Web Consortium (W3C) (2018). Web Content Accessibility Guidelines (WCAG) 2.1. https://www.w3.org/TR/WCAG21/

La gestione del nuovo sapere digitale contemporaneo. Scenari, criticità, sfide, prospettive

Nicola Barbuti[1]
[1] Dipartimento di Ricerca e Innovazione Umanistica DIRIUM – Università degli Studi di Bari Aldo Moro, Italia – nicola.barbuti@uniba.it

ABSTRACT (ITALIANO)
Il contributo presenta una riflessione sugli attuali scenari della digitalizzazione e dei processi di digitizzazione, focalizzando l'attenzione sulle criticità e sulle urgenze determinate dalla crescita esponenziale di materia digitale. Evidenziando la necessità di provvedere all'elaborazione di un impianto teoretico che sia di riferimento per attivare la sistematizzazione delle manifestazioni ed espressioni del nuovo sapere digitale che ci definisce e identifica, si discute l'urgenza di creare una cultura digitale consapevole sia negli operatori da impegnare nelle sfide con la dimensione digitale, sia delle comunità di utenti digitali. A riguardo, si ripropongono alcuni estratti di un documento prodotto nel 2019, la cosiddetta Carta di Pietrelcina, che presentava riflessioni e proposte in linea con il tema trattato e secondo noi oggi tornate attuali, nell'auspicio che possano fornire spunti di interesse da cui muovere per ulteriori, indispensabili iniziative di elaborazione teoretica di una nuova digiteconomia.
Parole chiave: Digiteconomia, Digitizzazione, Digitalizzazione Cultura digitale consapevole, Digital Heritage

ABSTRACT (ENGLISH)
Managing the New Contemporary Digital knowledge. Scenarios, Issues, Challenges, Perspectives
The paper presents a reflection on the current scenarios of digitalization and digitization processes, focusing on the critical issues and emergencies arising from the exponential growth of digitized resources. Highlighting the need to develop a theoretical framework that serves as a reference to systematize the manifestations and expressions of the new digital knowledge that defines and identifies us, we discuss the urgency of creating an aware digital culture both among operators engaged in challenges with the digital dimension, and within digital user communities. In this regard, some excerpts from a document produced in 2019, the so-called "Carta di Pietrelcina," are revisited, which presented reflections and proposals in line with the topic addressed, and that we think could be relevant even today, hoping that they can provide points of interest from which to move forward with further, indispensable theoretical elaborations of a new "digiteconomia".
Keywords: Digiteconomia, Digitization, Digitalization, Aware Digital Culture, Digital Heritage

1. INTRODUZIONE

Nell'ipercomplessità della trasformazione digitale (DT) (Dominici 2019; Dominici 2023), amplificata dalla recente deflagrazione degli LLMs "a portata di tutti" con ChatGPT (Ciotti 2023), la digitalizzazione e i processi di digitizzazione sono diventate una delle sfide più significative e potenzialmente foriere di prospettive e ricadute promettenti, se reinterpretate alla luce della loro recente evoluzione. Infatti, soprattutto i processi di digitizzazione sono diventati da tempo la manifestazione della trasformazione profonda nei metodi e nelle tecniche con cui stiamo inconsapevolmente generando un potenziale nuovo patrimonio culturale, composto esclusivamente di risorse digitali.
Sappiamo che le manifestazioni ed espressioni dell'agire e del sapere umano generano quasi esclusivamente da necessità correnti e circostanziali, che ne motivano e giustificano l'esistenza. Quasi nulla di quanto le comunità antropiche producono nei loro cicli di vita è generato in base ad ambizioni o presupposti necessariamente culturali. Nella millenaria evoluzione delle società umane, l'istanza di "culturale" ha sempre caratterizzato entità tangibili o intangibili, materiali e immateriali, di cui riconosciamo, leggiamo, classifichiamo e condividiamo le narrazioni mediate dalla comunicazione strutturata, che ne testimoniano il ciclo di vita e l'uso e riuso che le comunità antropiche ne hanno fatto nello spazio e nel tempo (Barbuti, 2022).

Va da sé che questi presupposti sono applicabili anche al sapere digitale che già da tempo ci definisce e identifica in quanto comunità sempre più immerse nella compulsione computazionale[1]. Anche i dati sono sempre generati da processi di digitizzazione in riscontro a bisogni e per fini concreti. Se registrassero diacronicamente anche le interazioni delle comunità di utenti che ne fruiscono, acquisirebbero nel tempo l'istanza culturale sopra definita, restituendo gradualmente la narrazione della memoria collettiva contemporanea. Le informazioni andrebbero registrate non secondo approcci casuali o circostanziali, ma in base a linguaggi e regole che ne garantiscano l'intellegibilità e consentano di veicolarne la conoscenza e l'esistenza oltre l'angusto spazio/tempo dell'utilizzo quotidiano (Tomasi, 2017; Barbuti e De Bari, 2024). Una parte della riflessione scientifica (Macrì e Cristofaro, 2021) ritiene che i processi di digitizzazione così sostanziati possano rappresentare un motore fondamentale per lo sviluppo socioculturale sostenibile. Sostanziare i dati registrando le informazioni sulle comunità di inter-attori digitali amplierebbe le connessioni e le relazioni potenziali tra i saperi dell'ecosistema digitale contemporaneo, e le comunità di utenti che in futuro ne vorranno esplorarne e comprenderne i processi accedendo alle narrazioni che lo raccontano, generando una graduale simbiosi tra i concetti di "utente" e di "istanza culturale". Invece, ancora oggi nei processi di digitizzazione la registrazione di queste informazioni è materia oscura e sottovalutata nella sua importanza. Il che rappresenta una vera e propria emergenza, in quanto, di fatto, l'intero ecosistema socioculturale contemporaneo che ci definisce e ci identifica da oltre mezzo secolo potrà affidare alla storia tracce di sé poco o nulla significative, e per lo più inintelligibili ai non addetti ai lavori (Bailey 2015; Ghosh 2016; Cosimi 2022).

2. STATO DELL'ARTE

Delineare metodi efficaci di riconoscimento e trasmissione delle istanze culturali potenzialmente insite nelle risorse generate dalla digitizzazione presuppone un impianto teoretico che supporti l'organizzazione, gestione, archiviazione, trasmissione e valorizzazione del nuovo sapere sotteso alla produzione quotidiana di miriadi di dati digitali granulari. Il confronto su questa urgenza versa da tempo in una fase embrionale dalla quale fatica a evolvere. La riflessione dovrebbe muovere da una posizione di osservazione prospettica, il cui punto di fuga sia una graduale disseminazione di una cultura digitale consapevole in tutte le diverse comunità di persone. È innegabile, infatti, che la trasformazione digitale abbia di fatto modificato già da tempo il concetto stesso di "retaggio culturale". Oggi, esso include pratiche sociali e relazioni non-tangibili peculiari alla dimensione digitale, generate senza intermediazioni istituzionali e disseminate in esperienze e condivisioni soggette a letture simultanee e multiple. Tra le manifestazioni e relative espressioni classificate come culturali, gli oggetti digitali possiedono un'esistenza propria, pur rimanendo concettualmente e materialmente connessi alla dimensione fisica «quali nodi di interscambi semantici» (Digital Library Italia, 2022). È perciò indispensabile emancipare i processi di digitizzazione dagli schemi rigidi delle dottrine con cui ci relazioniamo al patrimonio analogico, e focalizzare i progetti culturali digitali su prospettive user-oriented, interattive, dinamiche, diacroniche, sostenibili e inclusive. Nel nostro Paese, nonostante il Piano Nazionale di Digitalizzazione (PND) in corso[2] miri ad allineare istituzioni e cittadini nella titolarità responsabile della digitizzazione del patrimonio culturale, permangono due criticità, di fatto emerse in tutta la loro urgenza solo quando sono partiti i cantieri operativi: la mancanza di professionisti adeguatamente formati all'esecuzione e gestione dei processi, che sta sollevando criticità all'esecuzione delle scansioni; la metadatazione, ancora una volta focalizzata sugli artefatti originali digitizzati e non sulle risorse e collezioni digitali.
In un'intervista recente per il blog dell'AIUCD (Bolioli, 2022), Klaus Kempf ha riaffermato la necessità di considerare la digitizzazione come «un processo completo di trasformazione di contenuti analogici in contenuto digitale». Kempf ritiene che solo attraverso la «contestualizzazione delle grandi masse di dati si possano ottenere scoperte e conoscenze inaccessibili ai ricercatori», evidenziando l'importanza di costruire modelli concettuali di metadati basati su relazioni, provenienza e ciclo di vita delle risorse. Le sue asserzioni confermano l'urgenza di definire i requisiti culturali che caratterizzino e identifichino l'imponente materia digitale in generazione dal PND – circa 76 milioni di artefatti da riversare in rete entro il 2026 – in quanto manifestazione ed espressione di un nuovo sapere che, un domani, possa essere identificabile quale testimone e memoria della rivoluzionaria fase storica che stiamo vivendo. Questo imponente magma

[1] Singolare notare quanto i due lemmi siano assonanti e simili nella composizione strutturale, ma riferiti all'interazione con il digitale esprimano appieno il paradosso tra il rigoroso presupposto concettuale che presiede alla sua generazione e il modo con cui ancora oggi la maggioranza assoluta dei fruitori interagisce.
[2] https://digitallibrary.cultura.gov.it/il-piano/; https://digitallibrary.cultura.gov.it/notizie/come-cambia-la-visione-del-patrimonio-culturale-nellambiente-digitale

di materia digitale non-tangibile necessita di essere allocato in uno spazio virtuale – ma sempre fisico – del quale vanno definite con accuratezza consistenza, funzioni e accessibilità. La Commissione Europea ha stanziato 110 milioni di euro per sviluppare entro il 2025 un'infrastruttura cloud collaborativa dedicata digitizzazione di artefatti culturali, ricerca scientifica e documentazione sul patrimonio culturale digitale[3]. L'obiettivo è agevolarne la classificazione, organizzazione, archiviazione, gestione e valorizzazione nella sua dimensione di appartenenza, svincolandola definitivamente dagli approcci dannosamente conservatori (Barbuti e De Bari, 2024).
L'impellenza non più procrastinabile di un approccio sistemico alla digitizzazione impone di risalire a monte dell'impianto concettuale che vi presiede, per focalizzare la riflessione sulla necessità di intraprendere l'elaborazione di un impianto teoretico di riferimento, che raccolga, organizzi e renda disponibili i saperi e gli strumenti con cui affrontare e sostenere le difficili sfide ancora irrisolte a essa connesse. In primis, dare sostanza al concetto di digital heritage sancito dall'UE nel 2014, che dovrebbe includere tutte le nuove forme di manifestazioni ed espressioni culturali da essa generate, riconoscendone il valore intrinseco di entità culturali autonome. Un'istanza ancora avvolta in una spessa nebulosità concettuale, proprio perché non esiste ancora alcuna sistematizzazione che ne definisca manifestazioni ed espressioni, né un insieme di regole necessarie a riconoscerlo e organizzarlo in un nuovo sapere organico e coerente (Duranti e Shaffer, 2012; Chrysakis et al., 2018).

3. DISCUSSIONE

Quanto sopra delineato porta inevitabilmente a riaffermare ancora una volta (Barbuti 2022) l'urgenza di formare nel brevissimo tempo una prima generazione di professionisti, in grado di confrontarsi con il nuovo sapere digitale avendo consapevolezza sia dei metodi, che delle tecniche necessarie a governarlo e trasmetterlo. Il tema è argomento di confronto quanto mai pressante e improcrastinabile, sebbene sia già da tempo oggetto di riflessione scientifica soprattutto in ambito biblioteconomico (Myburg & Tammaro, 2013; Del Rosso & Lampert, 2013; Tammaro, Casarosa & Madrid, 2013; Colombati & Giusti, 2016; Tammaro, 2016; Walek, 2018; Barbuti, 2022; Barbuti et alii, 2022). Secondo l'opinione maggiormente condivisa, il processo di apprendimento dovrebbe puntare a innescare nei curatori del nuovo sapere digitale un graduale cambiamento di mentalità, istanziata dalla consapevolezza che digitalizzazione e digitizzazione sono esse stesse manifestazioni ed espressioni generative di nuova cultura. A tal fine, è indispensabile porsi da un punto di osservazione che si focalizzi sulla ricerca di un equilibrio quanto più stabile possibile tra tradizione e cambiamento, basato su un approccio all'innovazione aperto e flessibile che permetta ai curatori del nuovo sapere digitale di affrontare consapevolmente un cammino ancora non ben delineato.
Partendo da questa prospettiva, riteniamo che non si possa più rimandare oltre l'elaborazione di un impianto teoretico che sia di riferimento e supporto ai futuri curatori professionisti che saranno impegnati a censire, selezionare, organizzare, archiviare, gestire e rendere accessibile quanto saranno abilitati a riconoscere e validare come digital heritage nel magma di dati che quotidianamente produciamo. Pensiamo a una digiteconomia[4] che, muovendo dai presupposti teoretici e applicativi della sua ascendenza biblioteconomica, sistematizzi regole e procedure (nomos) di riferimento per i futuri curatori culturali chiamati a relazionarsi al nuovo sapere digitale (digit), classificandolo e rendendolo fruibile agli utenti in modo organizzato, coerente, intellegibile e interattivo (teche).
Alcune prime riflessioni su direzioni e contenuti formativi in tema di digitalizzazione e digitizzazione sono state articolate già alcuni anni fa, nell'ambito confronto scientifico sulla cultura digitale affrontato nell'ambito della "Scuola a Rete DiCultHer", e raccolte in un manifesto *intitolato Carta di Pietrelcina sull'Educazione all'Eredità Culturale Digitale*[5]. Il documento delinea un quadro di riferimento per intraprendere il necessario rinnovamento dei processi educativi e formativi sia dei futuri professionisti del digital heritage, sia delle comunità di utenti potenziali e reali soprattutto giovani, nell'ottica di generare di una cultura digitale consapevole e inclusiva in tutti gli attori coinvolti nella trasformazione in corso.

[3] https://www.dataspace-culturalheritage.eu/en
[4] Il neologismo è nostro. Il lemma recepisce le istanze concettuali espresse dalla Biblioteconomia, in quanto riteniamo che questa rappresenti il principale impianto teoretico di riferimento – non unico, si intende – da cui trarre gran parte delle premesse e della sostanza necessarie ad articolare un valido sistema di organizzazione, archiviazione, gestione e valorizzazione del sapere digitale.
[5] https://www.diculther.it/. Il documento, redatto congiuntamente da Carmine Marinucci e Nicola Barbut prende il nome dalla sede nella quale il documento è stato presentato ufficialmente e sottoscritto tra la Scuola a Rete DiCultHer e l'allora MiBACT nel 2019, immediatamente prima della pandemia, nell'ambito della manifestazione nazionale "JazzIn" realizzata dalla Fondazione Ampioraggio.

Ne riproponiamo di seguito alcuni estratti che, a nostro parere, risultano oggi essere attuali quanto mai prima, con l'auspicio che possano essere di stimolo per riprendere con la dovuta consapevolezza la riflessione sulla necessità di attivare un'elaborazione teoretica, che sia di riferimento per quanti inevitabilmente saranno impegnati nella sfida di definire l'autentica sostanza culturale del nuovo sapere digitale che ci definisce e ci identifica.

4. LA "CARTA DI PIETRELCINA SULL'EDUCAZIONE ALL'EREDITÀ CULTURALE DIGITALE"

La Carta prende le mosse dall'urgenza di «restituire ai giovani la consapevolezza di quanto sia importante riappropriarsi della titolarità partecipata dell'eredità culturale, ripartendo proprio dal riconoscimento del valore della cultura digitale». Allo scopo, si richiama un precedente documento, il **"**Manifesto Ventotene Digitale" redatto in occasione dell'Anno Europeo della Cultura (2018), nel quale erano già evidenziate le principali sfide da affrontare per interagire consapevolmente con il nuovo sapere digitale:

- co-creare un sistema di conoscenze e competenze digitali consapevoli, abilitate ad assicurare conservazione, fruizione ampia, interattiva, partecipata e consapevole, sostenibilità, valorizzazione, promozione e presentazione del nuovo Digital Cultural Heritage;
- sviluppare la Cultura Digitale quale espressione dell'Eredità Culturale nella quale dobbiamo riconoscere e identificare l'ecosistema dei metodi, processi, fenomeni e risorse singole o complesse che identificano il nuovo Digital Cultural Heritage, la cui essenza, manifestazione ed espressione risiede nella trasferibilità e replicabilità nello spazio e nel tempo delle entità digitali che identificano, categorizzano e qualificano la storia e l'esistenza delle comunità contemporanee nei loro contesti sociali, culturali, economici.

Nei due punti si ritrovano articolate alcune delle istanze oggi divenute urgenti, portate alla riflessione collettiva in un periodo in cui, come detto, digitalizzazione e digitizzazione del patrimonio culturale erano del tutto estranee a qualsiasi dibattito sul loro impatto sociale e sull'eventuale valore culturale. Nel seguito, la Carta prova a sintetizzare in un insieme organico di riferimenti tematici le tante indicazioni emerse da consultazioni pubbliche e sperimentazioni svolte in quegli anni a livello nazionale. Ne sintetizziamo di seguito alcune riflessioni a nostro parere oggi quanto mai attuali.

«Eredità Culturale ed inclusione sociale»

In ambito di valorizzazione e fruizione dell'Eredità Culturale, si sottolinea come l'inclusione sociale si sviluppi effettivamente quando il luogo culturale diventa luogo educativo, nel quale tutte le categorie sociali riescono ad interagire proattivamente e a sviluppare conoscenze e competenze. L'eredità culturale, infatti, è sempre espressione e rappresentazione delle creazioni tangibili, intangibili e, oggi, digitali delle comunità che vivono i territori. Tuttavia, permangono molti contesti in cui gruppi sociali stabili non autoctoni non sono ancora del tutto integrati nella storia culturale e artistica e, conseguentemente, non riescono a contribuire attivamente alla creazione e all'evoluzione di una rinnovata storia sociale collettiva. Risulta, perciò, necessaria la riflessione sulla didattica inclusiva anche per l'apprendimento di metodi e tecniche di valorizzazione e fruizione dell'Eredità Culturale.

«Titolarità culturale e disponibilità e accessibilità pubbliche dei dati»

Il diritto all'istruzione, l'accesso alla cultura, la sovranità culturale-epistemologica sono identificate quali istanze da cui genera il principio fondamentale della Titolarità Culturale dei dati, con riferimento in particolare alla gestione pubblica e sostenibile dei luoghi dove si preserva ed è fruibile pubblicamente e interattivamente il Digital Heritage. La "presa in carico" di una responsabilità comune e condivisa rispetto a un bene comune, corrispondente ad un processo di acquisizione di "titolarità culturale" esercitata con diritto, si applica sia a ciò che ereditiamo dal passato, sia a ciò che abbiamo la possibilità di progettare e co-creare nell'ambito degli ecosistemi digitali in cui viviamo, sperimentiamo ed esercitiamo, con la prospettiva di lasciare a nostra volta questa eredità a chi verrà dopo di noi.

«Educazione all'Eredità Culturale Digitale»

L'Educazione all'Eredità Culturale Digitale è fondamentale per generare conoscenze e competenze di cittadinanza globale. Avendo la sua valorizzazione nell'eredità materiale, immateriale e digitale, è per sua natura multi-, trans- e interdisciplinare, e si fonda su metodologie condivise attive e partecipative che richiedono forti sinergie tra i territori, gli attori del sistema formativo istituzionale (scuola, università) e quanti operano nella valorizzazione e conservazione del patrimonio culturale.

«La nuova Cultura Digitale»

Provvedere all'integrazione fra saperi umanistici tradizionali e conoscenze di metodi e tecniche computazionali è fondamentale per generare cultura digitale consapevole, elaborando modelli formativi che

puntino a creare conoscenze e competenze trasversali, abilitate a generare quel digital knowledge design system necessario a un sistema educativo sostenibile. Il processo pone al centro la creatività dei giovani per affrontare, mediante l'uso consapevole del digitale e con approcci innovativi, la conoscenza, l'accesso partecipato, la gestione, la tutela la presentazione, la fruizione, la conservazione e valorizzazione del Digital Cultural Heritage.

«Convenzione di Faro»

La Convenzione di Faro[6] ha innescato una profonda rivisitazione del concetto di Eredità Culturale legandola indissolubilmente alle comunità, assumendo un ruolo cruciale nel favorire un uso critico e consapevole di metodi e tecnologie digitali per favorire l'inclusione sociale di componenti culturali diverse. Si evidenzia come le metodologie e tecnologie digitali offrano la possibilità di raccogliere, condividere e archiviare/conservare qualsiasi forma espressiva, realizzando l'obiettivo di selezionare e valorizzare ciò che ci rappresenta. Il sapere digitale offre occasioni di riconfigurazione complessiva delle entità e dei luoghi culturali come 'eredità comuni', assumendo sostanza metodologica ed epistemologica, strutturale e di contesto, da cui generare una nuova ermeneutica della cultura per riorganizzare i saperi.

«La Carta dell'Educazione all'Eredità Culturale Digitale»

Si delineano i requisiti per un uso responsabile del digitale, che non può assolutamente prescindere dal coinvolgimento dei giovani e di quanti provvedono a vario titolo alla loro istruzione e formazione, in quanto potenziali protagonisti nella realizzazione di quel digital knowledge design system applicato all'educazione al patrimonio culturale. Un processo indispensabile per stimolare nei giovani il bisogno di conoscenza e di rigenerazione della memoria storica, sviluppando in loro una coscienza critica che li induca a oltrepassare le semplici erudizioni per riappropriarsi dell'importanza di saper leggere, interpretare e gestire le nuove fonti di conoscenza con autonomia intellettuale.

«Digital STHEAM»

Per la prima volta si afferma la necessità di evolvere le STEM in STHEAM, nell'ottica di affrontare consapevolmente l'ipercomplessità evolutiva che sta caratterizzando questo inizio di Ventunesimo secolo, definita da nuove alfabetizzazioni e, soprattutto, da conoscenze e competenze inter-, trans- e multidisciplinari. La transizione "from STEM to STEAM", dove la lettera "A" di "Art" identifica la creatività artistica e progettuale, e da STEAM a STHEAM, con l'inserimento delle Humanities, rimuoverebbe le barriere disciplinari e guiderebbe il cambiamento verso la consapevolezza che il digitale, dopo esserne stato una formidabile leva, può diventarne il motore alimentato da un'energia realmente sostenibile: la conoscenza. Una transizione che genera da un'idea rinnovata di "spazi di apprendimento", intesi non come luoghi fisici, ma come agorà virtuali dove interagire con l'innovazione, piattaforme in cui sperimentare la riappropriazione di conoscenze e competenze digitali non più solo astratte, ma anche e soprattutto officinali, dove studenti e docenti possano sviluppare collaborativamente percorsi cognitivi che favoriscano la reciproca condivisione di nuovi saperi. In questo paradigma, le metodologie e tecnologie digitali diventerebbero strutturali per la formazione e l'apprendimento condiviso, e scuola e università rappresenterebbero l'ecosistema formativo in grado di fornire ai giovani le chiavi di lettura del presente, con le quali aprire le porte di accesso alla transizione verso il futuro.

«Comunicazione eticamente efficace

La Carta dell'Educazione all'Eredità Culturale Digitale si propone quale premessa utile a favorire la conoscenza approfondita dell'uso corretto e consapevole della rete e degli strumenti e tecniche di comunicazione rese disponibili dalle nuove tecnologie digitali. Per conoscere l'Eredità Culturale Digitale, è fondamentale che ciascun cittadino la sappia trasmettere, comunicare, presentare e promuovere. Occorre, perciò, garantire a ciascuno la *media education* necessaria a conoscere e a interagire consapevolmente con metodi, strumenti, tecniche e piattaforme digitali, restituendo ai giovani il significato autentico di una "comunicazione eticamente efficace".

La Carta si chiude con 13 proposte di attività da avviare per sostanziare la riflessione teorica elaborata negli anni. Se ne riportano alcune, che a nostro rappresentano un utile riferimento anche oggi.

«Le prime Proposte [estratto]:

1. Considerare la presente Carta sull'educazione all'eredità culturale quadro di riferimento per un'idea rinnovata di "spazi di apprendimento", intesi come agorà virtuali dell'innovazione pedagogica.

2. Promuovere una alfabetizzazione critica digitale nelle scuole come materia propedeutica e obbligatoria a ogni attività. Una sorta di corso 0, com'è la grammatica o l'aritmetica. Il nucleo di un nuovo Trivium. Se non sai come funziona un motore di ricerca o non sai che cosa è una codifica digitale è come se non sapessi leggere.

[6] https://rm.coe.int/1680083746

6. Intensificare le occasioni di collaborare tra le Istituzioni scolastiche e le Istituzioni culturali afferenti al MiBAC, sia a partire dal piano dell'offerta formativa triennale che ogni scuola deve predisporre, sia per la realizzazione di programmi specifici, quali l'alternanza scuola lavoro.
7. Superare il riduttivismo culturale del passaggio da STEM a STEAM verso le Digital STHEAM.
8. Sostenere "meta incubatori officinali" per la formazione del Corpo docente e degli Studenti delle Scuole italiane sull'innovazione digitale per il Cultural Heritage con riferimento al modello di meta incubatore adottato a Matera (OFFICINE DI CULTURA DIGITALE).
9. Incentivare corsi di aggiornamenti professionale per il Corpo Docente, per gli operatori e professionisti delle Istituzioni culturali pubbliche e/o private […].
11. Sostenere la formazione professionale superiore, istituendo nuove tipologie di Istituti Tecnici Superiori (ITS) nel settore del Digital Cultural Heritage, declinati nei diversi ambiti operativi e di settore al sostegno delle industrie creative.
12. Promuovere e sostenere l'educazione e la formazione al Digital Cultural Heritage quale vantaggio competitivo per la creazione e il sostegno alle industrie culturali e creative.
13. Promuovere la media Education nelle scuole e, in maniera specifica, strumenti, tecniche e linguaggi che possano valorizzare il nostro patrimonio culturale digitale».

5. CONCLUSIONI

Un contest didattico nazionale, promosso dalla Scuola a Rete DiCultHer in collaborazione con Europeana, INDIRE e l'Istituto Centrale per il Catalogo Unico del MiC, ha rappresentato nel 2019 un'importante occasione per sperimentare e valutare l'applicazione "sul campo" delle istanze della Carta di Pietrelcina. All'interno di tale iniziativa, l'Istituto Comprensivo "Umberto I - San Nicola" di Bari, comprendente la Scuola Secondaria di I grado San Nicola della Città Vecchia e le Scuole Primarie "F. Corridoni" e "N. Piccinni", è stato selezionato per la realizzazione di un progetto di co-creazione digitale. L'intervento è stato svolto in collaborazione con il Dipartimento di Studi Umanistici DISUM (oggi DIRIUM) dell'Università degli Studi di Bari "Aldo Moro", gli studenti del corso di ITS per "Tecnico Superiore per la promozione e valorizzazione del turismo culturale digitale ed esperienziale" attivato dalla Fondazione ITS IOTA di Lecce, Vostok 100k del videomaker Lorenzo Scaraggi e Quorum Italia del visual designer Paolo Azzella.
Nel progetto, denominato "San Nicola salva la scuola!"[7], si è sperimentato il metodo living lab Crowddreaming quale modello di inclusione sociale fondato su metodologie partecipative e pratiche di creatività digitale. L'intervento si è focalizzato principalmente sulla realtà scolastica della Scuola secondaria inferiore San Nicola, la più antica scuola media della città di Bari, un tempo riferimento educativo d'eccellenza a livello cittadino, dotata di laboratori e strutture didattiche all'avanguardia, nel 2019 a rischio chiusura per mancanza di iscrizioni. Nondimeno, l'istituto conservava e ancora conserva molti degli antichi strumenti didattici e una biblioteca storica, preziose testimonianze materiali e immateriali dell'eredità culturale dell'istituto. Nell'ottica di muovere un primo passo verso il recupero e la valorizzazione di tale patrimonio in chiave digitale e partecipativa e, nel contempo, di riattivare l'interesse a iscriversi rilanciando il ruolo educativo dell'Istituto con una spinta verso una didattica innovativa, il progetto ha previsto la realizzazione di diversi contenuti digitali originali, tra cui:
- un'applicazione in AR interattiva, sviluppata dagli alunni della Scuola Secondaria San Nicola in collaborazione con i corsisti dell'ITS e con il DISUM di UNIBA utilizzando la piattaforma Metaverse, finalizzata a presentare la storia dell'Istituto attraverso dinamiche di gamification e modalità immersive di visita multimediale;
- un video talk documentaristico, a cura del videomaker Lorenzo Scaraggi;
- una collezione fotografica digitale corredata da metadati rappresentativa della storia e dell'identità dell'Istituto, realizzata con il supporto del graphic-maker Paolo Azzella e con la partecipazione attiva degli studenti della Scuola media San Nicola e dei bambini delle Scuole Primarie Corridoni e Piccinni.
Il progetto è stato presentato ufficialmente durante il Contest nazionale #HackCultura 2019, tenutosi nei giorni 15 e 16 aprile 2019 presso l'I.I.S. "G.B. Pentasuglia" di Matera. All'evento hanno partecipato figure istituzionali e accademiche di rilievo, tra cui la Dott.ssa Isabel Crespo, allora Responsabile Scientifica di Europeana Educational, il Dott. Carmine Marinucci per la Scuola a Rete DiCultHer, la Dott.ssa Elisa Sciotti del'ICCU, la Dott.ssa Irene Zoppi del'INDIRE e il Prof. Nicola Barbuti in rappresentanza del DISUM di UNIBA e del Polo Apulian DiCultHer.
Si era nel 2019. Da allora, a nostro parere, non molto è cambiato. Forse, sarebbe il caso di ammettere responsabilmente il ritardo e di attivarsi con la dovuta consapevolezza.

[7] https://bari.repubblica.it/cronaca/2019/05/31/news/bari_app_per_salvare_scuola_san_nicola-227644443/

BIBLIOGRAFIA

Barbuti, N. (2022). La digitalizzazione documentale. Metodi, tecniche, buone prassi. Milano: Editrice Bibliografica.

Barbuti, N., & De Bari, M. (2024). Libri e biblioteche tra museabilità e musealizzazione digitale: sogno o realtà? In Atti del convegno AIUCD 2024: Il futuro del patrimonio digitale. Università di Catania. https://aiucd2024.unict.it/atti-del-convegno/.

Barbuti, N., De Bari, M., Kameas, A., Chiotis, Th. (2022). New job role profiles to bridge the digital skills gap in the cultural heritage sectors: The BIBLIO project, «Umanistica Digitale», 6(13), pp. 97–115

Bailey, L. (2015) Digital Orphans: The Massive Cultural Black Hole on Our Horizon. Techdirt. <https://www.techdirt.com/ arti-cles/20151009/17031332490/digitalorphans-massive-cultural-blackhole-our-horizon.shtml>.

Bolioli, A. (2022). La trasformazione digitale delle biblioteche: intervista a Klaus Kempf sulle Digital Libraries, «Hypotheses. Il blog di Digitl Humanities di AIUCD – "Leggere, scrivere e far di conto"», marzo 2022 <https://infouma.hypotheses.org/999>.

Ciotti, F. (2023). Minerva e il pappagallo. IA generativa e modelli linguistici nel laboratorio dell'umanista digitale. Testo e Senso, 26. https://doi.org/10.58015/2036-2293/671; https://testoesenso.it/index.php/testoesenso/article/view/671.

Cosimi, S. (2015). Vint Cerf: ci aspetta un deserto digitale, in «Wired.it», <http://www.wired.it/attualita/2015/02/16/vint-cerf-futuro-medievale-bit-putrefatti/>

Chrysakis, I., Harami, L., Angelakis, D., and Bruseker, G. (ed. by, 2018). Generating and tracing the provenance of knowledge. Book of Abstract of CIDOC Conference. <http://www.cidoc2018.com/sites/default/files/CIDOC2018-BookOfAbstracts-Final-v-1-2.pdf>.

Dominici, P. (2019). Dentro la società interconnessa. La cultura della complessità per abitare i confini e le tensioni della civiltà ipertecnologica. Milano: Franco Angeli.

Dominici, P. (2023). From emergency to emergence: Learning to inhabit complexity and to expect the unexpected. Salute e Società, 22(1), 135–151. https://doi.org/10.3280/SES2023-001010.

Duranti, L. & Shaffer, E. (ed. by, 2012), The Memory of the World in the Digital Age: Digitization and Preservation. An international conference on permanent access to digital documentary heritage, in "UNESCO Confer-ence Proceedings". <http://ciscra.org/docs/UNESCO_ MOW2012_Proceedings_FINAL_ENG_Compressed.pdf>.

Ghosh, P. (2016), Google's Vint Cerf warns of 'digital Dark Age'. BBC News, Science & Environment <http://www.bbc.com/news/science-environment-31450389>.

Istituto Centrale per la Digitalizzazione del Patrimonio Culturale – Digital Library. (n.d.). Il Piano nazionale di digitalizzazione del patrimonio culturale. Ministero della Cultura. https://digitallibrary.cultura.gov.it/il-piano/

Macrì, E., & Cristofaro, C. L. (2021). The Digitalisation of Cultural Heritage for Sustainable Development: The Impact of Europeana. In Cultural Initiatives for Sustainable Development (pp. 373–400). Springer.

Myburgh, S., Tammaro, A.M. (2012). Education for Digital Librarians: Some European Observations, in: Amanda SPINK, Jannika HEINSTRÖM (ed. By), Library and Information Science Trends and Research: Europe, Vol. 6, Emerald Group Publishing Limited, 2012, p. 217-245.

Tammaro, A. M., Madrid, M., Casarosa, V., Digital curators' education: Professional identity vs. convergence of LAM (libraries, archives, museums). In: Digital Libraries and Archives: 8th Italian Research Conference, IRCDL 2012, Bari, Italy, February 9-10, 2012, Revised Selected Papers 8. Springer Berlin Heidelberg, 2013. pp. 184-194.

Tomasi Francesca, La preservazione del contenuto degli oggetti digitali: formalizzare la provenance, «Bibliothecae.it», 2017, 6, p. 17-40 https://cris.unibo.it/retrieve/handle/11585/611249/303579/paper-2017.pdf.

Wagner, A., & de Clippele, M.-S. (2023). Safeguarding Cultural Heritage in the Digital Era – A Critical Challenge. International Journal for the Semiotics of Law Revue internationale de Sémiotique juridique, 36, 1915–1923.

Archives and Digital Editions

Learner Corpus of Creative Writing: An interdisciplinary challenge

Ioanna Tyrou1, Katerina Florou2,
1 National and Kapodistrian University of Athens, Greece, iotyrou@ill.uoa.gr
2 National and Kapodistrian University of Athens, Greece, kathyflorou@ill.uoa.gr

ABSTRACT (ENGLISH)

Over the past two decades, corpora have become fundamental tools in computational linguistics and foreign language teaching, giving rise to the field of Corpus Linguistics. Learner corpora, collections of texts produced by foreign language learners, have been extensively studied to enhance language teaching methodologies. At the University of Athens, five years ago, an innovative project was initiated to integrate creative writing into learner corpora. This resulted in the creation of a corpus designed to address research questions related to teaching Italian as a foreign language, creative writing structures, and cross-linguistic analysis. This Learner Corpus consists of written texts produced by adult students of Italian, primarily native Greek speakers, in the context of creative writing activities. The corpus is systematically enriched through assignments aligned with specific themes, genres, and techniques, such as point-of-view shifts, material symbolism, and the "What if..." scenario exploration. It currently includes eight sub-corpora based on thematic units developed over six academic years, with a total size that reflects variations in student participation and task nature. The pedagogical applications of SCRICREA Corpus (derived from *"Scrittura Creativa"*) are extensive, encompassing language proficiency enhancement, creative expression development, and interlanguage analysis. Through activities such as narrative continuations, genre transformations, and music-inspired writing, learners develop linguistic accuracy, creativity, and critical thinking. Additionally, SCRICREA serves as a repository for research on linguistic phenomena, such as n-gram frequency analysis, further contributing to the field of corpus linguistics and foreign language pedagogy.

Keywords: Creative writing, Corpus Linguistics, Italian as a Foreign Language

ABSTRACT (ITALIANO)

Corpus degli Apprendenti di Scrittura Creativa: Una Sfida Interdisciplinare. Negli ultimi due decenni, i corpora sono diventati strumenti fondamentali nella linguistica computazionale e nell'insegnamento delle lingue straniere, dando origine al campo della Linguistica dei Corpora. I corpora degli apprendenti, collezioni di testi prodotti da studenti di lingue straniere, sono stati ampiamente studiati per migliorare le metodologie didattiche. Cinque anni fa, all'Università di Atene, è stato avviato un progetto innovativo per integrare la scrittura creativa nei corpora degli apprendenti. Questo ha portato alla creazione di un corpus progettato per affrontare domande di ricerca relative all'insegnamento dell'italiano come lingua straniera, alle strutture della scrittura creativa e all'analisi cross-linguistica.Questo Corpus degli Apprendenti consiste in testi scritti prodotti da adulti che apprendono l'italiano, principalmente madrelingua greci, nel contesto di attività di scrittura creativa. Il corpus viene arricchito sistematicamente attraverso compiti allineati a temi specifici, generi e tecniche, come i cambiamenti di punto di vista, il simbolismo materiale e l'esplorazione dello scenario "E se...". Attualmente include otto sub-corpora basati su unità tematiche sviluppate nel corso di sei anni accademici, con una dimensione totale che riflette le variazioni nella partecipazione degli studenti e nella natura dei compiti.Le applicazioni pedagogiche di SCRICREA (derivato da "Scrittura Creativa") sono estensive e comprendono il miglioramento della competenza linguistica, lo sviluppo dell'espressione creativa e l'analisi interlinguistica. Attraverso attività come continuazioni narrative, trasformazioni di genere e scrittura ispirata alla musica, gli studenti sviluppano accuratezza linguistica, creatività e pensiero critico. Inoltre, SCRICREA funge da repository per la ricerca sui fenomeni linguistici, come l'analisi della frequenza degli n-grammi, contribuendo ulteriormente al campo della linguistica dei corpora e della pedagogia delle lingue straniere.

Parole chiave: Scrittura creativa, Linguistica dei Corpora, Italiano come Lingua Straniera

1. INTRODUCTION

Since the new millennium, corpora have established themselves as a cornerstone in the fields of computational linguistics and foreign language teaching, forming an independent discipline known as Corpus Linguistics (McEnery &Hardie, 2012). Different types of corpora are employed depending on the specific needs of each researcher. Learner corpora, for instance, are extensive collections of oral or written texts produced by students of a foreign language (Granger, 2004). As a result, an increasing number of

studies now rely on corpora as a primary source of data. Although corpus-based methodologies have grown in sophistication, the use of corpus data remains associated with several unresolved challenges (Arppe et al, 2010). In addition, creative inspiration often originates from analyzing textual corpora, which provides the vocabulary, expressions, and thematic ideas necessary for creative writing. This approach enhances lexical appropriateness and encourages linguistic complexity, as creative language tasks require users to "reanalyze and combine known utterances and structures to create new ideas and forms" (Tin, 2012, 2011). Using corpora in activities like unfinished stories, inventive problem-solving, and transitioning from controlled to free writing fosters gradual improvement in creative expression (McDonald et al., 1997, cited by Arshavskaya, 2015; Mansoor, 2010; Melvita, 2023). The pedagogical use of corpora bridges creative exploration and structured learning by encouraging the synthesis of ideas, cultural expression, and divergent thinking (Herawati, 2021). Through systematic study and innovative strategies like the "six words" technique or narrative perspective shifts, students can significantly enhance their creative writing abilities (Tin, 2011; Le, 2018). Language learners can transform their expressive abilities into an authentic, stimulating, and autonomous process through activities that develop both language proficiency and creative thinking. In foreign language education, we aim to inspire students with meaningful, authentic material, encouraging them to express ideas and emotions effectively while using language with precision and originality. Emphasizing authenticity, communication, interaction, research, and creativity, we can foster innovative teaching methods and collaboration, helping students create unique, meaningful work in the foreign language (Tyrou, 2024).
In Greece, within the context of higher education, and specifically at the National and Kapodistrian University of Athens, in the Department of Italian Language and Literature, an initiative was launched to integrate these two fields. As a result, learner corpora were transformed into learner corpora of creative writing, aimed at addressing research questions related to the teaching of Italian as a foreign language, the structure of creative writing in a foreign language, and cross-linguistic analysis. This corpus, the collection and development of which began as a project in 2019 and continues to the present day, is continually enriched with data derived from written productions collected in the Creative Writing course. Hence the name of the corpus, SCRICREA, which is derived from the Italian term "Scrittura Creativa" (Creative Writing).

2. CONSTRUCTION OF THE CORPUS

The case of SCRICREA represents an instance of a Learner Corpus (LC) that aligns with the parameters of learner corpora. Specifically, it documents the interlanguage of learners while simultaneously serving as a resource for extracting samples of authentic language use in the context of creative writing. It consists exclusively of written output produced by students of Italian as a foreign language, created within the framework of creative writing activities conducted in an academic environment.

2.1 The learners

The dataset comprises 339 individuals. The students' ages range from 18 to 52, indicating an adult learner population. These individuals are students of Italian Language and Literature at the University of Athens. All participants are either native Greek speakers or have Greek as their first language. Regarding specific audience characteristics, all students are proficient in at least one foreign language, typically English, which they learned prior to studying Italian. A significant proportion also knows an additional foreign language, making Italian their second or third foreign language. Approximately one third of the students hold prior university degrees. They are enrolled in the 4th semester of their studies, and their language proficiency level in Italian language ranges between B2 and C1 according to the Common European Framework of Reference for Languages (CEFR).

2.2 The texts of the corpus

The written outputs in creative writing stem from a prompt. The topic is provided and discussed with the students, and the writing is produced subsequently. Given this structure, there is no fixed time frame for completing the tasks; however, students are allotted one week from the assignment of the topic to submit their written production. Similarly, there are no specific restrictions on text length, such as a prescribed word count. Nevertheless, there are shared elements, including the language, the genre of the weekly text, and the initial prompt.
The commonality within texts from the same unit lies in the textual genre (e.g., fairy tale, song, poem) and the writing technique employed (e.g., continuation of a story, adaptation with new characters). The

overarching shared characteristic across all units is the authentic language produced by students under the same educational conditions within the framework of creative writing.

2.2.1 The topics/prompts

From the above, it becomes evident that it is crucial to divide SCRICREA into sub-corpora based on the topic of each unit, as this determines both the type of discourse and the textual genre. The division is made according to the weekly topic, which also corresponds to the subject of each lesson, as outlined below (see List 1):

LESSON	Topic/prompt
1st lesson	Continuation of the story (The Young Crab- "il giovane gambero")
2nd lesson	Inventing" a little man made of any material and making him act, creating relationships and randomness depending on the material he is made of (e.g., glass, plastic, wood, ice cream, etc.) (James of Crystal- "Giacomo di Cristallo")
3rd lesson	Imaginary interview of the protagonist or another character from the story, or a brief calendar with text elements, or the recording of the autobiography of the story's hero (The Well of Cascina Piana- "Il pozzo di Cascina Piana")
4th lesson	Changing the point of view, or interpolation within the story (The Beautiful Stranger- "la bella sconosciuta")
5th lesson	A short musical-themed film for writing a story (based on the theme of the video) aimed at young or older children. Alternatively, the choice of a "friend-object" for the protagonist of the story and writing a text for children, or narrating the story with the "voice" of the object-protagonist (a balloon) (Way on Clouds)
6th lesson	Genre change, or using techniques of insertions/additions in the story, or the "What would happen if..." technique (The Man Who Stole the Colosseum- "l'uomo che rubava il Colosseo")
7th lesson	Text alteration technique - replacing words (changing the verb in the title) (The King Who Had to Die- "il re che doveva morire")
8th lesson	Selection of a song and writing a story that precedes or follows the theme of the song, or free-writing techniques

List 1:Topics and tasks and prompt

2.2.2 Outcomes of the activities

In the first activity, students created their own endings to a given story to enhance language proficiency and creative expression. The task encouraged innovation by integrating personal experiences, imagination, emotional insights and expectations into the narrative. By incorporating diverse perspectives, the activity aimed to reveal how students interpreted the story and emphasized different aspects. Ultimately, it sought to foster narrative control, boosting creative confidence and ease in using the foreign language.
The second activity aimed to enhance students' language, critical thinking, and creativity. Students were expected to use descriptive language to detail material properties (e.g., "hard as stone," "fragile as glass") and action-oriented vocabulary (e.g., "the wooden figure creaks" or "the straw figure is vulnerable to wind"). They were to create stories that highlighted the figures' unique properties (e.g., a glass figure avoiding conflict due to fragility). Symbolic meanings related to materials (e.g., marble for strength, wood for naturalness) were also expected to emerge. Finally, students were encouraged to explore deeper themes, such as using strengths and managing weaknesses.
In the third activity, students were expected to demonstrate an understanding of the protagonist's motives, thoughts, and emotions through their writing, highlighting key aspects of their life or actions. If they choose to write an interview, they should explore unknown aspects of the story or imagine post-story events, using vocabulary related to emotions and philosophy. In writing pages from the protagonist's

diary, students were expected to adopt a direct, emotional tone, reflecting on their thoughts and dilemmas. If writing an imaginative autobiography, students were to create a cohesive narrative linking events from the story with fictional ones, showcasing character evolution and adding details about the protagonist's childhood, goals, or life after the story.
In the fourth activity, students explored the narrative by altering the point of view or intervening in the story. Changing the point of view required understanding different characters' perceptions, adjusting language, tone, and details to fit the narrator's personality while offering alternative explanations. Interventions involved creating new scenes, dialogue, or internal thoughts to deepen character insights or influence the plot. Students could alter the story's flow, add new meanings, or develop secondary characters, while ensuring coherence with the original story and introducing original ideas that enhanced the narrative.
The fifth activity combined creativity, imagination, music, and storytelling to achieve diverse educational objectives. Students used music as inspiration to convey emotions, tone, and plot, reflecting the mood, style, or message of the piece (e.g., joy, nostalgia, adventure). They employed metaphors, descriptions, and dialogues or monologues to connect emotions with actions and scenes. For narratives involving a companion-object or the "voice" of an object, students personified the object by attributing human qualities (e.g., personality, emotions, and voice). These stories explored relationships and themes such as loneliness, friendship, loss, or dreams. Students were expected to create original, engaging stories that combined music, imagination, and emotions, with a logical narrative flow. Additionally, themes relevant to children (e.g., dreams, friendship, adventure, innocence) were anticipated to emerge.
The "What if..." technique encourages students to explore alternative scenarios, fostering both creative and critical thinking. Students are expected to develop varied plots (e.g., a thief stealing a different historical monument, or the protagonist stealing the Colosseum to save Rome), moral dilemmas (e.g., choosing between saving the Colosseum or protecting their family), and shifts in time (e.g., the Colosseum already being stolen in a past era or the story set in the future). Through this activity, students will practice hypothetical reasoning and imaginative language, demonstrating originality and creativity in their alternative narratives.
In the seventh activity, students use verb substitution, particularly in the title, to change the premise, tone, and direction of the story. This technique sparks creativity and leads to entirely new narratives. Students are expected to adjust the plot to align with the new title, introducing new themes that may shift the narrative's message or focus. Additionally, students will explore various genres (e.g., drama, fantasy, comedy, thriller), examining the consequences of the verb change on plot, characters, and tone. They will also create original scenarios and future possibilities for the story, ensuring coherence, originality, linguistic accuracy, and thematic development in their writing.
In the final activity, students combine music and creative writing to explore the emotions, narratives, and ideas evoked by a song. They are expected to create an original story linked to the song, imagining its beginning, emotions, or what happens after its end. Students should draw themes and images from the song and transform abstract feelings into specific scenes, characters, or dialogues. Alternatively, through free writing, students can write without constraints, letting their thoughts flow based on musical inspiration. This encourages the expression of personal feelings and ideas in various styles (e.g., narrative, poetic, or essayistic), fostering creativity without concern for right or wrong.

2.3 The size of the Corpus

The size of SCRICREA is approximately 1 million words. Its articulation and the sizes of its sub-corpora are presented in the table below. As shown, the corpus is composed of sub-corpora derived from eight different topics, as they have been developed over the past six academic years (see Table 1):

ACAD. YEAR	1° unità	2a unità	3° unità	4° unità	5a unità	6a unità	7a unità	8a unità (canzoni)	TOTAL
2019	13.951	34.995	36.664	34.400	33.261	25.101	19.600	26.451	224.423
2020	17.584	33.853	38.404	37.250	36.893	28.681	37.878	29.451	259.994
2021	19.018	37.399	40.765	41.853	42.008	34.416	23.121	39.965	278.545
2022	8.485	20.524	19.966	18.302	20.244	17.396	17.390	16.307	138.614

2023	6.974	10.833	10.415	11.507	11.489	5.552	10.308	8.411	75.489
2024	7.255	9.716	10.878	8.693	7.562	8.934	9.470	9.248	71.756
TOTAL	73.267	147.320	157.092	152.005	151.457	120.080	117.767	129.833	1.048.821

Table 1: SCRICREA, Corpus and subcorpora

Although the number of students remains constant for each task within the same academic year, the size of the sub-corpora varies due to differences in the nature of the tasks. Regarding the overall size of each corpus per year, it differs because of the varying number of students who enroll and submit assignments. In recent years, there has been a declining rate of student admissions to the department, which is also reflected in the size of the corpus.

3. APPLICATIONS

The SCRICREA corpus is a paradigmatic example of interdisciplinarity applied to language research and education. The project integrates the perspectives of corpus linguistics, language education and creativity studies, enabling an articulated analysis of language learning processes. The incorporation of narrative techniques, imagination exercises and a variety of textual genres serves to activate students' cognitive, linguistic and expressive skills, thus laying the foundation for a holistic and innovative teaching approach. The pedagogical and instructional applications of corpora have been extensively explored in prior research. Such studies have investigated, for instance, the frequency of verb usage (Ringdom, 1998), the use of connectives (Altenberg & Tapper, 1998), vocabulary (Lenko-Szymanska, 2005), prepositions (Diez Bedmar & Casas Pedrosa, 2006), and modal verbs (Aijmer, 2002). A corpus of this kind, however, offers a range of additional advantages. It acts as a repository of data and ideas for students of Italian and creative writing, while simultaneously functioning as a learner corpus that facilitates the analysis of students' interlanguage and serves as a basis for error analysis. Notably, parts of SCRICREA have already been employed for such purposes. For the analysis of n-grams and their frequency, we employed sub-corpora corresponding to topics 2, 5, 6, and 7, covering the period 2019–2022 (Florou & Tyrou, 2024). Some initial conclusions have already been drawn regarding the usefulness of n-gram measurements and analyses in the study of learner interlanguage, both from the aforementioned study and from other research focused on learner corpora (Gries, 2015; Pezic, 2025). Nevertheless, this specific corpus is not only composed of texts produced by learners of Italian language but is also intended for their use. It could serve as a space for peer experimentation through contrastive text analysis and error analysis. Furthermore, if shared with other educational settings (e.g., secondary education schools where Italian is taught), the corpus may function as a valuable resource: both as a model for creative writing and as a repository of interlanguage data from learners of Italian whose first language is Greek. Finaly, SCRICREA, due to its extensive nature, gives us the opportunity to use NLP tools to statistically investigate the corpus as well as the language of students in Italian. Indicators such as readability and type/token ratio can serve as supplementary measures not only in recognizing the student's level of proficiency in the Italian language but also in detecting qualitative characteristics such as imagination and creativity.

ACKNOWLEDGEMENTS

We would like to express our heartfelt gratitude, above all, to the students of our department for granting us permission to use their written works and for the opportunity they provide us to research the subject that will be taught to future generations of students. We would also like to thank the anonymous reviewers for their comments.

REFERENCES

Aijmer, K. (2002). Modality in advanced Swedish learners' written interlanguage. In S. Granger, J. Hung, & S. Petch-Tyson (Eds.), *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching, pp. 55-76.* John Benjamins Publishing Company, Amsterdam/ Philadelphia. https://doi.org/10.1075/lllt.6.07aij

Altenberg, B., & Tapper, M. (1998). *The use of adverbial connectors in advanced Swedish learners'*. In S. Granger (Ed.), *Learner English on Computer*, pp. 80-93. Longman, London and New York.

Arppe, A., Gilquin, G., Glynn, D., Hilpert, M., & Zeschel, A. (2010). Cognitive Corpus Linguistics: Five points of debate on current theory and methodology. In Corpora 5. pp. 1-27. https://doi.org/10.3366/cor.2010.0001

Arshavskaya, Ekaterina. (2015). Creative Writing Assignments in a Second Language Course: A Way to Engage Less Motivated Students. *InSight: A Journal of Scholarly Teaching*. 10. pp. 68-78. 10.46504/10201506ar.

Diez Bedmar, M., & Casas Pedrosa, V. (2006). The use of prepositions by Spanish learners of English at University level: Main problems. In *7th Conference on Teaching and Language Corpora (TaLC) Proccedings*, pp. 42-44, Paris.

Florou, K., & Tyrou, I. (2024). La particella "ne" nella scrittura creativa in italiano LS. *Indagine su un corpus di apprendenti greci. Umanistica Digitale*, (18), pp. 87-101. http://dx.doi.org/10.6092/issn.2532-8816/19939

Granger, S. (2004). Computer Learner Corpus Research: Curent Status and Futur Prospects. In M. Conn, & T. Upton (Eds.), *Applied Corpus Linguistics: A Multidimensional Perspective*, pp. 123-145. Amsterdam and Atlanda: Rodopi. https://doi.org/10.1163/9789004333772_008

Gries, S. T. (2015). Statistics for learner corpus research. *The Cambridge handbook of learner corpus research*, pp. 159-182.

Herawati, H. (2021). Learners as Story Writers: Creative Writing Practices in English as a Foreign Language Learning in Indonesia. In: Bao, D., Pham, T. (eds*) Transforming Pedagogies Through Engagement with Learners, Teachers and Communities*. Education in the Asia-Pacific Region: Issues, Concerns and Prospects, 57. Springer, Singapore. https://doi.org/10.1007/978-981-16-0057-9_5

Le, P. T. (2018). Using six-word stories to trigger EFL learners' creative writing skills. Indonesian Journal of English Language Teaching, 13, pp. 175-188. https://doi.org/10.25170/ijelt.v13i2.1456

Lenko-Szymanska, A. (2005). The role of L1 influence and L2 instruction in the choice of rhetorical strategies by EFL learners. In B. Lewandowska-Tomaszczyk (Ed.), *PALC'2005*. Practical Applications in Language Corpora. Peter Lang.

Mansoor, A. (2014). Ekphrastic practices in catalysing creative writing in undergraduate ESL classrooms. *New Writing: The International Journal for the Practice and Theory of Creative Writing*. https://doi.org/10.1080/14790726.2014.904887

McDonald, B. A., Rosselli, J. A., & Clifford, J. E. (1997). Journal writing: Learning, reflections, and adjustments to American life. *ERIC Document Reproduction Service* No. ED 423 709.

Melvita, S. (2023). Implementing unfinished story to build students' creative writing. *Journal of Educational Study*, 3, pp. 25-31. https://doi.org/10.36663/joes.v3i1.436

McEnery, T., & Hardie, A. (2012). *Corpus Linguistics: Method, theory and practice*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511981395

Pęzik, P. (2015). Using n-gram independence to identify discourse-functional lexical units in spoken learner corpus data. *International Journal of Learner Corpus Research (IJLCR)*, *1*(2).

Ringdom, H. (1998). Vocabulary frequencies in advanced learner English: A cross-linguistic approach. In S. Granger (Ed.), *Learner English on Computer*, pp. 41-52, Longman, London and New York.

Tin, T. B. (2011). Language creativity and co-emergence of form and meaning in creative writing tasks. *Applied Linguistics*, 32 (2), pp. 215- 235.

Tin, T. B. (2012). Freedom, constraints and creativity in language learning tasks: new task features. *Innovations in Language Learning and Teaching*, 6 (2), pp. 177-186.

Tyrou, I. (2024). Suggestions and prompts of foreign language activities in expressive writing. *International Journal of Education*, 12 (2).

Retrieval-Augmented Generation systems for enhanced access to digital archives

Michele Ciletti[1]
[1] University of Foggia – michele_ciletti.587188@unifg.it

ABSTRACT (ENGLISH)

This study investigates the potential of Retrieval-Augmented Generation (RAG) to enhance access to digital archives related to humanities topics, using a case study on historical newspaper data. Over the past years, Large Language Models (LLMs) have shown remarkable capabilities in generating natural language; however, they still suffer from hallucinations and context window limitations. As a solution, RAG frameworks have emerged, linking a language model to external databases in order to provide grounded, reliable sources. In this work, an experimental RAG pipeline was tested on the "Foggia Occupator Dataset", which includes articles from a 1945–1946 periodical published by the US forces occupying Foggia, Italy, at the end of World War II. To assess performance, domain experts constructed a set of ten questions on topics such as historical figures, examined events, and stylistic aspects of the articles. The corresponding "ground truth" answers were used for both quantitative and qualitative evaluations. Results show that the RAG system performs well in precisely defined scenarios, retrieving accurate information and correctly identifying named entities. However, broader questions occasionally led to incomplete or slightly erroneous answers, pointing to potential areas for algorithmic refinement—particularly in optimizing retrieval for complex or multi-article queries. The study concludes that RAG can significantly improve the searchability and reliability of digital archives, but continued improvements in metadata enrichment, query strategies, and retrieval algorithms are needed.

Keywords: Retrieval Augmented Generation; Artificial Intelligence; archiving; newspapers; Large Language Models

ABSTRACT (ITALIAN)

Retrieval-Augmented Generation per un migliore accesso agli archivi digitali. Questo studio analizza il potenziale della Retrieval-Augmented Generation (RAG) per migliorare l'accesso ad archivi digitali relativi a tematiche umanistiche, utilizzando un caso di studio sui dati di un giornale storico. Negli ultimi anni, i Large Language Models (LLM) hanno dimostrato notevoli capacità di generare linguaggio naturale; tuttavia, soffrono ancora di allucinazioni e limitazioni delle finestre di contesto. Come soluzione, sono emersi i framework RAG, che collegano un modello linguistico a database esterni per fornire fonti fondate e affidabili. In questo studio, una pipeline RAG sperimentale è stata testata sul "Foggia Occupator Dataset", che comprende articoli di un periodico del 1945-1946 pubblicato dalle forze statunitensi che occupavano Foggia, in Italia, alla fine della Seconda Guerra Mondiale. Per valutarne le prestazioni, degli esperti hanno realizzato una serie di dieci domande su argomenti quali personaggi storici, eventi esaminati e aspetti stilistici degli articoli. Le corrispondenti risposte di "ground truth" sono state utilizzate per delle valutazioni quantitative e qualitative. I risultati mostrano che il sistema RAG ottiene ottimi risultati in scenari definiti con precisione, recuperando informazioni accurate e identificando correttamente le entità nominate. Tuttavia, domande più ampie hanno occasionalmente portato a risposte incomplete o leggermente errate, indicando potenziali aree di perfezionamento dell'algoritmo, in particolare nell'ottimizzazione del reperimento di query complesse o multiarticolate. Lo studio conclude che la RAG può migliorare significativamente la ricercabilità e l'affidabilità degli archivi digitali, ma sono necessari continui miglioramenti nell'arricchimento dei metadati, nelle strategie di interrogazione e negli algoritmi di recupero.

Keywords: Retrieval Augmented Generation; Intelligenza Artificiale; archivistica; giornali; Large Language Models

1. INTRODUCTION

In recent years, we are witnessing a growing interest in Large Language Models, both from the scientific community and the general population. Since OpenAI's GPT-3.5 model demonstrated unprecedented natural language generation capabilities in November 2022, considerable efforts in terms of time and funding have been undertaken to develop new, increasingly high-performance models. Today, the latest models obtain excellent results on specific tasks, including programming and solving mathematical problems (Jiang et al., 2024; Li et al., 2024).

The limitations of Large Language Models, however, persist. Parallel to the systematic expansion and improvement of the amount of training data, work is being done on the design of new frameworks and techniques that can mitigate the inherent flaws of LLMs. Sometimes accused of being little more than "stochastic parrots", Large Language Models produce content on the basis of statistical patterns derived from their training data (Bender et al., 2021). The absence of internal logic in favour of linguistic consistency results in a tendency to hallucinate, i.e. to produce untrue information confidently presented as factual (Perkovic et al., 2024). This leads to considerable risks, as users may fail to notice an unreported error in a large amount of text, and consequently spread misinformation.
Some companies have tackled the problem by expanding their context windows, or by equipping their models with the ability to surf the Internet in order to refine their output. Promising, but still embryonic, appears instead the new tendency to produce specialised models on reasoning, which reflect on their output and make corrections before presenting them to the user (Jaech et al., 2024). Another strategy, particularly appropriate for the application of LLM to third-party contexts (companies, public bodies), is Retrieval-Augmented Generation (Gao et al., 2023). The technique makes it possible to "link" a language model to a database of proprietary data, which may consist of a series of digitised books, a list of products on sale, a series of services available to a customer, or much more: what matters is that this database is converted into embeddings, multidimensional vectors useful for encoding natural language in a computational manner. In practice, when a user queries the model, their query is also converted into embeddings, which are then compared with the database: the most semantically similar documents - usually identified by calculating the cosine similarity between the vectors - are added to the input and passed to the model, which answers the query on the basis of the information contained in the documents themselves. This solution can provide a concrete and reliable database for the language model, without limiting its generative capabilities.
These considerations give rise to the research questions of this paper: can Retrieval-Augmented Generation be implemented in digital archives to facilitate the search and consultation of their materials? What are the guidelines to follow and good practices to adopt? Can sufficiently reliable and informative results be obtained in this way? To provide answers, it was decided to carry out a case study on a recently created database: the Foggia Occupator Dataset, which contains articles published between 1945 and 1946 in the Foggia Occupator, a generalist newspaper printed by the US forces that occupied the city of Foggia in the closing years of World War II (Ciletti et al., 2024). While similar experiments have been tried across various fields (Gupta et al., 2024), this study aims to provide a specific perspective on archives related to the humanities, while also focusing on elaborating insights for practical implementations.

2. METHODOLOGY

The chosen dataset contains 874 articles, totalling 216,015 words. Its content, which is not only historical, but also social and customary, lends itself favourably to thorough investigations and presents elements that require in-depth analysis to be fully interpreted. Its very recent publication also means that the likelihood of an LLM containing fragments of it in its training dataset is almost non-existent. However, possible limitations lie in the narrow scope of the articles, which deal with a short period, with events that took place in a restricted geographical area and are written entirely in a single language (English). The aim, therefore, is to offer an example case that can provide guidelines and practical feedback in terms of performance for those wishing to implement similar solutions in other contexts, rather than to fully evaluate the effectiveness and convenience of RAG across multiple realities.
The articles constituting the dataset, originally collected in a JSON file with simple metadata, were embedded using the text-embedding-3-large model (OpenAI, 2024) and stored in a Chroma database, an AI-native open-source vector storage solution (Chroma, 2025). The database was then linked to the OpenAI API. The RAG system was implemented via a script, which, once a query or request was received as input, converts it into embedding via the same model and identifies the five documents with the closest cosine similarity to the input. The number is indicative and was chosen in consultation with the experts involved in the creation of the dataset, based on an estimate of a reasonable number of relevant documents for each given topic. The request and documents are then sent together with a pre-determined system prompt to OpenAI's GPT-4o model (Hurst et al., 2024), which produces a response. It was decided to choose one of the best performing models at the time of writing, with the aim of testing the real capabilities of an advanced LLM under optimal conditions. This was done in the belief that, in a short time, the general performance of LLMs will be further improved and that equally satisfactory results will be

obtainable at a lower cost and, perhaps, even through open-source models. However, future studies comparing multiple models are certainly desirable.
To test the effectiveness of the answers generated, experts were asked to produce a set of questions and answers concerning certain topics covered in the dataset, accompanied by a variable number of specific relevant articles. In drafting the question-and-answer pairs, the need to investigate multiple aspects was taken into account, namely:

- The ability to identify and provide information on real people, political affairs and social events;
- The ability to reflect on the causes of events and political and social situations in a holistic manner;
- The ability to analyse and provide information on linguistic and stylistic choices.

An attempt was also made to produce questions on specific topics related to the history of the city and the newspaper, avoiding in most cases overly general topics on which there might be information in the model's traning dataset. In any case, a very low temperature was set in the model settings to ensure that the answers were as deterministic and reproducible as possible, as well as based as much as possible on the information contained in the documents. The set of questions and answers (n=10) is outlined in Table 1.

Query ID	Question	Expert answer	Model answer
1	Who was Savino Bufalo?	Savino Bufalo was the photographer of the Foggia Occupator.	Savino Bufalo was a staff photographer who took photos at events such as the Foggia Fair and other local happenings.
2	Who authored the column 'Static from your AES'?	The column 'Static from your AES' was authored by Ken Eill, however, in a single issue it was authored by Bob Dewart.	The column 'Static from your AES' was authored by Cpl. Ken Eill.
3	What were the main events of the 1945 Christmas celebrations in Foggia?	During the 1945 Christmas celebrations, Christmas trees were brought into town, soldiers dressed up as Santa Claus and gave gifts to children staying at the hospital. The EM Red Cross organised most of these celebrations. A party at the gym was also organised. An orchestra led by Frank De Grazia performed musical pieces across those days. The amount of children showing up to the gift distributions created some unrest.	The main events of the 1945 Christmas celebrations in Foggia included parties for children, where gifts were given to the sons of prisoners of war, and these celebrations continued through Christmas Day. The American celebrations featured Franck DeGrazia and his orchestra, who performed at various venues around town, including night spots and serenades on Christmas Eve. There were also Christmas dinners and dances that involved many local women and guests, including important Foggia civil officials and high-ranking Allied Forces officers. Additionally, Frank De Grazia and his group went on a serenade, spreading the Spirit of Christmas with their music throughout the town.
4	How was the Italian educational system structured in 1945-1946?	In 1945-1946, the Italian educational system saw children start elementary school at age 6, possibly after kindergarten. Then, there was the possibility to attend a vocational or classical school. Vocational schools were attended by the poorer children, who were then ready to go to work. Poverty, especially in Southern Italy, meant that many	The main events of the 1945 Christmas celebrations in Foggia included parties for children, where gifts were given to the sons of prisoners of war, and these celebrations continued through Christmas Day. The American celebrations featured Franck DeGrazia and his orchestra, who performed at various venues around town, including night spots and serenades on Christmas Eve. There were also Christmas dinners and dances that involved many local women and guests, including important Foggia civil officials and high-ranking Allied Forces officers. Additionally, Frank De Grazia and his group went on a serenade, spreading the Spirit of Christmas with their music throughout the town.

Query ID	Question	Expert answer	Model answer
		children did not attend compulsory school.	
5	What kinds of sports were practised in Foggia in 1945-1946?	The main sports practiced in Foggia in 1945-1946 were baseball, basketball and boxing, all of which had structured tournaments and thorough match reports in the Foggia Occupator. However, softball and handball were practiced too.	The main events of the 1945 Christmas celebrations in Foggia included parties for children, where gifts were given to the sons of prisoners of war, and these celebrations continued through Christmas Day. The American celebrations featured Franck DeGrazia and his orchestra, who performed at various venues around town, including night spots and serenades on Christmas Eve. There were also Christmas dinners and dances that involved many local women and guests, including important Foggia civil officials and high-ranking Allied Forces officers. Additionally, Frank De Grazia and his group went on a serenade, spreading the Spirit of Christmas with their music throughout the town.
6	Did protests, riots or violent events ever happen in Foggia, and if so, why?	Protests and riots happened, mainly because of rifts between locals and soldiers, strikes, food prices, requisitions and grain shortages.	Yes, protests and riots have occurred in Foggia. On March 31, there was an all-day rioting incident where one Italian was killed and eighteen were injured. Thousands of people, led by Foggia building workers, attacked a train and searched the city for black-market grain. This unrest was caused by an extreme shortage of grain in Southern Italy. The building workers were protesting for food and work, and the demonstrators also entered the local ration board and burned the records.
7	What was the political situation like in Italy, and especially in Foggia, in June 1946, during the days of the institutional referendum?	In June 1946, Italy voted for the establishment of the republic in a narrow referendum. The outcome was celebrated in Foggia, even though the monarchy won in the province by a considerable margin. While there was civic unrest in Italy, the situation in Foggia remained peaceful. Representatives from several parties were elected to the National Constituent Assembly. The Christian Democrats had the most, but there were also Socialists, Qualunquists and Communists.	In June 1946, Italy was undergoing a significant political transition. The country was recovering from World War II and the fall of the Fascist regime. A key event during this period was the institutional referendum held on June 2, 1946, where Italians voted to decide between maintaining the monarchy or establishing a republic. This was a crucial moment in Italian history, as it led to the abolition of the monarchy and the establishment of the Italian Republic. In Foggia, the political situation reflected the broader national context. The region participated in the elections for the National Constituent Assembly, which was tasked with drafting a new constitution for Italy. The election results in Foggia Province showed a diverse political landscape, with representation going to Christian Democrats, Communists, Socialists, and Qualunquists. Despite the national vote favoring the monarchy, Foggia Province had a majority voting for the monarchy by more than two to one. However, the transition to a republic proceeded without major incidents in the area, as indicated by the peaceful nature of public events like parades.
8	What does ‘Siponto’ mean, and in what context is it mentioned?	‘Siponto’ is the name of a maritime location near Foggia. It is mentioned in the Foggia Occupator because it was a popular destination for civilians and soldiers looking for leisure, who	"Siponto" refers to a location where there are clubs, specifically the Siponto Officers Club and the Siponto Beach Club. In the context provided, it is mentioned as a place where club managers are preparing for the summer season by constructing outdoor terraces for relaxation and building a new pier at the Siponto Beach Club. Additionally, discussions and activities related to the Italian Peace Treaty and Army vehicles are taking place at the Siponto Beach Club.

Query ID	Question	Expert answer	Model answer
		often went to the Siponto Beach Club.	
9	What are the contents of the column 'Nights' Round Foggia, and who is its author?	The column 'Nights 'Round Foggia' gives information about various entertainment events in Foggia, such as club events, concerts, parties and shows.	The column "Nights 'Round Foggia" is authored by Eddie M. Dunaj. It includes an account of the grand opening of the Castle Bar, a new officers' nightclub in Foggia. The column also features what club managers have to say about the appearance of "Nights 'Round Foggia" from the previous week.
10	How would you describe the writing style of the 'Girl About Town' column?	The column 'Girl About Town' is written in an informal and playful tone, reflecting its contents which concern social life in Foggia and local gossip.	The writing style of the column "Girl About Town" appears to be light-hearted and conversational, with a focus on social tidbits and gossip. The column seems to have a playful tone, as indicated by the repetition of the title and the mention of the writer's eagerness to please the editors or fill up space. It suggests a casual and engaging approach, likely aimed at entertaining readers with social news and anecdotes.

Table 1. The queries and answers set

All queries were subjected to the model. To assess the effectiveness in selecting relevant documents, classic measures such as precision and recall appeared inappropriate due to the incompatibility between the fixed number of documents selected for each query (5) and the variable number of relevant documents for each specific query selected by the experts. To overcome this, the selection of documents for each application was classified according to the following formula: (Number of relevant documents retrieved) / (min(Total number of relevant documents, 5)).
Secondly, the quality of the responses generated was assessed in relation to those provided by the experts. A first assessment was made by calculating the cosine similarity between the embedding of the answers generated by the model and those of the experts. Subsequently, the experts themselves qualitatively rated each question generated by the model on a scale from 1 (not at all satisfactory) to 5 (completely satisfactory), also providing a brief comment on the strengths and weaknesses of each one. Automatic evaluation systems, such as RAGAs (Es et al., 2024), were considered but ultimately discarded because of the manageable size of the questions and answers set. However, such approaches may be beneficial when dealing with larger datasets and future studies are encouraged to implement them.

3. RESULTS AND DISCUSSION

The performance of the RAG system is displayed in Figure 1. Overall, the cosine similarity between the model and expert responses is consistently high, with a mean of 0.92668 and a standard deviation of 0.031383. Retrieval performance is perfect in half of the cases, while it has a value of 0.8 in two others and 0.6 in three others. Three answers were rated as completely satisfactory by the experts (5 out of 5), four others scored 4 out of 5, two scored 3 out of 5 and one scored 2 out of 5. No response was rated as not at all satisfactory (1 out of 5).
Analysing the answers given by the model in relation to ground truth and the experts' comments, clear strengths and weaknesses appear. The model excels in the more specific answers, with clearly delineated areas and themes: query 4, which refers to a long article on the Italian education system, is particularly precise and detailed. On the contrary, broader queries, which ask for an overall analysis on phenomena spanning several articles and periods, sometimes lack precision or completeness. This is the case with query 3, which received the lowest score from the experts because, in the course of a long and otherwise valid answer, it placed a social event of February 1946 during the Christmas celebrations of 1945. This is the only case of completely wrong information that was found. Query 6, on the other hand, produced an accurate answer, but limited in its analysis, as it analysed a single protest event, when there were many others in the database. Problems of this kind are certainly attributable to the retrieval algorithm, which in this experiment is rather basic and for demonstration purposes: a weighting system of certain keyword-related embeddings could certainly mitigate the situation. The word 'Foggia', for example, should be weighted negatively, since almost all articles refer to the city context, but the algorithm, when the name of

the city is specified in the query, is inclined to prioritise documents that explicitly mention it. Not being able to rely so much on the users' ability to construct optimal prompts (which is why there was no intervention on the queries written by experts), algorithmic performance is certainly a point to invest in. Likewise, it was observed that the addition of a system prompt to guide the model significantly improves understanding of the general context in which the model operates.
On the other hand, an excellent performance was observed in interpreting stylistic elements of a given group of articles, especially with regard to query 10. Furthermore, a better performance was noted in queries that referred to specifically named headings, which leads one to think that the results can be further improved with the implementation of accurate and complete metadata, a practice that is already widely encouraged in the construction of digital archives.

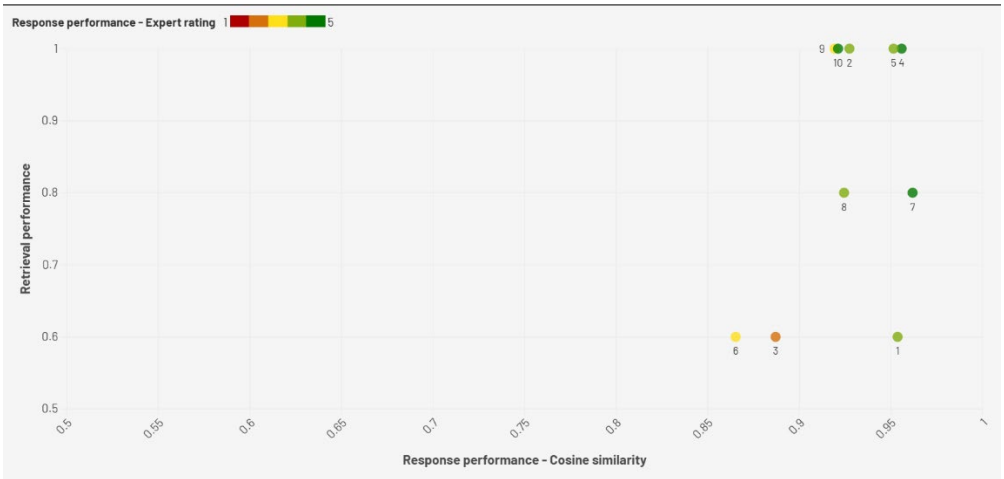


Figure 1. Performance analysis of the RAG system

4. CONCLUSIONS

Through the application of a RAG system to the Foggia Occupator dataset, the strengths and weaknesses of such approaches in making digital archives more accessible and easily searchable were reflected upon. Clear opportunities arise in making the contents of archives more immediate and effective, but doubts remain about accuracy and reliability. Points to work on to mitigate these problems are highlighted, mainly the enrichment of metadata, algorithmic refinement and the judicious choice of contexts to which to apply the systems. Further research testing this method against larger, more diverse text corpora are also desirable.
From the point of view of practical implementation, substantial obstacles remain in the costs of using LLM and the technical requirements needed to build, and then host, RAG-based systems. However, it has been demonstrated that moderate initial deployments are feasible even with relatively modest resources, and the hope is that costs and computational requirements will continue to decrease. Likewise, the proliferation of educational content on machine learning in general, which is often freely accessible, can only foster the development of digital skills necessary for the application of LLM-based systems, even by people who are not necessarily formally trained in computer science, such as staff in the GLAM sector. Effective implementation still requires competency across multiple domains including search engines, embeddings, document preprocessing, and LLM configuration—which is why collaboration between domain experts and technical scientist remains fundamental to bridge knowledge gaps. This collaborative approach becomes more feasible as technical literacy becomes increasingly attainable through accessible learning resources. It is hoped that future research will continue to analyse techniques and strategies to foster the implementation of machine learning technologies in humanistic contexts, with a focus on practical aspects and possible methodological improvements.

ACKNOWLEDGEMENTS

The authors have no competing interests to declare that are relevant to the content of this article.

REFERENCES

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? 🦜. In Proceedings of the 2021 ACM conference on fairness, accountability, and transparency (pp. 610-623).

Ciletti, M., di Furia, M., Guarini, P., & Toto, G. A. (2024). Foggia Occupator: a case study on the creation of an Open Educational Resource through the digitization of a historical newspaper. In HELMeTO 2024 Book of Abstracts (p. 235).

Chroma. (2025). Chroma: The AI-native open-source embedding database [Computer software]. GitHub. https://github.com/chroma-core/chroma

Es, S., James, J., Anke, L. E., & Schockaert, S. (2024). Ragas: Automated evaluation of retrieval augmented generation. In Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations (pp. 150-158).

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Guo, Q., Wang, M., & Wang, H. (2023). Retrieval-Augmented Generation for Large Language Models: A Survey. ArXiv, abs/2312.10997. https://doi.org/10.48550/arXiv.2312.10997.

Gupta, A., Shirgaonkar, A., Balaguer, A. D. L., Silva, B., Holstein, D., Li, D., ... & Benara, V. (2024). RAG vs Fine-tuning: Pipelines, Tradeoffs, and a Case Study on Agriculture. arXiv preprint arXiv:2401.08406.

Hurst, A., Lerer, A., Goucher, A. P., Perelman, A., Ramesh, A., Clark, A., ... & Kivlichan, I. (2024). Gpt-4o system card. arXiv preprint arXiv:2410.21276.

Jaech, A., Kalai, A., Lerer, A., Richardson, A., El-Kishky, A., Low, A., ... & Kaiser, L. (2024). OpenAI o1 System Card. arXiv preprint arXiv:2412.16720.

Jiang, J., Wang, F., Shen, J., Kim, S., & Kim, S. (2024). A Survey on Large Language Models for Code Generation. ArXiv, abs/2406.00515. https://doi.org/10.48550/arXiv.2406.00515.

Li, Q., Cui, L., Zhao, X., Kong, L., & Bi, W. (2024). GSM-Plus: A Comprehensive Benchmark for Evaluating the Robustness of LLMs as Mathematical Problem Solvers., 2961-2984. https://doi.org/10.48550/arXiv.2402.19255.

OpenAI. (2024). GPT-4o API. Retrieved from https://platform.openai.com/docs/models/gpt-4-and-gpt-4-turbo.

OpenAI. (2024). text-embedding-3-large. Retrieved from https://platform.openai.com/docs/guides/embeddings.

Perkovic, G., Drobnjak, A., & Boticki, I. (2024). Hallucinations in LLMs: Understanding and Addressing Challenges. 2024 47th MIPRO ICT and Electronics Convention (MIPRO), 2084-2088. https://doi.org/10.1109/MIPRO60963.2024.10569238.

Preserving Clarity: The MAGIC project approach to ancient manuscripts

Yahya Momtaz[1], Stefania Conte[2], Guido Russo[3]
[1] University of Naples Federico II, Italy – yahya.momtaz@unina.it
[2] University of Naples Federico II, Italy – stefania.conte@unina.it
[3] University of Naples Federico II, Italy – guido.russo@unina.it

ABSTRACT (ENGLISH)
The MAGIC project, launched in 2023, integrates advanced AI methodologies and web development to open access to the cultural assets, mainly ancient books. The project maintains access to early printed books and manuscripts through the use of high-resolution photography and bleed-through correction techniques. The project's web platform is multilingual and provides responsive design in order to make access to all the digital devices and allow global audiences to access historical documents with ease. This study explains the techniques of the MAGIC project, which connect historical and today's environments while promoting a simple and accessible digital humanities environment.
Keywords: cultural heritage; Bleed-through; React web development; digital preservation; manuscript digitization

ABSTRACT (ITALIANO)
*Preservare la chiarezza: l'approccio del progetto MAGIC ai manoscritti antichi.*Il progetto MAGIC, lanciato nel 2023, integra metodologie di intelligenza artificiale avanzate e sviluppo web per aprire l'accesso ai beni culturali. Il progetto mantiene l'accesso ai primi libri e manoscritti stampati attraverso l'uso di fotografie ad alta risoluzione e tecniche di correzione bleed-through. La piattaforma web del progetto è multilingue e offre un design reattivo per rendere accessibili tutti i dispositivi digitali e consentire al pubblico globale di accedere facilmente ai documenti storici. Questo studio spiega le tecniche del progetto MAGIC, che collegano gli ambienti storici e odierni promuovendo al contempo un ambiente di digital humanities semplice e accessibile.
Parole chiave: patrimonio culturale; Bleed-through; sviluppo web React; conservazione digitale; digitalizzazione di manoscritti

1. INTRODUCTION
Digitizing historic manuscripts is crucial for keeping hold of cultural heritage, increasing access to historical texts, and protecting them for future generations. However, various challenges remain, particularly regarding the restoration of damaged texts and maintaining their accessibility through current digital channels (Russo et al., 2020; Conte et al., 2024). The MAGIC project addresses these issues by combining advanced picture restoration techniques with a dynamic, React-based web platform to build an inclusive and interactive digital library. In order to preserve historical writings and provide access to cultural information, the project combines advanced AI approaches with accessible web development. React, a popular JavaScript library for developing interactive UI, is used to develop a web platform that supports global access to these digitized assets. Primary features include high-resolution image display via the International Image Interoperability Framework (IIIF), enabled by Mirador, a viewer designed for the exploration of digitized manuscripts. Additionally, the website offers multilingual support for diverse audiences and has a responsive design suitable for various digital devices. The research focuses on a newly started study addressing the cleanup of manuscripts affected by bleed-through, the ongoing deterioration issue caused by ink spreading through paper. This issue often harms text clarity, reducing the readability of scanned images. The MAGIC project uses a multi-phase, AI-driven approach that integrates Gaussian Mixture Models (GMMs) for segmentation and Non-Local Means (NLM) denoising for text reconstruction. This research emphasizes the combination of AI-based image processing next to recent methods to produce accessible digital copies of ancient books and manuscripts. Beyond restoring the images, this project aims to make these digital assets more accessible and usable to both the academic community and the general public through the use of appropriate technologies. In addition, it employs AI and image processing techniques to improve text readability while reducing potential harm to the source materials.

2. WEBSITE DEVELOPMENT AND FEATURES

React.js is a popular JavaScript library known for its component-based architecture. This architecture allows developers to break down the user interface into reusable, self-contained units called components, which manage their own state and logic. Components can be composed together to build complex UIs, promoting modularity, code reusability, and easier maintenance. React.js also leverages a virtual DOM (Document Object Model) to optimize performance (see Fig. 1). The virtual DOM acts as an intermediary representation of the actual DOM, allowing React.js to efficiently update only the necessary parts of the UI when changes occur (Kolomoyets & Kynash, 2023, p. 3).

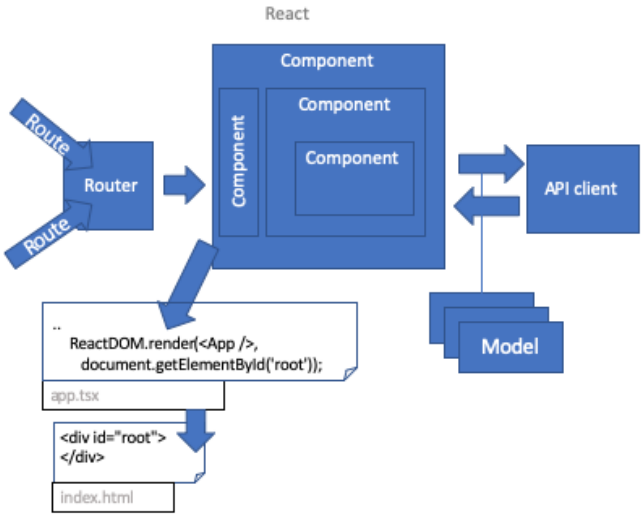


Figure 1. React architecture
(Copyright of https://handsonreact.com/docs/architecture)

The platform guarantees smooth functionality and accessibility. These architecture decisions boost the usefulness and scalability of the MAGIC platform while highlighting its dedication to preserving cultural heritage in an accessible manner for worldwide audiences. The Container and Presentational Components Pattern organizes components into logical (data-handling) and visual (UI-rendering) roles, providing clear and engaging descriptive sections for each manuscript. The Atomic Design Pattern organizes components hierarchically, facilitating fast rendering of multi-resolution JPEG 2000 images and allowing for the further analysis of text and ornaments.



Figure 2. The Magic website

The website is developed with a responsive design (see Fig. 2) and is using Bootstrap and Tailwind CSS for a consistent user experience across all devices. This responsiveness allows users to access the platform precisely from desktop computers, tablets, and smartphones. On the other hand, Framer Motion makes user interaction with the website more engaging by including smooth animations and transitions. The

platform includes Mirador, an advanced viewer for International Image Interoperability Framework (IIIF) resources, which enables users to view images at multiple resolutions, magnify small details, and look at specific sections of a text with maximum accuracy.

3. ARTIFICIAL INTELLIGENCE IN THE MAGIC PROJECT

The MAGIC project has just started using artificial intelligence (AI) to address key challenges in the protection and digitization of historical documents. A primary concern is to reduce bleed-through, a common issue when ink from one side of a page passes to the opposite side; it covers the text and decreases clarity. This issue is especially common in early manuscripts, where preservation quality deeply impacts scholarly research and public access (Conte et al., 2024). The project employs an AI-driven pipeline for bleed-through correction by improving image quality in addition to the historical integrity of the manuscripts. The procedure, now in an experimental stage, starts with segmentation, when the manuscript photos are divided into sections, separating text, background, and ornaments. Subsequently, thresholding is implemented through unsupervised machine learning methods such as Gaussian Mixture Models (GMMs) to evaluate and classify pixels according to their characteristics, which is essential for differentiating text from background noise and other components (Hanif et al., 2022). The third stage involves cleaning; it works with techniques like Non-Local Means (NLM) for denoising and inpainting to restore damaged areas by reconstructing the background texture and also to ensure a seamless restoration.

Nation	Number of Manuscripts
Vatican city	11
Italy	72
France	13
England	2
Spain	1
USA	1
Austria	1
Total	**101**

Table 1. dataset

Our dataset (see Table 1) is based on a collection of approximately 100 medieval manuscripts of Dante Alighieri's Divine Comedy that have been fully digitized over the past five years by the University of Naples. The manuscripts originate from various parts of the world, including the Vatican City, Italy, France, England, Spain, the United States, and Austria, each of which contributed one book.

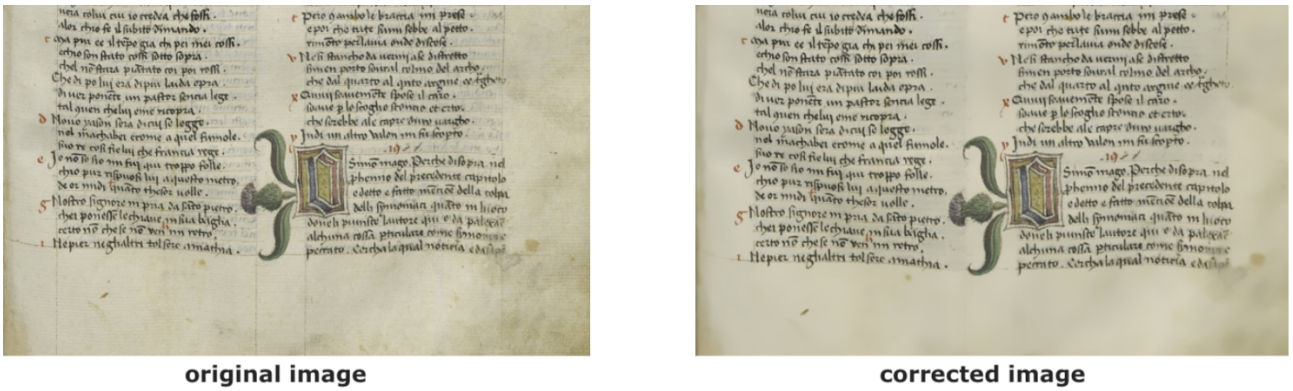


Figure 3. comparison of original image and the corrected image by bleed-through
(Original image: Firenze, Biblioteca medicea laurenziana, Plut.40.1, sheet 62r,
Corrected image is under Copyright of University of Naples Federico II, Dept. of Humanistic Studies)

These AI-driven techniques boost the visual clarity of text and their readability (see Fig. 3) for the public by combining image processing methodologies with a responsive React-based web platform. MAGIC

guarantees that restored texts are accessible to a varied, worldwide audience. The aspect of the MAGIC project highlights its value. The project integrates AI and digital humanities to tackle significant issues in manuscript preservation while promoting diversity. This method preserves cultural heritage while simultaneously expanding access to historical information for academics, educators, and the public.

4. CONCLUSION

The MAGIC project applies AI-driven image processing and React-based web development to improve historical document accessibility and preservation. The research boosts the clarity of scanned manuscripts by combining segmentation, thresholding, and cleaning methods, therefore addressing issues including bleed-through while preserving their visual and written integrity. With its simple interface, the React-based platform lets global viewers investigate these cultural gems without consider for geography.
Beyond preservation, MAGIC supports academic study, multidisciplinary cooperation, and public interaction with historical materials. This project shows how modern technology may transform digital humanities to guarantee the availability and lifetime of old knowledge for the next generations.

ACKNOWLEDGEMENTS

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

REFERENCES

Novac, C. M., Novac, O. C., Sferle, R. M., Gordan, M. I., Bujdosó, G., & Dindelegan, C. M. (2021). Comparative study of some applications made in the Vue.js and React.js frameworks. *2021 16th International Conference on Engineering of Modern Electric Systems (EMES)*, Oradea, Romania, pp. 1–4. https://doi.org/10.1109/EMES52337.2021.9484149

Kolomoyets, M., & Kynash, Y. (2023). Front-End web development project architecture design. *Proceedings of the Conference on Computer Science and Information Technology (CSIT)*, pp. 1–5. https://doi.org/10.1109/CSIT61576.2023.10324238

Hanif, M., Tonazzini, A., Hussain, S. F., Habib, U., Salerno, E., Savino, P., & Halim, Z. (2022). Blind bleed-through removal in color ancient manuscripts. *Multimedia Tools and Applications, 82*(8), 12321–12335. https://doi.org/10.1007/s11042-022-13755-6

Russo, G., Aiosa, L., Alfano, G., Chianese, A., Cornevilli, F., Di Domenico, G. M., Maddalena, P. M., Mazzucchi, A., Muraglia, C., Russillo, F., Salvi, A., Spisso, B., Trombetti, G., & Zollo, G. (2020). MA.G.I.C.: Manuscripts of Girolamini in cloud. IOP Conference Series: Materials Science and Engineering, 949, 012081, 1–8. https://doi.org/10.1088/1757-899X/949/1/012081

Conte, S., Mazzucchi, A., Russo, G., Tortora, A., & Tortora, G. (2024). The organization and management of the MAGIC project for ancient manuscripts digitization: connections between Mediterranean cultures. *MeTe Digital: Mediterranean Networks Between Texts and Contexts. Proceedings of the XIII AIUCD Conference*, Catania, Italy, pp. 38–42.

Conte, S., Ferrante, G., Laccetti, L., Mazzucchi, A., Momtaz, Y., & Tortora, A. (Eds.). (2024). Content representation and analysis: The Magic Project and the Illuminated Dante Project integrated systems for multimedia information retrieval. In K. Roitero, M. Viviani, E. Maddalena, & S. Mizzaro (Eds.), Proceedings of the 14th Italian Information Retrieval Workshop (pp. 66–69). CEUR Workshop Proceedings.

Digital Explorations of Historical Multilingual Practices. The Challenges of the HyperAzpilcueta Project

Manuela Bragagnolo[1], Marcus Pöckelmann[2], Polina Solonets[3], Andreas Wagner[4]
[1] University of Trento, Italy – manuela.bragagnolo@unitn.it
[2] Martin Luther University Halle-Wittenberg, Germany – marcus.poeckelmann@informatik.uni-halle.de
[3] Max Planck Institute for Legal History and Legal Theory, Germany – solonets@lhlt.mpg.de
[4] Max Planck Institute for Legal History and Legal Theory, Germany – wagner@lhlt.mpg.de

ABSTRACT (ENGLISH)
Based on the example of an early modern case of self-translation and collaborative translation, the contribution seeks to open discussion about the relation between multilingual practices embodied in textual research objects on the one side and the "user experience" of digital platforms and their affordances on the other.
Keywords: text collation; digital multilingual practices; HyperAzpilcueta

ABSTRACT (ITALIANO)
Esplorazioni digitali di pratiche multi linguistiche storiche. Le sfide del progetto HyperAzpilcueta
Basandosi sull'esempio di un caso di autotraduzione e traduzione collaborativa della prima età moderna, il contributo cerca di aprire una discussione sulla relazione tra le pratiche multilinguistiche incarnate negli oggetti di ricerca testuale, da un lato, e la "user experience" delle piattaforme digitali e delle loro possibilità, dall'altro.
Parole chiave: collazione di testi; pratiche digitali multilinguistiche; HyperAzpilcueta

1. INTRODUCTION
Historical multilingual practices involve much more than mere linguistic translation. For this reason, conceptualizing the main aspects of those practices and their application in different textual witnesses is a challenge in itself. But the question then is: how can digital environments and methods facilitate investigation of the various aspects of these practices and their differences?
As an illustration of the general problem, we present "HyperAzpilcueta", a digital synoptic edition of several printed editions of an early modern bestseller, *Manual de Confessores* (Manual for Confessors). Its author, the Spanish canon law professor Martín de Azpilcueta (1492-1586), translated and simultaneously substantially revised his work several times, incorporating different types of authorities, examples and arguments in the different editions. Together with his printers, he even organized typographical devices for his respective target audience to better orient themselves in the text. With the help of translations and revisions, he thus devised new ways of producing different types of knowledge of (religious, spiritual, but also legal) normativity.

2. MARTÍN DE AZPILCUETA'S MANUAL DE CONFESSORES AS A CASE OF MULTILINGUAL PRACTICE
First printed in Portugal in 1549, the *Manual* was then transformed, updated, and translated into different languages by Azpilcueta himself, in a revision process that lasted for almost 40 years. Four of the editions showcasing the most important transformations are used in the HyperAzpilcueta project: the first edition, printed in Coimbra in Portuguese in 1549; the second Portuguese edition, with major transformations (Coimbra, Portuguese, 1552); the 1556 Spanish translation by Azpilcueta, largely enriched (Salamanca, 1556); and finally, the Latin translation, completely revised in structure (Rome, 1573).
As Azpilcueta was working on translation alone or in cooperation with other scholars and practitioners of the time, his work on the *Manual* is an example of both, "collaborative translation" (Cordingley & Frigau Manning, 2017) and "self-translation" practices (Miglietti, 2024). Such a multilingual practice was quite common in the early modern world, with every new edition used to update, restructure, and adapt the text (and the book as a material object) to new target audiences, new contexts of production and circulation, and to changes in societal practices that form the subject matter of many parts of the book.
In what dimensions can we study the translation practice of the *Manual*?
Firstly, the differences between the texts reflect the varying contexts of writing, book production and dissemination. They actually "accompanied" Azpilcueta's moving from Portugal (Coimbra, 1549; Coimbra,

1552) to Spain (Salamanca, 1556), and from Spain to Rome (Rome, 1573). Every new edition and "self-" or "collaborative" translation was produced in close contact with a different set of local actors (Bragagnolo, 2022; Bragagnolo, 2024): throughout his career, missionaries and confessors have contacted Azpilcueta with questions and doubts about the latest (or earlier) edition(s) of the *Manual*, as did kings and penitents, theologians and jurists, not to forget book printers and traders. Economic and political constellations varied locally and changed substantially during the 40 years of revising the *Manual*, and the Council of Trent (1545–1563) revised in important ways the official Catholic doctrine on many questions treated in it. Therefore, each new version incorporated the feedback of a particular community of practice in its revised and translated language.

Secondly, the different versions of the *Manual* aimed to reach different reader groups: The first edition (in 8° pragmatic vernacular handbook) was meant for simple and unlearned confessors, active in remote parishes of Portuguese provinces. By contrast, the very erudite, rich, in 4° Latin *Enchiridion* was intended to reach learned confessors and missionaries of the universal Church, widespread all over the world, via academic channels of communication and book distribution. The editions in between reflected this path in their language, from the vernacular to Latin, with injections of doctrine, and increasingly technical, complex and erudite use of language.

Those differences between contexts of production and target audiences found expression in concrete linguistic phenomena: different audiences demand different sets of references to authorities, different contexts provide different sets of examples (mundane or contrived, colonial or courtly), different languages have different proverbs and personae. And in terms of knowledge organization, Azpilcueta developed different ways of mapping and navigating the subject matter: even though the basic structure of 27 chapters was kept intact in all editions, each edition shows another ordering of arguments, examples and references, mostly with new sections added. Different structural and typographical finding devices such as sub-chapter headings and summaria, the wrapping of newly added passages in asterisks or their introduction by daggers etc. suggest very different uses of the book by the readers. Some of them, such as the marginal numbers, even seem to indicate an (again, multilingual) practice of navigating across and comparing the different editions on the part of the readers.

3. CONCEPTUALIZING THE COMPREHENSION OF MULTILINGUAL PRACTICE IN A DIGITAL PLATFORM

The question arises of how a digital platform can facilitate in the study of the various described aspects of the *Manual* and texts with a similar history of tradition. The following section discusses the main requirements and the possibilities already offered by existing approaches.

First, it is essential to allow researchers to closely examine the text itself by presenting it in full detail whenever necessary (close-reading). This should also take into account – may even by optional highlighting – their various special elements such as different fonts, marginal numbers, literature references or special characters (e.g. asterisks and daggers in our case). If the texts are encoded in TEI XML, this is supported by many tools for generating digital editions such as TEI-Publisher,[1] EVT[2] (Rosselli Del Turco et al., 2014) or the Versioning Machine[3] (Schreibman et al., 2003).

A second essential feature is the possibility to compare certain passages of the text witnesses with each other in order to allow researchers to identify similarities and textual variants. Typical close-reading representations for this purpose in collation tools highlight or arrange the words in specific ways. This includes classic synoptic views, which can be found in the already mentioned tools (TEI-Publisher, EVT and Versioning Machine) as well as in LERA (Pöckelmann et al., 2022), or variant graphs as in CollateX (Dekker et al., 2015) or TRAViz (Jänicke et al., 2015). However, the underlying collation algorithms in current tools are typically designed for witnesses of the same language and rather simple textual variants like the replacement of individual words or phrases with alternatives. While research for handling multiple languages in the domain of scholarly editions is presently going on (Balck et al,. 2025; Reboul, 2024; Hohenegger et al., 2023), available collation tools still lack the ability to automatically point the researcher to semantic similarities and differences on a word or phrase level when it comes to multilingual texts.

When dealing with larger amounts of text like the *Manual* has, the collation in hierarchical steps is reasonable (Pöckelmann et al., 2022), i.e., in a first step larger segments of the text such as sentences or

[1] https://teipublisher.com/exist/apps/tei-publisher-home/index.html (cons. 05.01.2025).
[2] http://evt.labcd.unipi.it (cons. 05.01.2025).
[3] http://v-machine.org (cons. 05.01.2025).

paragraphs are aligned before in a second step the collation on the word level is performed. This presupposes the constitution and alignment of those segments in the first place, which becomes harder the more substantial the revisions of the text witnesses are. Besides the presentation and user interface, features of automatic segmentation and detection of semantic similar phrases or sentences that differ only on the surface, i.e., mostly in a changed wording but not meaning, are relevant for the platforms or toolchains we are interested in. Promising recent solutions outside the domain of scholarly editing are based on word or sentence embeddings (e.g. Kusner et al., 2015). While these approaches have been intended for monolingual applications at first, they were soon be extended to the multilingual realm (Reimers & Gurevych, 2020; Ham & Kim, 2021), resulting in state-of-art approaches for effective sentence-alignment even across languages, e.g. SentAlign (Steingrímmsson et al., 2023). However, these have difficulties when dealing with major structural differences among the text witnesses, as it is the case with the *Manual*. Another problem lies in the training of the underlying embeddings, which requires a large amount of digitized text. This is simply not available as we face so-called *low-resource languages* in this project due to the historical state and the specific genre. In our alternative approach to embeddings, we apply a rule-based system for segmentation with an alignment algorithm that profits from the similarity of the languages in the *Manual* (Wagner & Bragagnolo, 2019) in combination with an evaluation of some of the particular features mentioned above and are currently extending the collation platform LERA[4] (Pöckelmann et al., 2022) accordingly.

Even if the used algorithms deliver good results, it is unlikely that any automatic segmentation or alignment will be perfect at extracting and evaluating all the relevant features nor that the results correspond exactly to the researcher's expectations. This weakness is amplified by the fact that often changes in one aspect (like segmentation) entail substantial further changes in another (alignment). Thus, a digital platform must provide means for researchers to correct both segmentation and alignment in the later close reading process. It should be possible to join or split segments, move them around (in the sense of adjusting their position in the synoptic view), and even place copies of existing segments – marked as such – at other positions to reflect transpositions.

Finally, the question arises as to how the entirety of the work and the identified revisions can be presented. With the large volume of text on the one hand and the structural changes in terms of both sequencing and volume on the other, there is a need for overview visualizations (distant-reading) in order to navigate through these volumes of text as well as to compare structural changes. The latter may include transpositions that span over long distances within the different text witnesses. A recent survey (Yousef & Jänicke, 2021) lists a huge variety of visualization methods that has been developed over the years to represent alignments, some of which are applicable to huge text witnesses for distant-reading. Furthermore, some older approaches have focused on conveying the dynamics of revision processes or structural aspects (e.g. Fry, 2009; Cheesman & Roos, 2017). However, due to the combination of various aspects of the *Manual* – multilinguality, restructuring and the mere existence of more than two witnesses – no existing visualization is sufficient. A flexible, zoomable, or elastic way of providing overviews of different levels of distance seems desirable (as in Bludau et al., 2020; Pöckelmann et al., 2024). This could include the text itself at the lowest zoom level, making it a scalable-reading approach.

4. CONCLUSION

As we have shown, for many of the mentioned requirements, some functionalities are, in principle, available in existing tools. Yet the combination of tasks (represent structural as well as semantic changes, establish alignment, facilitate close reading, process etc.), the volume of text to be processed and presented and the multitude of features to take into account – and the multilinguality affecting all the tasks individually and in combination – mean that there is, as of now, no platform that really does facilitate study of historical multilingual practices like we have presented with Apilcueta's *Manual*. Moreover, to the best of our knowledge, for the issue of distinguishing "salient" from "trivial" differences between texts in different languages (where practically everything is different in terms of linguistic surface forms), no solution at all is available. The state of affairs rather seems to demand a discussion about how to conceptualize "salient semantic similarity or difference" in a multilingual context, how to anchor any mechanism in this regard to concrete words or phrases in the source documents, and (only then) how to convey eventually acquired measurements in a user interface.

[4] https://lera.uzi.uni-halle.de (cons. 05.01.2025).

REFERENCES

Balck, S., Dähne, J., Etling, F., Fischer, F., Frenzel, S., Grote, B., Heße, S., Molitor, P., Pöckelmann, M., Ritter, J., Singh, Y. & Stede, M. (2025). *Collation of Multilingual Versions of a Text: Necessity, Approach, Challenges* [Conference presentation abstract]. ADHO DH2025. Lisbon; Portugal.

Bludau, M.-J., Brüggemann, V., Busch, A. & Dörk, M. (2020). Reading Traces: Scalable Exploration in Elastic Visualizations of Cultural Heritage Data. *Computer Graphics Forum, 39*(3), 77-87. https://doi.org/10.1111/cgf.13964.

Bragagnolo, M. (2022). Les voyages du droit du Portugal à Rome. Le Manual de confessores de Martín de Azpilcueta (1492-1586) et ses traductions. In J.-L. Fournel, I. Paccagnella (Eds.), *Traduire. Tradurre. Translating. Vie des mots et voies des œuvres dans l'Europe de la Renaissance* (pp. 209-228), Droz.

Bragagnolo, M. (Ed.). (2024). *The Production of Knowledge of Normativity in the Age of the Printing Press. Martín de Azpilcueta's Manual de Confessores in global perspective*, Brill. https://doi.org/10.1163/9789004687042

Cheesman, T. & Roos, A., (2017). Version Variation Visualization (VVV): Case Studies on the Hebrew Haggadah in English. In M. Büchler & L. Mellerin (Eds.), Computer-Aided Processing of Intertextuality in Ancient Languages [Special Issue]. *Journal of Data Mining & Digital Humanities*. https://doi.org/10.46298/jdmdh.3704

Cordingley, A. & Frigau Manning, C. (Eds.) (2017). *Collaborative Translation: from the Renaissance to the Digital Age*, Bloomsbury Academic. https://doi.org/10.5040/9781350006034

Dekker, R., Van Hulle, D., Middell, G., Neyt, V. & Van Zundert, J. (2015). Computer-supported collation of modern manuscripts: CollateX and the Beckett Digital Manuscript Project. *Digital Scholarship in the Humanities*, 30(3), 452-470. https://doi.org/10.1093/llc/fqu007

Fry, B. (2009) *On the Origin of Species: The Preservation of Favoured Traces.* https://benfry.com/traces/

Ham, J., Kim, E.-S. (2021). Semantic alignment with calibrated similarity for multilingual sentence embedding. In M.-F. Moens, X. Huang, L. Specia & S. W. Yih (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2021* (pp. 1781-1791). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-emnlp.153

Fry, B. (2009) *On the Origin of Species: The Preservation of Favoured Traces.* https://benfry.com/traces/

Hohenegger, H., Montemagni, S., Pasini, E., Ciotti, F., Mancinelli, T., De Longis, E., Chiaravalli, I., Boschetti, F., Del Grosso, A. M. & Meschini, F. (2023) *DiScEPT – Digital Scholarly Editions Platform and aligned Translations. https://www.studigermanici.it/progetto-discept*

Jänicke, S., Geßner, A., Franzini, G., Terras, M., Mahony, S. & Scheuermann, G. (2015). TRAViz: A Visualization for Variant Graphs. *Digital Scholarship in the Humanities*, *30*(suppl_1), i83–i99. https://doi.org/10.1093/llc/fqv049

Kusner, M. J., Sun, Y., Kolkin, N. I. & Weinberger, K. Q. (2015). From Word Embeddings to Document Distances. In F. Bach & D. Blei (Eds.), *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*. (pp. 957-966). https://dl.acm.org/doi/10.5555/3045118.3045221

Miglietti, S. (2024). Translations of self-translations in early modern Europe: preliminary remarks and exploratory case studies. *Philological Quarterly, 102*(2-3), 199-223.

Pöckelmann, M., Damm, A. & Rebiger, B. (2024). *An Interactive Visualization of Recurring Microforms Across Various Hebrew Works and their Witnesses* [Conference presentation abstract]. DH2024, 34. international annual conference of Digital Humanities, Washington D.C., United States.

Pöckelmann, M., Medek, A., Ritter, J. & Molitor, P. (2022). LERA—an interactive platform for synoptical representations of multiple text witnesses. *Digital Scholarship in the Humanities, 38*(1), 330–346. https://doi.org/10.1093/llc/fqac021

Reboul, M. (2024). Alignement sémantique et manque de données : l'apport des modèles de langue. Le cas du latin et du grec. In B. Bordalejo, R. Risam & E. Château-Dutier (Eds.), DH Unbound 2022, Selected Papers [Special Issue]. *Digital Studies/Le champ numérique 13*(3): 1–13. https://doi.org/10.16995/dscn.9669.

Reimers, N. & Gurevych, I. (2020). Making monolingual sentence embeddings multilingual using knowledge distillation. In B. Webber, T. Cohn, Y. He & Y. Liu. (Eds.), *Proceedings of the 2020*

Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 4512-4525). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.365

Rosselli Del Turco, R., Buomprisco, G., Di Pietro, C., Kenny, J., Masotti, R. & Pugliese, J. (2014). Edition Visualization Technology: A Simple Tool to Visualize TEI-based Digital Editions. *Journal of the Text Encoding Initiative, 8*. https://doi.org/10.4000/jtei.1077

Schreibman, S., Kumar, A. & McDonald, J. (2003). The versioning machine. *Literary and Linguistic Computing, 18*(1), 101–107. https://doi.org/10.1093/llc/18.1.101

Steingrímmsson, S., Loftsson, H. & Way, A. (2023). SentAlign: Accurate and scalable sentence alignment. In Y. Feng & E. Lefever (Eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 256-263) https://doi.org/10.18653/v1/2023.emnlp-demo.22

Wagner, A. & Bragagnolo, M. (2019). *Multimodale Versuche der Alignierung historischer Texte.* [Conference presentation abstract] DHd2019, 6. Jahrestagung der Digital Humanities im deutschsprachigen Raum. Frankfurt, Germany, 181–184. https://doi.org/10.5281/zenodo.4622224.

Yousef, Y. & Jänicke, S. (2021). A Survey of Text Alignment Visualization. *IEEE Transactions on Visualization and Computer Graphics*, *27*(2), 1149–1159. https://doi.org/10.1109/TVCG.2020.3028975

Digitalizzazione di un fondo archivistico per la creazione di un centro di documentazione digitale

Dario Baldini
Università degli Studi di Roma Tor Vergata/Fondazione Giannino Bassetti, Italia - baldini.dario98@gmail.com

ABSTRACT (ITALIANO)

Questo progetto di dottorato è incentrato sulla digitalizzazione di un fondo archivistico per creare un centro di documentazione consultabile con metodi informatici, legati principalmente a tecnologie web e di intelligenza artificiale. Innanzitutto, le carte sono state scansionate e i file di immagine sono elaborati da *Image* API dello standard IIIF per l'invio delle immagini e delle informazioni tecniche in JSON; esse sono state programmate con una applicazione Python scritta in Flask. Le scansioni sono state strutturate, descritte e annotate tramite *Presentation* API, file JSON che poi vengono elaborati dal *viewer* Mirador 3 per generare un'interfaccia web per la fruizione dei documenti digitalizzati; viene inizializzato e configurato dinamicamente in un *template* Jinja2. Dopo di che, il testo dei documenti è stato trascritto con sistemi OCR e codificato nello schema XML/TEI in modo molto granulare, marcando specifiche porzioni testuali per registrarne il valore semantico. È stato poi programmato un complesso *style sheet* XSL per la trasformazione dei file XML in XHTML, in modo che i documenti codificati possano essere visualizzati sul web, che il loro aspetto grafico possa essere definito con *style sheet* CSS e che possano essere arricchiti di funzionalità interattive con JavaScript. Successivamente, la ricerca si concentrerà sulle possibilità di applicazione dei sistemi di AI ai prodotti digitali ottenuti durante le fasi precedenti, utilizzando LLM *open-source* in locale grazie a specifiche piattaforme. Verranno implementate funzionalità interattive al centro di documentazione per elaborazioni testuali come traduzione e sintesi. Verrà programmata una interfaccia per il recupero delle informazioni basata su sistemi di RAG per estrarre i dati semantici della codifica testuale. Si valuterà la fattibilità del *fine-tuning* per permettere a un modello di riconoscere le componenti testuali, anche per l'automatizzazione delle operazioni di codifica.
Parole chiave: Archivio digitale; *International Image Interoperability Framework* (IIIF); *Text Encoding Initiative* (TEI); sviluppo web dinamico; intelligenza artificiale

ABSTRACT (ENGLISH)

Digitisation of an archive collection for the creation of a digital documentation centre

This PhD project focuses on the digitisation of an archive collection to create a documentation centre that can be consulted with IT methods, mainly related to web and artificial intelligence technologies. First, the papers were scanned, and the image files are processed by Image APIs of the IIIF standard to send images and technical information in JSON; they have been programmed using a Python application written in Flask. The scans have been structured, described and annotated via Presentation APIs, JSON files that are then processed by the Mirador 3 viewer to generate a web interface for the use of the digitised documents; it is initialized and configured dynamically in a Jinja2 template. After that, the text of the documents has been transcribed with OCR systems and encoded meticulously in the XML/TEI schema, marking specific textual portions to record their semantic value. A complex XSL style sheet has then been programmed to transform the XML files into XHTML, so that the encoded documents can be displayed on the web, their graphic appearance can be defined with CSS style sheets and they can be enriched with interactive features with JavaScript. Subsequently, the research will focus on the possibilities of applying AI systems to the digital products obtained during the previous phases, using open-source LLM locally thanks to specific platforms. Interactive features will be implemented at the documentation centre for text processing such as translation and synthesis. An information retrieval interface based on RAG systems will be programmed to extract semantic data from text encoding. The feasibility of fine-tuning will be evaluated to allow a model to recognize textual components, also to automate the encoding process.
Keywords: digital archive; International Image Interoperability Framework (IIIF); Text Encoding Initiative (TEI); dynamic web development; artificial intelligence

L'obiettivo di questo progetto di dottorato è la digitalizzazione di un fondo archivistico per la creazione di un centro di documentazione consultabile con metodi informatici, legati principalmente a tecnologie web e di intelligenza artificiale. Il processo di digitalizzazione consiste nell'applicazione di due standard

internazionali molto noti e utilizzati nel campo delle *Digital Humanities*, IIIF e TEI, e di una serie di sistemi di sviluppo web dinamico.

1. Digitalizzazione dei documenti

Il progetto può essere suddiviso in tre fasi ben distinte: la prima riguarda la scansione dei documenti con uno scanner professionale e la loro codifica negli standard IIIF e TEI, impiegati proprio per la conservazione e la trasmissione del patrimonio culturale. L'*International Image Interoperability Framework* (IIIF) è un insieme di API sviluppate per la trasmissione di immagini e contenuti multimediali di vario tipo ad ambienti web diversi, garantendo la massima compatibilità (Meschini, 2020, "Una domanda ancora aperta"). Dopo aver acquisito le immagini, i file vengono innanzitutto caricati su server e organizzati in *folder*, e per permetter la loro elaborazione sono stati programmati *Image* API IIIF grazie a un'applicazione Python; quest'ultima utilizza il *micro-framework* Flask, che attraverso il sistema del routing, ovvero mappando specifici URL alle funzioni (Pallets, 2024, "Routing"), permette lo scambio dei dati di tutte le componenti del progetto tramite richieste HTTP. In particolare, sono state programmate due funzioni per i due *web service* definiti dalle linee guida dello *Image* API: lo *Image Request*, per richiedere una rappresentazione dell'immagine, e lo *Image Information Request*, per richiedere informazioni tecniche sull'immagine in formato JSON-LD. Per il primo, è stato mappato uno URI conforme alla sintassi dello *Image Request* a una funzione per elaborare l'immagine. I parametri di prefisso — ossia il percorso — e identificatore vengono estratti dalla richiesta HTTP e utilizzati nella funzione per selezionare l'immagine, mentre gli altri parametri di regione, dimensioni, rotazione e qualità specificano quali operazioni di trasformazione eseguire: la porzione da rendere; come ingrandire o rimpicciolire l'immagine; come ruotarla o specchiarla; come modificarne i colori. L'immagine manipolata viene infine inviata in formato JPEG. Per il secondo, è stato mappato uno URI conforme alla sintassi dello *Image Information Request* a una funzione con un dizionario Python con delle proprietà preconfigurate, ovvero `@context` che fa riferimento allo URI del contesto dello Image API, la versione dell'*Image* API e il livello di *compliance*, l'insieme di possibilità di trasformazione che vengono garantite. Prefisso e identificatore vengono estratti dalla richiesta HTTP e inseriti nel dizionario, insieme alle dimensioni che vengono calcolate analizzando l'immagine. Tutte le informazioni tecniche vengono poi inviate in formato JSON (Appleby et al, 2020a). Dopo di che, le immagini vengono strutturate e descritte attraverso il *Presentation* API, anch'esso da programmare nel formato JSON-LD. La struttura tipica di un *Presentation* API è composta da un *manifest*, che codifica un intero oggetto composito, che contiene in rapporto di *embedding* una serie di *canvas*, dei contenitori digitali che offrono rappresentazioni di parti specifiche dell'oggetto composito. Nel progetto, lettere, note e documentazione varia vengono raccolti insieme, mentre testi più importanti hanno file dedicati. La prima proprietà di un *Presentation* API deve essere `@context`, che rimanda al contesto del *Presentation* API; le uniche proprietà obbligatorie in tutte le risorse sono `id` e `type`. È anche possibile programmare delle collezioni, risorse che raccolgono *manifest* e collezioni subordinate con un rapporto di *referencing*, cioè facendo riferimento alla loro proprietà `id`. All'interno della proprietà `items` delle *canvas* si trovano le pagine di annotazioni, la cui unica funzione è contenere le annotazioni, sistemi di codifica in JSON-LD definiti dal *Web Annotation Data Model* per legare informazioni aggiuntive a risorse online, o per descrivere le associazioni esistenti tra loro (Sanderson et al, 2017); esse sono composte da contenuti audiovisivi o testuali di vario tipo, il *body*, legati a una *canvas*, il *target*, per permetterne la visualizzazione. Per legare le singole scansioni alle *canvas* è necessario utilizzare il valore `painting` nella la proprietà `motivation` dell'annotazione, riportare le informazioni tecniche dell'immagine — tra cui URI di *Image* API — nel *body*, e riprendere la proprietà `id` delle *canvas* stesse come *target*; la proprietà service verrà usata all'interno del *body* per fare riferimento al servizio dell'*Image* API. È anche possibile annotare specifiche parti delle scansioni inserendo le pagine di annotazioni nella proprietà `annotations` delle canvas, utilizzando il valore `painting` nella la proprietà `motivation` dell'annotazione, scrivendo un contenuto testuale nel *body*, e selezionando una porzione di *canvas* come *target*, riprendendo la proprietà `id` seguita da un *FragmentSelector*. Questa tecnica è stata utilizzata per evidenziare e descrivere simboli, firme o altri elementi grafici dei documenti. Infine, le *canvas* possono essere elencate all'interno di un *manifest* in modi alternativi con i *range*. Nel progetto sono stati sfruttati per ricreare digitalmente i sommari dei documenti originali, o per raccogliere i documenti a seconda della tipologia (corrispondenza, note, etc.) (Appleby et al, 2020b).

Successivamente si procede alla trascrizione dei documenti con sistemi OCR e alla codifica testuale con il linguaggio di marcatura definito dalla *Text Encoding Initiative* (TEI), la comunità internazionale che si occupa della gestione, dello sviluppo e della promozione delle TEI *Guidelines*, un insieme standardizzato di

pratiche di codifica per materiale testuale pensato per la produzione di documenti archiviabili e trasmissibili con metodi informatici, soprattutto nell'ottica della conservazione a lungo termine di testi antichi (Sechi, 2010, "Text Encoding Initiative"). Un documento TEI riprende la struttura base e le regole di sintassi di XML: si inserisce la dichiarazione XML alla prima linea, seguita da un solo elemento radice (Sechi, 2010, "Struttura di XML"), marcato con il *tag* `<TEI>`, che racchiude un solo TEI *header*, in cui vengono registrati i metadati, e uno o più elementi in cui riportare il testo dei documenti (TEI Consortium, 2025, "Default Text Structure"). Come per i *manifest*, lettere, note e documentazione varia vengono raccolti in un unico file TEI, mentre testi più importanti hanno file dedicati. Il TEI *header* si compone di cinque parti: *file description*, in cui si descrivono il documento TEI e le fonti e che contiene gli unici elementi veramente obbligatori, ovvero `<titleStmt>`, `<publicationStmt>` e `<sourceDesc>`; *encoding description*, in cui si riassumono pratiche editoriali e metodologie di codifica; *text profile*, in cui si riportano informazioni classificatorie e contestuali; una parte marcata dal *tag* `<xenoData>` in cui inserire metadati in un formato non conforme alle specifiche TEI; *revision history*, in cui si elencano le revisioni del documento (*Ivi*, "The TEI Header"). La marcatura del testo può essere considerata un processo di modellizzazione per come viene concepito nelle *Digital Humanities*, pensato quindi per l'elaborazione elettronica e la rappresentazione dei testi (Ciula et al., 2023, "Modelling Practices"). La segmentazione nelle unità testuali di livello più generale prevede l'utilizzo dell'elemento *division* `<div>`, mentre le vere e proprie porzioni di testo vengono segmentate con elementi di varie categorie: *chunk*, *phrase-level*, *inter-level* e *component*. Tutti i vari fenomeni testuali che devono essere differenziati dal testo che li circonda saranno codificati con i relativi elementi *inline* (TEI Consortium, 2025, "The TEI Infrastructure"): nomi propri di persona, organizzazione e luogo; numeri, misure e date; termini stranieri e tecnici. Allo stesso modo, verranno impiegati anche tutti quegli elementi che rappresentano interventi editoriali, legati al trattamento di abbreviazioni, forme erronee, aggiunte e cancellazioni, ma anche interpretativi per casi in cui il testo risulta illeggibile. In tutti i casi si mira a una codifica abbastanza minuziosa dei fenomeni, e verranno quindi impiegati tutti gli eventuali elementi figlio per marcare le singole parti di testo e tutti gli attributi necessari per codificarne le caratteristiche.

2. Pubblicazione delle risorse digitali

La seconda fase è incentrata sulla pubblicazione online delle risorse digitali create durante la prima fase, e quindi sulla creazione del portale web che ospiterà il centro di documentazione digitale e sull'implementazione di funzionalità pensate per la ricerca. Le API IIIF vengono elaborate da uno specifico *client* per generare un'interfaccia web in cui visualizzare e navigare le scansioni (IIIF Consortium, s.d., "Viewing objects"). È stato definito un *template* Jinja2 in cui viene inizializzata un'istanza di Mirador 3 — *viewer* IIIF che utilizza la nota libreria React — inserendo un riferimento allo URL in cui è archiviato il codice JavaScript del client. Nel *template* viene chiamata la specifica funzione `Mirador.viewer()` passando come parametro un oggetto JavaScript le cui proprietà servono a configurare il *viewer*; qui si inseriscono variabili che saranno estratte dinamicamente dalla richiesta HTTP. Infatti, nell'applicazione Flask si mappa uno URL con i parametri di percorso e identificatore a una funzione per il *renderig* del *template* Jinja2; questi parametri vengono estratti per costruire dinamicamente lo URL di un *Presentation* API, il *template* viene renderizzato passando lo URL come variabile e il *viewer* viene visualizzato nel browser (Helm, 2024, "Embedding Mirador"). Alla stessa funzione è stato mappato anche uno URL simile al precedente ma con l'aggiunta di un frammento che indica l'identificatore di una singola *canvas*; esso permetterà di caricare il *viewer* su una specifica *canvas* grazie alla funzione `Mirador.actions.setCanvas()` (*Ivi*, "Embedding Mirador"). Per visualizzare le trascrizioni TEI in pagine web, è stato programmato un complesso *style sheet* XSL in cui si inserisce la dichiarazione XML e l'elemento radice marcato con il tag `<xsl:stylesheet>`. Si definiscono *template* per trasformare gli elementi XML dati in input con elementi `<xsl:template>`: un *template* generale seleziona l'elemento radice per produrre un documento XHTML; *template* specifici selezionano i singoli elementi, anche con istruzioni condizionali. Il contenuto testuale di elementi e attributi viene così estratto, elaborato e aggiunto all'output XHTML. I file vengono trasformati mappando uno URL con i parametri di percorso e identificatore nell'applicazione Flask a una specifica funzione che analizza e trasforma in oggetti XML i file XML e lo *style sheet* XSL. Dopo di che, l'oggetto dello *style sheet* XSL viene convertito in un oggetto "trasformatore XSLT" e applicato all'oggetto del documento XML (Behnel, 2024, "XPath and XSLT with lxml"). Il documento trasformato in XHTML viene così inviato all'utente e la trascrizione visualizzata nel browser. Per permettere la navigazione del centro di documentazione digitale è stata creata una *homepage* per il portale da cui selezionare i documenti, disponibile online al seguente link: https://centrodocumentazione.fondazionebassetti.org/. Essa si basa su un *template* Jinja2 che

permette la generazione dinamica di schede descrittive dei documenti analizzando i file TEI ed estraendo da essi informazioni come titolo, autore, data, lingua e numero di pagine, oltre ai dati necessari per costruire gli URL per visualizzare *viewer* e trascrizioni. Nel momento di scrittura di questo *paper*, la *homepage* presenta quattro raggruppamenti tematici predefiniti per l'accesso ai documenti digitalizzati: pensiero europeista, Comune di Milano, Commissione Trilaterale e IPALMO. Successivamente verranno aggiunte delle funzionalità di ricerca avanzata per permettere all'utente di cercare documenti in base a titolo, autore, *range* di data di pubblicazione o parole chiave.
Il progetto prevede di implementare nelle regole di trasformazione dello *style sheet* XSL una marcatura semantica avanzata per il portale, utilizzando quando possibile gli elementi HTML che hanno uno specifico valore semantico, come `<article>`, `<aside>` e `<time>` (MDN Team, s.d.-a, "Semantics"); a questo si aggiunge la codifica di metadati strutturati tramite set di *microdata*, facendo principalmente riferimento al vocabolario di Schema.org, descrivendo entità e documenti con etichette come `Person`, `Organization` e `CreativeWork` (MDN Team, s.d.-b, "Microdata"). Ciò sarà utile per migliorare l'accessibilità del portale e permetterne l'utilizzo con strumenti assistivi, altro importante obiettivo del progetto per il quale verranno implementati anche gli attributi WAI-ARIA, come cui `role` (MDN Team, s.d.-c, "WAI-ARIA basics"). Su questi principi si è basata anche la programmazione dello *style sheet* CSS che definisce l'aspetto grafico del portale: i caratteri dei documenti sono stati uniformati con un font ad alta leggibilità di giusta grandezza, sono stati scelti colori che garantiscono un buon contrasto e sono stati evidenziati link, pulsanti e altri elementi con cui l'utente può interagire (*Ivi*, "CSS and JavaScript accessibility best practices"). È inoltre stata ricostruita nel centro di documentazione l'identità grafica dell'istituzione che conserva il fondo archivistico, modificando anche i colori e i font dell'interfaccia di Mirador 3 configurando l'istanza del *client* e inserendo uno *style sheet* CSS interno nel *template* Jinja2. Il progetto fa poi ampio uso di JavaScript per rendere più interessante l'aspetto delle pagine web, aumentare la possibilità di interazione dell'utente e aggiungere funzionalità al portale. Alcuni elementi, come quelli marcati dai *tag* `<note>` e `<stamp>` che codificano annotazioni a mano e timbri, sono stati inseriti all'interno di componenti espandibili dall'utente. Sono state programmate delle funzioni per passare da uno all'altro dei due elementi figlio degli elementi `<choice>` dei file TEI; essi servono a raccogliere insieme due proposte di codifica per la stessa porzione di testo, come abbreviazioni e forme sciolte, forme erronee e correzione ed espressioni originali e normalizzate (TEI Consortium, 2025, "Elements Available in All TEI Documents"). In maniera simile, è possibile passare da una all'altra delle varie letture di certe porzioni di testo che differiscono in versioni dello stesso documento, marcate da diversi elementi `<rdg>` all'interno dello stesso `<app>` (*Ivi*, "Critical Apparatus"). Le pagine con le trascrizioni sono poi dotate di barre laterali espandibili, tra cui una che riporta alcune delle informazioni contenute nel TEI *header* dei file TEI. Una seconda barra laterale contiene delle schede informative su persone e organizzazione menzionate nei documenti generate dinamicamente scaricando i metadati da Wikidata. Nell'elemento che marca il nome di queste entità è stato inserito il codice identificativo della relativa raccolta di metadati di Wikidata tramite l'attributo `ref`. L'identificativo viene riportato nella trascrizione trasformata, dalla quale viene estratto da una funzione JavaScript quando la pagina viene caricata e utilizzato nell'applicazione Python per costruire lo URL di *Linked Data Interface*; esso permette di scaricare i dati in formato JSON (Wikidata:Data access, 2025, "Linked Data Interface (URI)"), che infine sono utilizzati nella funzione JavaScript per generare le schede. Una terza barra laterale contiene delle opzioni per modificare le impostazioni di lettura, tra cui passare alla modalità notte o schermo intero, e mostrare i caratteri originali dei documenti. Altre opzioni permettono, per esempio, di posizionare il testo aggiunto a margine o sopra la linea nella sua posizione originale e mostrare il testo cancellato in originale o le componenti tipografiche delle pagine. Sono state create anche delle barre orizzontali espandibili poste nella parte superiore e inferiore del *viewport* che riportano intestazioni e piè di pagina originali, codificati con l'elemento `<fw>` e utilizzando rispettivamente i valori `header` e `footer` per l'attributo `type` (TEI Consortium, 2025, "Representation of Primary Sources"). Si prevede anche di implementare delle funzionalità pensate principalmente per i ricercatori. La prima consiste in un sistema per generare automaticamente citazioni bibliografiche, mentre la seconda riguarda l'annotazione di porzioni testuali in un'ottica di *social scholarship*.

3. Applicazione di sistemi AI

La terza e ultima fase si focalizza sull'implementazione di sistemi di intelligenza artificiale al centro di documentazione, puntando principalmente a utilizzare *large language model* (LLM) *open-source* in locale per ragioni economiche e di sicurezza dei dati. L'approccio principale prevede l'utilizzo delle API di piattaforme come Hugging Face e Ollama, che permettono di scaricare gli LLM sulla macchina e utilizzarne

le funzionalità all'interno del codice (Hugging Face Team, s.d., "Transformers"; Chiang & Morgan, s.d., "Quickstart"); si valuterà anche l'implementazione di LLM in maniera più complessa tramite *framework* come LangChain, LlamaIndex o Haystack. Verranno sviluppate delle funzionalità interattive per le pagine delle trascrizioni per permettere agli utenti di tradurre o sintetizzare dinamicamente porzioni testuali. Si valuterà poi la programmazione di una interfaccia per il recupero delle informazioni dai file TEI tramite meccanismi di *retrieval augmented generation* (RAG), una tecnica che permette agli LLM di arricchire l'output generato con dati recuperati da specifici set di documenti (Nagoudi et al., 2024, p. 22). Sarà di massimo interesse garantire l'estrazione della ricchezza semantica della codifica nei processi di generazione. Infine, si studierà la fattibilità di utilizzare i file TEI per il *fine-tuning*, cioè una tecnica di *transfer learning* con la quale si riaddestrano modelli preesistenti con nuovi dati (Zhang et al., 2023, "Fine-Tuning"), per permettere a un LLM di riconoscere e descrivere le componenti testuali in file di testo semplice; questo verrà fatto anche con l'obiettivo di automatizzare il processo di codifica di nuovi documenti.

BIBLIOGRAFIA

Appleby, M., Crane, T., Sanderson, R., Stroop, J. & Warner, S. (2020a, 3 giugno). Image API 3.0. IIIF | International Image Interoperability Framework. Consultato il 23 marzo 2025, da https://iiif.io/api/image/3.0/.

Appleby, M., Crane, T., Sanderson, R., Stroop, J. & Warner, S. (2020b, 3 giugno). Presentation API 3.0. IIIF | International Image Interoperability Framework. Consultato il 23 marzo 2025, da https://iiif.io/api/presentation/3.0/.

Behnel, S. (2024, 10 agosto). lxml - XML and HTML with Python. lxml. Consultato il 23 marzo 2025, da https://lxml.de/.

Chiang, M. & Morgan, J. (s.d.). Documentation. GitHub. Consultato il 23 marzo 2025, da https://github.com/ollama/ollama/blob/main/docs/README.md.

Ciula, A., Eide, Ø., Marras, C. & Sahle, P. (2023). Modelling Between Digital and Humanities, Open Book Publishers, Cambridge, https://doi.org/10.11647/OBP.0369.

Helm, L. (Ed.). (2024, 22 maggio). Mirador 3 Documentation. GitHub. Consultato il 23 marzo 2025, da https://github.com/projectmirador/mirador/wiki.

Hugging Face Team (s.d.). Documentations. Hugging Face. Consultato il 23 marzo 2025, da https://huggingface.co/docs.

IIIF Consortium (s.d.). How It Works. IIIF | International Image Interoperability Framework. Consultato il 23 marzo 2025, da https://iiif.io/get-started/how-iiif-works/.

MDN Team (s.d..-a). MDN Web Docs Glossary: Definitions of Web-related terms. MDN Web Docs. Consultato il 23 marzo 2025, da https://developer.mozilla.org/en-US/docs/Glossary.

MDN Team (s.d.-b). Web technology for developers. MDN Web Docs. Consultato il 23 marzo 2025, da https://developer.mozilla.org/en-US/docs/Web.

MDN Team (s.d.-c). Learn web development. MDN Web Docs. Consultato il 23 marzo 2025, da https://developer.mozilla.org/en-US/docs/Learn_web_development.

Meschini, F. (2020). Oltre il libro. Forme di testualità e digital humanities. Editrice Bibliografica.

Nagoudi, E. B., Alcoba Inciarte, A. & Muhammad, A. (2024, 27 maggio). Improving Archives-Focused LLMs with Retrieval Augmented Generation. In Duranti, L. & Rogers, C. (Eds), Artificial Intelligence and Documentary Heritage. SCEaR Newsletter, Special issue 2024. 21–26.

Pallets (2024, 13 novembre). Flask Documentation (3.1.x). Flask. Consultato il 23 marzo 2025, da https://flask.palletsprojects.com/en/stable/.

Sanderson R., Ciccarese, P. & Young, B. (2017, 23 febbraio). Web Annotation Data Model. W3C. Consultato il 23 marzo 2025, da https://www.w3.org/TR/annotation-model/.

Sechi, L. (2010). Editoria Digitale, Apogeo.

TEI Consortium (2025, 24 gennaio). P5: Guidelines for Electronic Text Encoding and Interchange. Text Encoding Initiative. Consultato il 23 marzo 2025, da https://tei-c.org/release/doc/tei-p5-doc/en/html/index.html.

Wikidata:Data access. (2025, 4 gennaio). In Wikidata. Consultato il 23 marzo 2025, da https://www.wikidata.org/wiki/Wikidata:Data_access.

Zhang, A., Lipton, Z. C., Li, M. & Smola, A. J. (2023). Dive into Deep Learning. Cambridge University Press. Consultato il 23 marzo 2025, da https://d2l.ai/.

Risorgimento Digitale: Un progetto di hyperedizione per i testi risorgimentali. Le *Noterelle* di Abba come caso di studio

Vincent Mobilia[1]
[1] Università degli Studi di Catania, Italia – vincent.mobilia@phd.unict.it

ABSTRACT (ITALIANO)
Il presente contributo presenta il portale *Risorgimento digitale*, sviluppato all'interno del progetto di ricerca *L'impresa dei Mille nella Memorialistica: la Sicilia come punto di partenza del processo unitario.* Il progetto si inserisce nell'ambito della letteratura risorgimentale e si focalizza sull'attività artistica di Giuseppe Cesare Abba e sulle nuove possibilità offerte dalle Digital Humanities.
L'obiettivo principale è quello di realizzare l'edizione digitale, critica e commentata, delle *Noterelle* di Giuseppe Cesare Abba, affiancata da un portale open access dedicato alla letteratura risorgimentale.
Tale edizione mira a fornire un inedito studio sull'organicità del processo correttorio, mentre il commento (anch'esso inedito) ha come obiettivo quello di rendere accessibile il testo anche ad un pubblico non specialista.
Il suddetto studio ambisce alla promozione dell'opera di Abba attraverso la pubblicazione digitale, ad accesso libero, ampliandone la diffusione presso un vasto pubblico, anche non specialista.
Il portale potrà fornire l'impulso per la successiva applicazione dell'informatica umanistica all'opera di altri autori della letteratura risorgimentale.

Parole chiave: Giuseppe Cesare Abba; Risorgimento; edizione critica digitale; codifica XML-TEI; hyperedizione.

ABSTRACT (ENGLISH)
Risorgimento Digitale: a hyperedition project on Risorgimento literary texts. Abba's Noterelle *as a case study*
This paper presents the *Risorgimento digitale* portal, developed as part of the research project *The Expedition of the Thousand in the Memoirs Literature: Sicily as the starting point of the unification process*. The project focuses on Giuseppe Cesare Abba's literary activity and the new opportunities offered by Digital Humanities.
The primary goal is to create a digital, critical, and annotated edition of Giuseppe Cesare Abba's *Noterelle*, supported by an open-access portal dedicated to Risorgimento literature. This edition aims to provide an unpublished study on the organic nature of the corrective process, while the commentary (also unedited) aims to make the text accessible to a non-specialist readers.
This study aims to promote Abba's work through open-access digital publishing, making the text widely accessible. The portal will also serve as a model for future applications of Digital Humanities to other Risorgimento literary works.

Keywords: Giuseppe Cesare Abba; Risorgimento; digital scholarly edition; XML-TEI encoding; hyperedition.

1. INTRODUZIONE
Il Risorgimento, avvenimento cardine dal punto di vista storico e politico, rappresenta per l'Italia un periodo di forte vivacità culturale contraddistinto da diversi generi letterari. All'interno di questo dinamico universo, un ruolo cruciale spetta alla memorialistica garibaldina e all'esperienza artistica di Giuseppe Cesare Abba, autore di *Da Quarto al Volturno. Noterelle d'uno dei mille* (Abba, 1983), capolavoro della letteratura risorgimentale. Nonostante l'alto valore artistico dell'opera, Abba in passato ha ottenuto un'attenzione accessoria e incidentale da parte della critica: infatti, sono pochi i saggi e i volumi che attribuiscono ad Abba un ruolo primario dal punto di vista letterario. Oltre a ciò, anche da un punto di vista digitale, non sono ancora presenti progetti dedicati alla letteratura del Risorgimento.
Il progetto si propone di colmare questa lacuna, coniugando la tradizione degli studi filologici con le metodologie innovative proprie della Filologia Digitale e, più in generale, delle Digital Humanities. Esso mira a realizzare l'edizione digitale, critica e commentata, delle *Noterelle* di Giuseppe Cesare Abba (in formato XML-TEI) e lo sviluppo di un portale open access come primo approccio per la creazione di una hyperedizione (Di Silvestro et al., C2022) su modello dell'Edizione Nazionale dell'Opera Omnia di Luigi

Pirandello[1]: l'hyperedizione offre vantaggi significativi poiché permette di integrare al dato filologico fonti documentali, storiche e di archivio, rendendo fruibile il testo non soltanto ad un pubblico specialista. Inoltre, attraverso il commento interattivo, si mira a risolvere la difficoltà di accesso ad un testo a volte distante dal lettore contemporaneo (specialmente non specialista) ma estremamente attuale e interessante. Grazie al commento interattivo sarà possibile inserire sia commenti di carattere linguistico e letterario, sia collegamenti a fonti storiche, letterarie e ad altro materiale didattico, superando il limite delle tradizionali note a piè pagina.
In ultima analisi, il progetto *Risorgimento Digitale* intende eliminare la barriera tra lettore e Risorgimento, attraverso l'applicazione delle Digital Humanities in un'ottica orientata all'accessibilità, alla fruizione e alla valorizzazione delle opere legate al nostro Risorgimento.
Seguendo questa prospettiva, l'edizione digitale delle *Noterelle* rappresenta un passo fondamentale al fine di rinnovare l'interesse verso lo scrittore cairese e il Risorgimento, promuovendo al contempo un modello di inclusione culturale.

2. DESCRIZIONE DEL PROGETTO

Il progetto prevede la realizzazione dell'edizione digitale, critica e commentata, delle *Noterelle* di Giuseppe Cesare Abba. Questa edizione integrerà un commento interattivo e un apparato critico, al fine di analizzare, secondo rigorosi criteri filologici, le varianti grafiche, lessicali, tematiche e cronologiche tra le varie edizioni del testo. L'edizione così allestita costituirà il primo approccio per l'hyperedizione dell'opera di Abba e verrà poi pubblicata all'interno di un portale digitale, destinato alla raccolta e alla valorizzazione del Risorgimento.
Purtroppo, non sono disponibili manoscritti, integrali o parziali, delle *Noterelle* e, dunque, l'analisi dei testimoni si baserà sulle prime sette edizioni dell'opera, con particolare attenzione all'edizione del 1911, l'ultima pubblicata in vita dall'autore.
Cardine del progetto è un'integrazione tra le Digital Humanities e gli studi tradizionali. Infatti, non si utilizzeranno solo strumenti e software quali codifica XML-TEI, LERA, EVT2 ma, al fine di analizzare e comprendere i mutamenti tra le varie edizioni del testo e le motivazioni alla base del processo correttorio, si farà riferimento alle letture personali dello stesso autore (Abba, 1995) e alla fitta rete di epistole scambiate con Giosuè Carducci e con Mario Pratesi (Abba, 1999).
Allo studio delle varianti, seguirà l'esegesi del testo, attraverso un approfondito e inedito commento storico-letterario, configurato come strumento di supporto alla lettura. Il commento digitale permetterà di superare i limiti delle tradizionali note a piè di pagina, offrendo al lettore (specialista e non) approfondimenti e collegamenti ad altre risorse.
Esso segnalerà i numerosi riferimenti al panorama culturale di età romana e alla letteratura italiana, con uno spazio privilegiato riservato alla *Commedia* dantesca e al Manzoni. Tale commento sarà sostenuto dalla documentazione epistolare e libraria, oltre che da opportuni riferimenti bibliografici.

3. METODOLOGIA

Il nucleo fondamentale del progetto consiste nella realizzazione dell'edizione digitale del testo di Abba, codificata secondo lo standard XML-TEI (Tomasi, 2008, pp. 103-142) tramite l'uso di Oxygen XML Editor. L'edizione sarà poi pubblicata digitalmente all'interno di un portale open access (Eve, 2023) sviluppato con WordPress. La circolazione del testo in maniera digitale, si configura come fondamentale strumento di diffusione dell'opera di Abba, rendendola accessibile e fruibile anche al lettore colto ma non specialista e aumentandone, dunque, la ricezione rispetto ad una fondamentale ma dispendiosa edizione nazionale cartacea.
Come sopracitato, non esistono manoscritti dell'opera e, in via preliminare, si è proceduto al reperimento delle prime sette edizioni a stampa delle *Noterelle* in formato PDF e alla scansione dei file in oggetto utilizzando Tesseract, un software OCR open source, scelto per l'elevata precisione nel riconoscimento dei caratteri. Successivamente, i documenti digitalizzati sono stati convertiti in file XML e saranno confrontati utilizzando LERA, strumento sviluppato dall'Institute of Computer Science Martin Luther dell'Università Halle-Wittenberg, il quale dispone di vari tools per analizzare le differenze tra diverse versioni di un'opera.
La digitalizzazione dei testi, oltre a rappresentare un aspetto essenziale per l'allestimento dell'apparato critico, costituisce un'importante azione di conservazione del patrimonio culturale (Campagnolo, 2020).

[1] https://www.pirandellonazionale.it/ (cons. 28/12/2024).

La codifica è stata progettata non soltanto per rappresentare gli aspetti strutturali del testo, ma anche per marcare elementi specifici dell'opera, indispensabili per un approccio ermeneutico. Infatti, nella fase di modellazione (Ciula et al., 2023), al fine di coniugare ermeneutica, Digital Humanities e filologia, si è scelto di marcare i seguenti aspetti:

a. Avvenimenti storici: per identificare correttamente la cronologia degli eventi e permettere collegamenti con altre fonti storiche;
b. Rappresentazione di Garibaldi: per evidenziare la costruzione dell'immaginario mitico garibaldino;
c. Rappresentazione della Sicilia: al fine di sottolineare la percezione della Sicilia da parte di un autore settentrionale;
d. Rappresentazione del nemico: marcata per analizzare la costruzione letteraria dell'immagine dell'avversario politico e militare;
e. Rappresentazione delle donne: codificata per analizzare la percezione dell'universo femminile e la sua condizione nella memorialistica garibaldina, con particolare riferimento alla Sicilia di metà Ottocento;
f. Riferimenti letterari ed eruditi: per mettere in risalto la fitta rete di collegamenti intertestuali e di citazioni che caratterizza l'opera di Abba.

Questo livello di analisi più approfondito e la codifica così strutturata risulta essere a fondamenta del c.d. "circolo ermeneutico", in accordo al quale per la comprensione di un testo è necessario esaminarlo sia nella sua complessità sia nelle sue singole parti (Grondin, 2015).
Specificatamente, in linea con l'indirizzo di ricerca, si è scelto di utilizzare il tag `<ref>` per la marcatura dei personaggi, il tag `<seg>` per la codifica delle varie rappresentazioni ed il tag `<anchor>` per le note relative al commento.
Invece, per quanto concerne l'apparato critico si è scelto di impiegare il Parallel Segmentation Method e quindi di utilizzare i tag `<app>` e `<lem>` per indicare la lezione accolta e il tag `<rdg>` con l'attributo `@wit` per le varianti (fig. 1).

```
<div type="section" n="1">
<head><placeName>Parma</placeName>,<date when="1860-05-03"><date when="1860-05-03">3 maggio 1860</date>.</date><time>Notte.</time></head>

<p> Le ciance saranno finite. Se ne intesero tante che parevano persino accuse.
    — Tutta <placeName >Sicilia</placeName> è in armi; il <placeName><placeName>Piemonte</placeName></placeName> <app><lem>non si può movere</lem>
                    <rdg wit="#N1"> non può muoversi</rdg>
                    <rdg wit="#N2">non può muoversi</rdg>
                    <rdg wit="#N3">non si può muovere</rdg></app>
; ma <persName>Garibaldi</persName>? — Trentamila insorti accerchiano <placeName>Palermo</placeName>: non aspettano che un capo, Lui! Ed egli se ne sta
chiuso in <placeName>Caprera</placeName>? No, è in <placeName>Genova</placeName>. — E allora perché non parte? — Ma <placeName>Nizza</placeName> ceduta? dicevano
```

Figura 1. Esempio di una porzione di testo codificata in XML-TEI con apparato critico

Il testo, così codificato, sarà pubblicato sfruttando le potenzialità di EVT2 (Di Pietro & Rosselli Del Turco, 2018; Schäuble & Gabler, 2018) o della versione successiva (EVT3), la quale è ancora in fase di sviluppo e disponibile in beta version.[2] La scelta di utilizzare EVT si basa sulle possibilità offerte dal software come la visualizzazione simultanea di più edizioni del testo, la rappresentazione immediata delle varianti e la possibilità di filtrare queste ultime in base alle informazioni codificate nei file XML-TEI.
Come già illustrato, per la fruizione dell'edizione digitale delle *Noterelle* è stato progettato un portale open access, il quale potrà essere implementato in futuro raggiungendo l'auspicabile punto di contatto tra le Digital Humanities e la letteratura del Risorgimento: raccogliere testi, fonti e materiale d'archivio legato ad autori risorgimentali, sviluppando in futuro, come menzionato precedentemente, una hyperedizione, che possa rendere accessibili e conosciuti autori e opere che altrimenti godrebbero di scarsa attenzione. Oltre a ciò, crediamo che si possa anche raggiungere l'obiettivo di fornire al lettore versioni filologicamente attendibili e corrette di opere letterarie (Italia, 2020, cap. 1).
Il portale, dotato di apposito dominio e server, sarà sviluppato utilizzando il content management system (CMS) WordPress. Preliminarmente, è stata progettata un'ipotetica interfaccia grafica – secondo i requisiti elencati da Rosselli Del Turco (2012) – realizzando un wireframe (Pierazzo, 2016, pp. 103-109) con il software Figma.
Tale interfaccia, nella sua bozza, è costituita da una homepage composta dall'intestazione (header) in cui è presente il logo e cinque button che rimandano alle sezioni del portale denominate "Homepage", "Progetto", "Autori", "News" e "Contatti". Nella zona centrale (body) è presente una galleria d'immagini interattiva a scorrimento continuo, la quale successivamente potrà essere arricchita in itinere e

[2] https://visualizationtechnology.wordpress.com/ (cons. 28/12/2024).

permettere, cliccando sul nome di ogni autore, di accedere alla pagina dedicata (fig. 2). All'interno della sezione "Autori", cliccando sul singolo nome si aprirà una pagina dedicata al profilo dell'autore, costituita dalle nozioni biografiche più importanti e da un elenco interattivo delle opere digitalizzate. Cliccando sul singolo titolo si aprirà la pagina dedicata all'opera selezionata.
In futuro, crediamo che questo portale potrà costituirsi come luogo (virtuale) di incontro e di lavoro in cui studiosi, autori ed esperti potranno collaborare per lo sviluppo di un rigoroso e scientifico processo creativo ed ermeneutico, indispensabile per ravvivare il patrimonio culturale e produrre conoscenza (Burdick et al., 2012, pp. 86-93).
In aggiunta a ciò, a titolo esemplificativo e non esaustivo, si prevede altresì la creazione di mappe della letteratura interattive dove localizzare episodi e battaglie presenti nei testi in esame (Karsdorp et al., 2021, pp. 33-64), anche sfruttando la tecnologia GIS (Cooper et al., 2016).



Figura 2. Homepage

In conclusione, come evidenziato dallo stesso Price (2013) "electronic editing can be daunting – financially, technically, institutionally, and theoretically – but it is also a field of expansiveness and tremendous possibility" (p. 439).

4. IMPATTO ATTESO

Nell'ambito dei risultati attesi, la ricerca presenta una duplice ricaduta scientifica e tecnologica, a beneficio della comunità studentesca e universitaria.
In primo luogo, la realizzazione dell'edizione critica e commentata intende promuovere l'autore, con conseguente riscoperta della sua opera. Oltre a ciò, quest'attività di ricerca si propone di colmare la lacuna presente in letteratura e fornire uno studio organico a chi, in futuro, vorrà accostarsi alla letteratura risorgimentale e, segnatamente, alla figura di Abba.
In secondo luogo, per ciò che concerne la prospettiva informatica e l'accessibilità culturale, con l'applicazione delle Digital Humanities supponiamo una maggiore diffusione dell'opera. Infatti, dalla pubblicazione dell'edizione digitale succitata, è ipotizzabile una maggiore fruizione del testo anche tra lettori colti ma non specialisti. Infatti, lo studioso potrà focalizzare la propria attenzione sull'apparato critico e le varianti, mentre il lettore comune potrà usufruire del commento per una maggiore comprensione del testo.
Parimenti, l'abbattimento dei costi di produzione, potrebbe aumentare una riflessione sull'esperienza artistica dell'Abba, e in senso lato sul Risorgimento, nel mondo scolastico e accademico.
Oltre a ciò, in considerazione dello sviluppo del portale in cui accogliere queste edizioni digitali, è ipotizzabile un aumento di simili ricerche e pubblicazioni. Inoltre, l'incremento della collaborazione tra ricercatori e studenti, consentirebbe anche l'attuazione di progetti che prevedano l'applicazione delle tecnologie informatiche per studi specifici sul Risorgimento, a beneficio dell'intera comunità scientifica.
Alla luce del mutevole e volatile contesto sociopolitico, riteniamo indispensabile una riscoperta di testi dedicati ad un evento cardine per l'Italia, con una riflessione che sappia coinvolgere tradizione e Digital Humanities.

5. CONCLUSIONI

L'edizione digitale delle *Noterelle* di Giuseppe Cesare Abba, pubblicata all'interno del portale *Risorgimento Digitale*, aprirà nuove prospettive sia per la fruizione di testi letterari, sia per collaborazione tra ricercatori, enti e università. Infatti, la natura open access del portale, in linea con i principi delle Digital Humanities, riflette questo spirito di cooperazione, essenziale nella prospettiva di una hyperedizione: l'apporto di diverse figure sarà indispensabile non soltanto ai fini dell'edizione stessa, ma anche per stimolare nuove letture critiche e incoraggiare anche future sperimentazioni didattiche sul Risorgimento.
In conclusione, la digitalizzazione delle *Noterelle* diverrà un prototipo per le successive edizioni digitali di testi risorgimentali, dove far collimare tradizione e innovazione, rinnovando l'interesse per la letteratura del periodo risorgimentale e, in ultimo, configurandosi come attività di promozione del nostro patrimonio letterario e identitario.

RINGRAZIAMENTI

Si ringrazia la biblioteca digitale dell'Università di Toronto, la Biblioteca dell'Archiginnasio di Bologna e la Biblioteca Multimediale Arturo Loria di Carpi per la concessione delle scansioni PDF delle diverse edizioni delle *Noterelle*.

BIBLIOGRAFIA

Abba, G. C. (1983). Da Quarto al Volturno. Noterelle d'uno dei Mille. In L. Cattanei, E. Elli, & C. Scarpati (A cura di), *Scritti garibaldini* (Edizione Nazionale delle opere di Giuseppe Cesare Abba, Vol. 1, pp. 301–457). Brescia: Morcelliana.

Abba, G. C. (1995). *Scritti vari. Critica letteraria e storica: note e appunti* (Edizione Nazionale delle Opere di Giuseppe Cesare Abba, Vol. 5.1, a cura di M. Candiani). Brescia: Morcelliana.

Abba, G. C. (1999). *Epistolario* (Edizione Nazionale delle Opere di Giuseppe Cesare Abba, Vol. 8, a cura di L. Cattanei & E. Costa). Brescia: Morcelliana.

Burdick, A., Drucker, J., Lunenfeld, P., Presner, T., & Schnapp, J. (2012). *Digital_Humanities*. Cambridge (MA) - London: The MIT Press.

Campagnolo, A. (2020). Conservation and Digitization: A Difficult Balance? In *Book conservation and digitization. The challenges of dialogue and collaboration* (pp. 49-87). Leeds: Arc Humanities Press.

Ciula, A., Eide, Ø., Marras, C., & Sahle, P. (2023). *Modelling Between Digital and Humanities: Thinking in Practice* (1ª ed.). Open Book Publishers. https://doi.org/10.11647/OBP.0369.

Cooper, D., Donaldson, C., & Murrieta-Flores, P. (Eds.). (2016). *Literary Mapping in the Digital Age*. London-New York: Routledge.

Di Pietro, C., & Rosselli Del Turco, R. (2018). Between Innovation and Conservation: The Narrow Path of User Interface Design for Digital Scholarly Editions. In R. Bleier, M. Bürgermeister, H. W. Klug, F. Neuber, & G. Schneider (Eds.), *Digital Scholarly Editions as Interfaces* (Schriften des Instituts für Dokumentologie und Editorik ed., Vol. 12, pp. 133-163). Norderstedt: BoD.

Di Silvestro, A., D'Agata, C., & Sichera, A. (2022). Edizione critica, edizione digitale, hyperedizione. "Il fu Mattia Pascal" come paradigma dell'edizione digitale dell'Opera Omnia di Luigi Pirandello. *Bollettino. Centro di studi filologici e linguistici siciliani, 33*, 263–280.

Eve, M. P. (2023). Open Access in the Humanities Disciplines. In J. O'Sullivan (Ed.), *The Bloomsbury Handbook to the Digital Humanities* (pp. 223-231). London: Bloomsbury Academic.

Grondin, J. (2015). What is the hermeneutical circle? In N. Keane & C. Lawn (Eds.), T*he Blackwell companion to hermeneutics* (pp. 299–305). Chichester: Wiley-Blackwell. https://doi.org/10.1002/9781118529812.ch34.

Italia, P. (2020). *Editing Duemila*. Roma: Salerno.

Karsdorp, F., Kestemont, M., & Riddel, A. (2021). *Humanities data analysis. Case Studies with Python*. Princeton, NJ-Oxford: Princeton University Press.

Pierazzo, E. (2016)*. Digital Scholarly Editing. Theories, Models and Methods*. London-New York: Routledge.

Price, K. M. (2013). Electronic Scholarly Editions. In R. Siemens, & S. Schreibman (Eds.), *A Companion to Digital Literary Studies* (pp. 434-450). Malden: Blackwell Publishing.

Rosselli Del Turco, R. (2012). After the editing is done: Designing a Graphic User Interface for digital editions. *Digital Medievalist*, *7*. https://doi.org/10.16995/dm.30.

Schäuble, J., & Gabler, H. W. (2018). Encodings and Visualisations of Text Processes across Document Borders. In R. Bleier, M. Bürgermeister, H. W. Klug, F. Neuber, & G. Schneider (Eds.), *Digital Scholarly Editions as Interfaces* (Schriften des Instituts für Dokumentologie und Editorik ed., Vol. 12, pp. 165-191). Norderstedt: BoD.

Tomasi, F. (2008). *Metodologie informatiche e discipline umanistiche*. Carocci: Roma.

SpaceLat
La geografia della letteratura latina tardoantica

Riccardo Consolini
Università del Piemonte Orientale, Italia, consoliniriccardo@gmail.com

ABSTRACT (ITALIANO)
In questo breve articolo vorrei discutere gli obiettivi e i metodi del progetto SpaceLat, futuro archivio digitale di testi della letteratura latina tardoantica, e mostrarne le potenziali applicazioni.
Parole chiave: letteratura latina; edizioni digitali; tarda antichità; cartografia.

ABSTRACT (ENGLISH)
SpaceLat: the geography of late antiquity Latin literature. In this brief article I would like to discuss the goals and methods of the SpaceLat project, future digital archive of late antiquity Latin literature texts, and show its potential applications.
Keywords: Latin literature; digital editions; late antiquity; cartography.

1. INTRODUZIONE
Di qualsiasi ambito disciplinare si tratti, l'esistenza di spazi virtuali che mettano a disposizione testi e fonti in formato digitale, specialmente se di libero accesso, è un fattore di grande utilità per la diffusione e la democratizzazione del sapere che favorisce la ricerca, l'istruzione e la divulgazione. In particolare, nell'ambito della letteratura latina, da anni la biblioteca digitale DigilibLT dell'Università del Piemonte Orientale fornisce un archivio liberamente consultabile e in continuo ampliamento che raccoglie ormai più di 400 opere composte in periodo tardoantico e relative a molteplici generi letterari.[1] Proprio grazie all'esperienza maturata nella gestione di questo ambiente digitale, nel 2023 nasce il progetto SpaceLat, coordinato dalla professoressa Alice Borgna, che ha come obiettivo coadiuvare da una parte le istanze delle Digital Humanities e dall'altra la ricerca archeologica e geografica.[2] Questa esposizione vuole mostrare i primi risultati del progetto (ancora in fase di prototipo) e discuterne i fini e le potenzialità.

2. SCOPI E METODI
Il principale obiettivo del progetto resta quello di mettere a disposizione del pubblico un'ampia selezione di fonti primarie della letteratura latina tardoantica in formato digitale; la novità sta invece nell'accompagnare e integrare i testi con l'ausilio di mappe interattive.
La prima fase del lavoro riguarda la digitalizzazione delle opere scelte per il corpus a partire dalle edizioni critiche di riferimento. Questa operazione consiste nella marcatura XML dei file di testo secondo lo standard internazionale TEI che garantisce un'elevata riproducibilità, diffusione e accessibilità delle informazioni testuali e metatestuali delle opere.[3] Completata la marcatura ordinaria si procede poi con una marcatura specifica che codifica all'interno del file le varie informazioni di carattere geografico e archeologico fornite dal testo. Tramite l'ausilio di un software queste informazioni sono successivamente raffigurate sotto forma di mappe interattive basate sullo standard WGS 84[4] e riproducono il contenuto dell'intero testo o di una sua porzione, a seconda delle scelte dell'editore. Il programma è versatile è può generare diversi tipi di visualizzazione.

·Ecco un esempio un "blocchetto geografico" singolo:

```
<listPlace>
        <place xml:id="place1" rend="full">
                <placeName xml:id="placeName1" type="city">
                        <name></name>
```

[1] https://digiliblt.uniupo.it (cons. 10/01/2025).
[2] https://research.uniupo.it/it/projects/spaces-and-paths-in-late-antiquity-a-digital-connection-for-liter (cons. 10/01/2025).
[3] TEI: Guidelines for Electronic Text Encoding and interchange, https://tei-c.org/release/doc/tei-p5-doc/en/html/index.html (cons. 10/01/2025).
[4] https://earth-info.nga.mil/?dir=wgs84&action=wgs84 (cons. 10/01/2025).

```
                <region type="XX" n="PMNT"></region>
                <country key="XX"></country>
                <bloc type="XX"></bloc>
            </placeName>
            <location type="city" n="10">
                <geo decls="#WGS">0.00000 0.00000</geo>
                <desc></desc>
                <label></label>
            </location>
        </place>
</listPlace>
```

Una codifica di questo genere identifica un luogo specifico e le sue relative informazioni geografiche, rappresentandolo sulla mappa come un unico punto (con icone, colori e dimensioni variabili).

·Ecco invece un esempio di percorso:

```
<listPlace>
        <place xml:id="test_multiple_location_line" rend="fullPath">
            <district type="XX"></district>
            <region type="XX" n="PMNT"></region>
            <country key="XX"></country>
            <bloc type="XX"></bloc>
            <location type="mountain">
                <geo decls="#WGS">0.00000 0.00000</geo>
                <desc></desc>
                <label></label>
            </location>
            <location type="city">
                <geo decls="#WGS">0.00000 0.00000</geo>
                <desc></desc>
                <label></label>
            </location>
            <location type="village">
                <geo decls="#WGS">0.00000 0.00000</geo>
                <desc></desc>
                <label></label>
            </location>
        </place>
</listPlace>
```

Una codifica di questo genere indentifica un itinerario tra diverse località e le loro relative informazioni geografiche, rappresentando contemporaneamente sulla mappa sia luoghi sia il percorso che li unisce (con icone, colori e dimensioni variabili).

L'augurio è che l'archivio digitale di SpaceLat possa rivelarsi un ausilio utile e fruttuoso sia nel campo della didattica che della ricerca. L'identificazione delle entità geografiche del mondo antico non è sempre agevole; a volte è dubbia o addirittura impossibile. L'archivio SpaceLat potrebbe diventare un punto di riferimento in cui trovare informazioni bibliografiche e rimandi a contributi specialistici di ambito archeologico e geografico.

3. PROTOTIPI

I primi testi di cui ci siamo occupati appartengono al corpus degli *Itineraria Hierosolymitana*,[5] racconti di pellegrinaggio verso i luoghi di culto della Terra Santa prodotti in periodo tardoantico e medievale, che, proprio per l'argomento trattato, si prestano molto bene ad una rappresentazione geografica. Tenendo

[5] Sull'argomento vedi Maraval 1985, 1995 e 1996; Natalucci 1991; Wilkinson 1981 e 2002; Aist 2018.

conto che non è possibile mostrare in modo accurato quelle che sono a tutti gli effetti delle cartine interattive (in cui è quindi possibile spostarsi, aumentare la risoluzione etc.) riporto qui di seguito come esempi alcune porzioni di testo insieme alle immagini delle loro relative rese grafiche (chiaramente non definitive).

Breviarius de Hierosolyma[6]

(forma a)

2. Et inde intrans in **Golgotha** est ibi atrium grande ubi crucifixus est Dominus...
6. Quomodo discendis ad **Silua**, ibi est ille lacus, ubi missus est sanctus Hieremias...

(forma b)

7. ... Ad dextera parte ibi est **uallis Iosaphat**, ibi iudicaturus est Dominus iustos et peccatores...

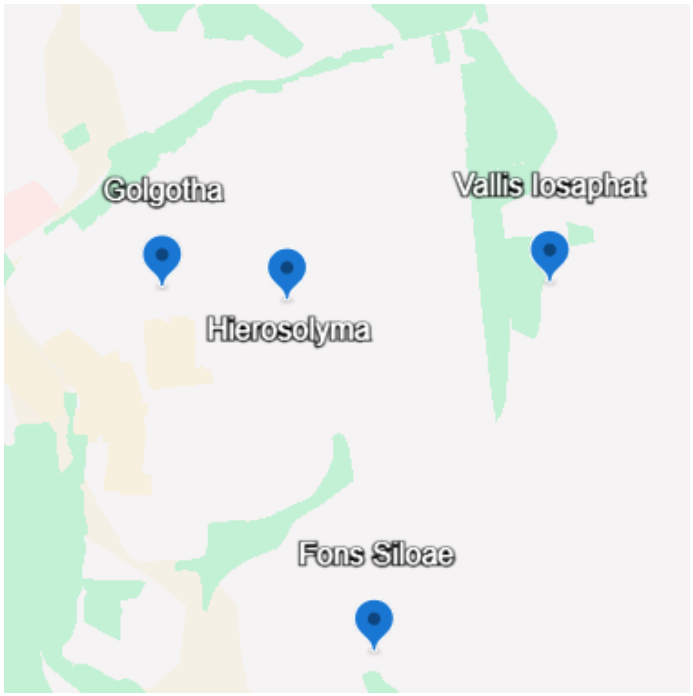


Fig 1. I luoghi del *Breviarius de Hierosolyma*

Eucherii De situ Hierusolimae epistula ad faustum presbyterum[7]

11. Sex milibus **Bethlehem** ab **Hierusolima** in meridiano latere secedit...
12. **Hiericho** uero ab Hierusolima in orientem aestiuum decem et octo milibus excurrit...
13. **Chebron** ciuitas quondam gigantum lapide ferme uicesimo et secundo ab Hierusalem infra meridianam plagam distat...

[6] Per le edizioni critiche del testo vedi Tobler 1879, 55-59; Gildemeister 1882, 33-35; Geyer 1898, 151-155; Willmart 1928, 101-106; Weber 1965, 105-112 (l'edizione di riferimento, a cui risale la divisione del testo in due forme distinte). Per un compendio sulla tradizione vedi Pellegrini 2012a. Per le traduzioni vedi Stewart 1890, 13-16; Heisenberg 1908, 111-122; Maraval 1996, 179-184; Wilkinson 2002, 117-121; Whalen 2011, 40-41.
[7] Per le edizioni critiche vedi Tobler 1879, 49-54; Geyer 1989, 125-134; Fraipont 1965, 235-241 (l'edizione di riferimento). Per le traduzioni vedi Stewart 1890, 7-12; Heisenberg 1908, 129-137; Maraval 1996, 169-178; Wilkinson 2002, 94-98.



Fig 2. Alcuni luoghi del *De situ Hierusolimae*

Thedosii De situ Terrae Sanctae[8]

3. De porta Purgu usque ubi pugnauit Dauid cum Golia in monte Buzana, quod interpretatur lucerna, milia XV. De Buzana usque **Eleuteropoli** milia XV. De Eleuteropoli usque in loco, ubi requiescit sanctus Zacharias, milia VI, et de ipso loco usque ad **Ascalona** milia XX. De Ascalona usque ad **Gaza** milia XII. Inter Ascalonam et Gazam ciuitates duas, id est **Antedona** et **Maioma**. De Gaza usque ad **Rafia** milia XXIIII. De Rafia usque ad **Betuliam**, ubi Olofernis mortuus est, milia XII.

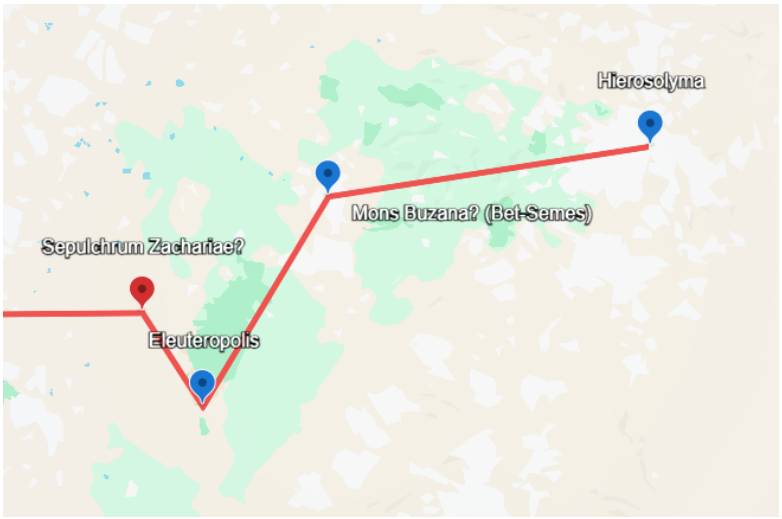


Fig 3a. Il par. 3 del *De situ Terrae Sanctae*
(In rosso sono indicati i luoghi non identificati)

[8] Per le edizioni critiche vedi Tobler 1879, 61-88; Gildemeister 1882, 15-30; Pitra 1888, 118-121; Geyer 1898, 135-150 (l'edizione di riferimento). Per un compendio sulla storia delle edizioni vedi Pellegrini 2012b. Per le traduzioni vedi Bernard 1893, 7-19; Heisenberg 1908, 106-110; Maraval 1996, 185-202; Wilkinson 2002, 103-116.

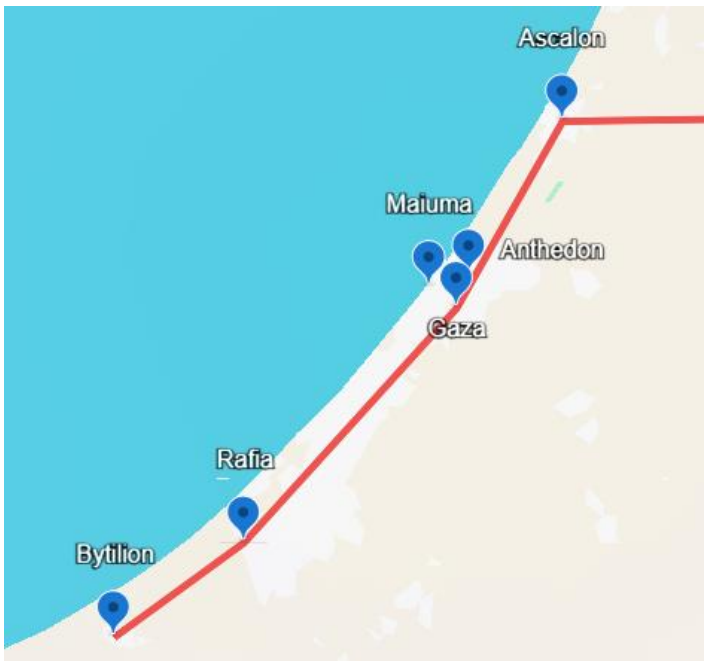


Fig 3b. Il par. 3 del *De situ Terrae Sanctae*

32. In prouincia Cilicia **Aegea** dicitur ciuitas, ubi XL dies commercia geruntur et nemo de eis aliquid requirit; si post XL dies inuentus fuerit negotium gerere, fiscalia reddit. In prouincia Cilicia ciuitas **Tharso**, inde Apollonius fuit. De Tharso usque **Adana** ciuitate milia XXX. De Adana usque ad **Masista** XXX milia. De Masista usque ad **Anasta** <...> usque **Aegeas** LX milia. De Aegeas usque ad **Alexandria Scabiosa** LX milia. Ab Alexandria Scabiosa usque **Antiochia** LX milia. De Antiochia usque in **Quiro**, ubi sunt sanctus Cosmas et Damianus, qui ibi et percussi sunt, milia LX. De Quiro usque **Barbarisso**, ubi sunt percussi sanctus Sergius et Bacco, milia LX. De Barbarisso usque ad **Eneapoli** <...> in **Calonico** milia LXXX. De Calonico usque in **Constantina** milia LX. De Constantina in **Edessa** LXXX milia, ubi Abgarus rex, qui domino Christo scripsit, manebat. De Edessa usque in **Dara** CXX milia. De Dara in **Amida** LXXX milia, quae est ad fines Persarum. De Amida usque **Ramusa** milia XVIII.



Fig 4a. Il par. 32 del *De situ Terrae Sanctae*
(In viola sono indicati i luoghi di incerta identificazione)

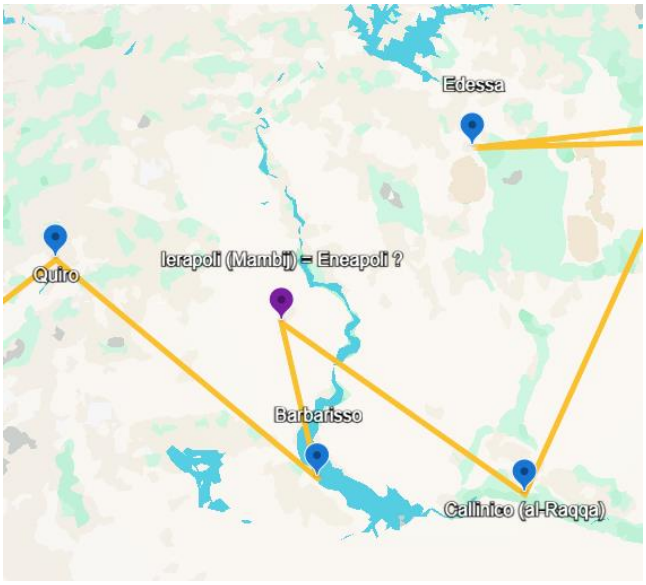


Fig 4b. Il par. 32 del *De situ Terrae Sanctae*
(In viola sono indicati i luoghi di incerta identificazione)

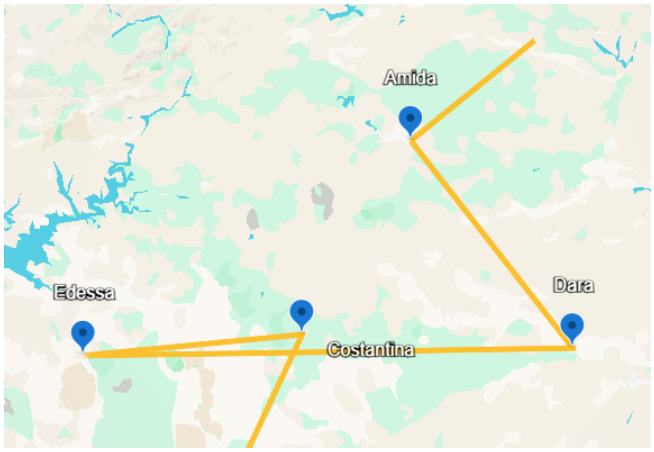


Fig 4c. Il par. 32 del *De situ Terrae Sanctae*

RINGRAZIAMENTI

Ringrazio sentitamente la prof.ssa Alice Borgna, la prof.ssa Eleonora Destefanis, il professor Maurizio Lana e la prof.ssa Raffaella Afferni per avermi dato la possibilità di partecipare al progetto SpaceLat; ringrazio inoltre il dottor Gianmario Cattaneo, la prof.ssa Nadia Rosso e il dottor Massimo Ghisalberti per i preziosi consigli.

BIBLIOGRAFIA

AIB, *Manifesto per l'information literacy*, https://www.aib.it/documenti/ilmanifesto
IFLA, *Media et Information Literacy recommendations*, https://www.ifla.org/publications/ifla-media-and-information-literacy-recommendations
Aist 2018 = R. Aist, *From Topography to Text. The Image of Jerusalem in the Writings of Eucherius, Adomnàn and Bede*, Turnhout 2018.
Bernard 1893 = J. H. Bernard, *Theodosius - On the topography of the Holy Land*, in *Palestine Pilgrims Text Society*, *Theodosius. (A. D. 530)*, vol. 2^2, London 1893.

Fraipont 1965 = I. Fraipont, *Eucherii <quae fertur> De situ Terrae Sanctae epistula ad Faustum presbyterum*, in CCSL, vol. 175, Turnhout 1965, 235-241.
Geyer 1898 = P. Geyer, *Itineraria Hierosolymitana saeculi IIII-VIII,* CSEL vol. 39, Wien 1898.
Gildemeister 1882 = J. Gildemeister, *Theodosius de situ Terrae Sanctae im ächten text und der Breviarius de Hierosolyma vervollstandigt*, Bonn 1882.
Heisenberg 1908 = A. Heisenberg, *Grabeskirche und Apostelkirche. Zwei basiliken Konstantins: Untersuchungen zur Kunst und Literatur des ausgehenden Altertums*, vol. I, Leipzig 1908, pp. 111-122.
Maraval 1985 = P. Maraval, *Lieux saints et pèlerinages d'Orient. Histoire et géographie. Des origines à la conquête arabe,* Paris 1985.
Maraval 1995 = P. Maraval, *Les itinéraires de pèlerinage en Orient (entre le 4e et le 7e s.)*, in *Akten des XII. Internationalen Kongresses für christliche Archäologie*, 1995 pp. 291-300.
Maraval 1996 = P. Maraval, *Récits des premiers pèlerins chrétiens au Proche-Orient (IVe- VIIe siècle),* Paris 1996.
Molinier 1882 = A. Molinier, *Theodosius. De situ Terrae Sanctae*, in «Revue Critique» 13 (1882) pp. 328-34.
Natalucci 1991 = N. Natalucci (a cura di), *Egeria. Pellegrinaggio in Terra Santa*, Firenze 1991.
Pellegrini 2012a = M. Pellegrini, *Breviarius de Hierosolyma* in *La Trasmissione dei testi latini del Medioevo - Mediaeval Latin Texts and their Transmission* Te.Tra. 4, Firenze 2012, pp. 76-80.
Pellegrini 2012b = M. Pellegrini, *Theodosius archidiaconus* in *La Trasmissione dei testi latini del Medioevo - Mediaeval Latin Texts and their Transmission* Te.Tra. 4, Firenze 2012, pp. 507-521.
Pitra = J. Pitra, *Vergilii de situ Terrae Sanctae*, in *Analecta sacra et classica*, V, Paris 1888 pp. 118-121.
Stewart 1890 = A. Stewart, *The Epitome of S. Eucherius about certain holy places (ca. A.D. 440) and the Breviary or short description of Jerusalem (ca. A.D. 530)*, in *The Library of the Palestine Pilgrims' Text Society*, vol. 2^1, London 1890.
Tobler-Molinier 1879 = T. Tobler, A Molinier, *Itinera Hierosolymitana et descriptiones Terrae Sanctae bellis sacris anteriora*, Genève, 1879.
Tsafrir 1986 = Y. Tsafrir, *The Maps used by Theodosius. On the Pilgrim Maps of the Holy Land and Jerusalem in the Sixth Century C. E.*, in «Dumbarton Oaks Papers» 40 (1986), pp. 129-145.
Weber 1965 = R. Weber, *Breviarius de Hierosolyma,* in CCSL vol. 175, Turnhout 1965, pp. 105-112.
Whalen 2011 = B. E. Whalen, *Pilgrimage in the Middle Ages: A Reader*, Toronto 2011, pp. 40-41.
Wilkinson 1981 = J. Wilkinson, *Egeria's Travels to the Holy Land*, London 1981.
Wilkinson 2002 = J. Wilkinson, *Jerusalem Pilgrims Before the Crusades*, Oxford 2002.
Willmart 1928 = A. Willmart, *Un nouveau témoin du 'Breviarius de Hierosolyma'*, in «Revue Biblique», XXVII, Paris 1928, pp. 101-106.

Verso l'edizione digitale del carteggio Canneti-Fiacchi

Chiara Manca[1], Fiammetta Sabba[2], Bianca Sorbara[3], Silvia Tripodi[4]
[1] Alma Mater Studiorum - Università di Bologna, chiara.manca5@unibo.it
[2] Alma Mater Studiorum - Università di Bologna, fiammetta.sabba@unibo.it
[3] Alma Mater Studiorum - Università di Bologna, bianca.sorbara2@unibo.it
[4] Università di Napoli Federico II, silvia.tripodi@unina.it

ABSTRACT (ITALIANO)

Il contributo intende presentare i primi risultati relativi alla realizzazione dell'edizione scientifica digitale del carteggio intercorso negli anni 1711-1730 tra Pietro Canneti e Mariangelo Fiacchi, oggi conservato presso la Biblioteca Classense di Ravenna. All'interno del contributo saranno illustrati gli obiettivi e le metodologie adottate dal gruppo di lavoro, selezionate in modo da rendere l'edizione digitale un valido strumento per studi interdisciplinari e accessibile per ogni tipologia di utenza. In seguito alla presentazione del carteggio, del suo valore storico e delle tappe costitutive indispensabili per la creazione dell'edizione digitale (digitalizzazione, metadatazione, trascrizione), verranno presentate le scelte relative al lavoro di individuazione degli elementi culturali caratterizzanti emersi dalle lettere, con particolare focus su quelli di interesse specificatamente bibliografico e storico-bibliotecario. Successivamente saranno fornite le specifiche tecniche e le scelte relative all'elaborazione digitale del testo, quali modellizzazione, marcatura, allineamento dei metadati descrittivi e progettazione della pubblicazione di esso.
Parole chiave: edizione scientifica digitale; TEI-XML; carteggio; Istituzione Biblioteca Classense; XVIII secolo.

ABSTRACT (ENGLISH)

Towards the digital edition of the Canneti-Fiacchi correspondence

In this work, we present the preliminary results on the implementation of the Digital Scholarly Edition of the correspondence between Pietro Canneti and Mariangelo Fiacchi, which took place from 1711 to 1730. This correspondence is currently preserved in the Classense Library in Ravenna. Throughout this work, we discuss the main objectives and methods adopted by the team selected to make the Digital Scholarly Edition a viable tool for cutting-edge interdisciplinary studies in an accessible way to a broad audience. Firstly, the historical relevance of the correspondence is outlined, followed by a discussion of the essential steps needed for the creation of the digital edition, i.e. the digitalization, as well as the metadata and transcript generation. Subsequently, the selection of the key cultural elements emerging from the correspondence is analysed with a particular emphasis on those pertaining to the realm of bibliography and historical librarian studies. In addition, an extensive overview on the set of techniques employed to process the text digitally, namely modelling, marking-up, the alignment of the descriptive metadata and the relative publication plans.
Keywords: scholarly digital edition; TEI-XML; correspondence; Classense Library; 18th century.

1. INTRODUZIONE

Attraverso la presentazione dell'edizione scientifica digitale del carteggio Canneti-Fiacchi, attualmente in corso di realizzazione presso il Laboratorio LUDI del Dipartimento di Beni Culturali dell'Università di Bologna,[1] il paper intende sottolineare l'importanza della creazione di un'infrastruttura digitale che sia allo stesso tempo corretta dal punto di vista metodologico e filologico, e accessibile per ogni tipologia di utenza. Il contributo presenta i primi risultati del progetto *DigiLet Class* (*Digitizing Letters of Classense Library*), avente come obiettivi sia la valorizzazione della storia della Biblioteca Classense di Ravenna, delle lettere manoscritte e del relativo contenuto attraverso le trascrizioni interattive della corrispondenza fra l'abate Pietro Canneti (1659-1730) e il padre bibliotecario Mariangelo Fiacchi (1688-1777); sia la creazione di uno strumento valido per lo studio storico del mercato librario tra XVII e XVIII secolo, in particolare dell'area adriatica e dello Stato Pontificio, attraverso la navigazione degli indici di personalità, luoghi e opere citate all'interno delle lettere.

Si tratta di un carteggio avente un importante valore documentale e documentario, dal momento che è testimone unico della costituzione della prima raccolta libraria della Biblioteca Classense, originaria dell'Abbazia di Classe, e presenta elementi di notevole rilevanza per la storia del libro, della stampa e del contesto culturale del XVIII secolo.

[1] LUDI - Laboratorio Universitario di Documentazione e Informazione.

Il carteggio Canneti-Fiacchi, intercorso negli anni 1711-1730, è costituito da un totale di 599 lettere manoscritte, di cui 438 lettere scritte da Canneti a Fiacchi e 161 lettere scritte da Fiacchi a Canneti, tutte conservate presso la Biblioteca Classense di Ravenna.[2]
Le tematiche presenti nelle lettere sono principalmente legate alla ricerca, all'acquisto e allo scambio di edizioni a stampa ed esemplari manoscritti che Pietro Canneti desiderava per la *costruenda* biblioteca dell'Abbazia, ma emergono anche vicende personali e interpersonali dei due monaci, appartenenti all'Ordine Camaldolese.
La realizzazione dell'edizione digitale del carteggio si configura quindi come un'importante opportunità di valorizzazione di tali contenuti, poiché i carteggi, in qualità di fonti documentarie, sono rivelatori di contesti culturali di tipo nazionale o locale e permettono studi di tipo comparativo e interdisciplinare (Sabba 2016).
La prima fase di *DigiLet Class* ha previsto la riproduzione digitale e la relativa metadatazione delle missive da parte dell'Istituzione Biblioteca Classense, con la collaborazione di tirocinanti del Corso di Laurea Magistrale in Scienze del libro e del documento. A questa è seguita una riflessione metodologica del gruppo di ricerca sulle scelte da applicare alla seconda fase del progetto, costituita dalla trascrizione interpretativa del testo delle lettere e dall'individuazione delle tipologie di entità considerate significative.
La terza fase tecnico-informatica, al momento in corso di svolgimento, riguarda la codifica semantica del carteggio, ovvero la base dell'edizione digitale destinata a essere successivamente pubblicata su una piattaforma web.
Le scelte tecniche alla base dell'edizione digitale inseriscono pienamente il progetto all'interno dell'attuale panorama nazionale e internazionale di iniziative di valorizzazione di carteggi, epistolari e raccolte. In questa prospettiva, si è scelto di utilizzare lo schema TEI (Text Encoding Initiative), considerato lo standard per eccellenza per la codifica e la descrizione di testi digitali e anche valido strumento per la creazione di edizioni digitali di alta qualità (TEI Consortium 2023).

2. METODOLOGIA

2.1 Dalla trascrizione delle lettere all'identificazione delle entità

Se la trascrizione è intesa come tappa indispensabile di ogni edizione, essa si configura allo stesso tempo come gesto culturale, atto storico e di pensiero (Jodogne 2018). Numerose sollecitazioni sui problemi relativi ai criteri di trascrizione sono state in parte stimolate anche dall'incremento dello sviluppo delle edizioni digitali, in genere le regole sono dettate dalla natura stessa del carteggio al quale si adattano le norme del sistema prescelto (Vetrugno 2018).
All'interno della fitta rete di scambi eruditi e letterari che interessò Pietro Canneti, il carteggio con le sue centinaia di lettere permette di far luce sui numerosi contatti che l'abate intrattenne negli anni e sulle modalità di ampliamento della raccolta libraria classense e della sua biblioteca personale.
La scelta di affiancare l'immagine digitale della lettera originale alla trascrizione consente di restituire visivamente le convenzioni degli scriventi anche in termini di resa grafica. Ogni lettera è stata trascritta fedelmente senza eccezioni mantenendo abitudini e usi scrittori, anche se non più correnti, e la suddivisione in paragrafi; si è intervenuto sullo scioglimento delle abbreviazioni senza adottare l'uso di parentesi basandosi sulle forme piene corrispondenti nei casi di parole abbreviate per troncamento, contrazione o per i segni abbreviativi di significato proprio. È stato mantenuto l'uso moderno delle lettere maiuscole per quanto riguarda i nomi propri, le lettere dopo il punto fermo e per i titoli di dignità poiché negli originali risulta convenzionale il frequente ricorso a tali formule. Si ricorre all'uso dei parentesi quadre per indicare le lacune materiali determinate da perdita del supporto o del testo causata da macchie di varia natura o abrasioni, le stesse parentesi sono state utilizzate per racchiudere il testo eventualmente integrato; inoltre, le parti testuali presenti sui margini sono state riportate in calce tra parentesi quadre rispettando l'ordine di successione.
Pur comprendendo diverse centinaia di missive e responsive, il carteggio Canneti-Fiacchi non rappresenta una serie completa; nonostante alcune lacune e la conseguente difficoltà di comprendere pienamente il contesto e la realtà di riferimento, il *corpus* annovera eruditi e personalità molto note nel mondo intellettuale e apre a numerosi scambi informativi e letterari che rivelano, inoltre, nomi di professionisti e intermediari coinvolti nei processi di ampliamento delle raccolte, facendo emergere allo stesso tempo non

[2] Nel dettaglio le lettere sono suddivise come segue: Biblioteca Classense Ravenna, Carteggio P. Canneti - M. Fiacchi, Lettere, Buste, 10-12 [Busta 10 (anni 1711-1720) - 133 lettere; Busta 11 (anni 1721-1724) - 141 lettere; Busta 12 (anni 1725-1730) - 140 lettere]; Miscellanea Mob. 3. 5 I^2 (Mob. 3. 5 I^2, n. 4 Fiacchi a Canneti, 1724, 1 lettera; Mob. 3. 5 I^2, nn. 5-14, Canneti a Fiacchi, 1725-1725, 12 lettere; Mob. 3. 5 I^2, nn.16-19, Canneti a Fiacchi, 1725-1726, 4 lettere; Mob. 3. 5 I^2, nn.21-27, Canneti a Fiacchi, 1729- 1730, 7 lettere). Per quanto riguarda invece le lettere spedite da Fiacchi a Canneti si rimanda a Biblioteca Classense Ravenna, Carteggio M. Fiacchi-P. Canneti, Lettere, Busta, 24, fasc. 2.

solo le modalità di acquisizione e commercio librario, ma anche le diverse località interessate dai traffici commerciali.
Alla trascrizione è seguita una fase di analisi testuale al fine di identificare e distinguere le varie tipologie di entità presenti nelle lettere per permettere di evidenziare le numerose informazioni e trasformarle in dati strutturati. Per esplicitare e agevolare la comprensione della fitta rete di conoscenze, si è scelto di identificare e distinguere le entità in quattro macrocategorie: personaggi, luoghi, organizzazioni ed entità bibliografiche.
In una fase intermedia tra la trascrizione delle lettere e la codifica dei testi tramite lo standard TEI-XML sono stati predisposti diversi fogli Excel dedicati a ogni tipologia di entità citate, identificando ognuna di queste con un codice univoco nella seguente forma: "DLCL_CF_X0000", in cui "X" è una lettera diversa in base alla tipologia di entità e le cifre seguenti rappresentano il numero progressivo della singola entità. Al fine di favorire collegamenti tra l'edizione digitale e altre basi di conoscenza consolidate, i codici univoci del progetto di edizione digitale si affiancano ad ulteriori identificativi specifici per ogni tipologia di entità, descritte, quando possibile, tramite informazioni strutturate nelle varie colonne del foglio Excel. Per i personaggi citati si è riportato l'identificativo VIAF (The Virtual International Authority File)[3] e altre informazioni come nome, cognome, altro nome o attributo, data e luogo di nascita, data e luogo di morte, occupazione principale, genere e affiliazione; mentre per i luoghi è stato utilizzato il codice GeoNames,[4] che ha lo scopo di referenziare ogni luogo con le relative coordinate geografiche. Anche per le organizzazioni è stato utilizzato il codice VIAF, accompagnato dalla relativa denominazione e dalle informazioni sulla posizione geografica. Qualora sia stato possibile procedere all'identificazione delle edizioni a stampa, oltre alle informazioni specifiche di edizione (autore, titolo, data e luogo di stampa, tipografo) che permetteranno nuovi riscontri sugli esemplari classensi, è stato associato il codice SBN,[5] mentre per i manoscritti la segnatura è stata scelta come elemento distintivo, accompagnata da altri dati come autore, titolo, data e luogo di copia, provenienza, copista.

2.2 Il modello di codifica TEI per il carteggio Canneti-Fiacchi

Come evidenziato da Hankins (Hankins 2013), la riproduzione digitale di lettere consente sia di rappresentare le loro speciali caratteristiche grafiche e testuali sia di chiarire i contesti sociali in cui sono state prodotte, permettendone, inoltre, una fruizione che, per modalità e per granularità, risulta difficilmente replicabile nelle edizioni convenzionali.
Lo sviluppo dell'edizione digitale del carteggio è iniziato con la codifica TEI-XML delle trascrizioni delle lettere della Busta 10, scritte da Pietro Canneti a Mariangelo Fiacchi, ovvero la base di un'edizione digitale vera e propria. La metodologia di seguito esposta si è consolidata durante questa prima fase del progetto e vuole porsi come modello per la marcatura del corpo restante delle lettere. In seguito a un attento studio delle linee guida di TEI (TEI Consortium 2023), sono stati selezionati gli elementi più idonei a rappresentare le caratteristiche strutturali e semantiche del *corpus*. Questa selezione è stata esplicitata all'interno di uno schema ODD (One Document Does-it-all) generato mediante l'editor Roma di TEI.[6] Il progetto TEI generato da questo schema (utilizzando il software Oxygen XML Editor)[7] è stato versionato e pubblicato in una repository del servizio di hosting di software Github.[8]
Dal momento che il carteggio è costituito da tre buste e una miscellanea, si è scelto di utilizzare un formato di codifica che permettesse di risalire alla segnatura specifica. Per ogni lettera della Busta 10 si è quindi creato un file XML rinominato da un codice univoco nella forma "DLCL_CF_EXXYYY", dove le due "X" sono le cifre che rimandano al numero della busta e le tre "Y" quelle per il numero della lettera all'interno di essa. Lo stesso codice si trova all'interno dell'attributo *@xml-id* dell'elemento radice *TEI* del documento.
La costruzione di questi documenti è iniziata dall'intestazione *teiHeader*, ovvero l'elemento TEI che descrive i metadati. Al suo interno vi sono elementi come: *correspDesc*, parte di un'integrazione recente da parte di TEI (Stadler, Illetschko e Seifert 2016), che ha consentito di codificare informazioni sulle lettere e i loro metadati quali mittente, destinatario, luoghi e date di invio e ricezione della lettera; *revisionDesc*, invece, fornisce i log delle modifiche fatte ai documenti nel tempo, importante sia per tenere traccia della gestione del consistente numero di file, sia per garantire la riproducibilità del lavoro fatto su questi; gli *xi:include* inseriti all'interno di *settingDesc* e *particDesc* hanno permesso di includere le liste che

[3] VIAF: The Virtual International Authority File.
[4] GeoNames.
[5] OPAC SBN: Catalogo collettivo delle biblioteche del Servizio Bibliotecario Nazionale.
[6] Roma: ODD Editor.
[7] Oxygen XML Editor.
[8] Repository Github progetto TEI-XML Carteggio Canneti-Fiacchi.

descrivono le entità identificate nelle lettere con una sola riga di codice per ogni file (uno per ogni tipologia di entità, in modo analogo a quanto fatto su Excel). Ogni menzione all'interno del testo di una di queste entità è stata marcata da un elemento apposito in base alla tipologia: *persName* per le persone nominate direttamente, l'elemento *rs* con valore "person" nell'attributo *@type* per le persone nominate indirettamente, *placeName* per i luoghi, *orgName* per le organizzazioni e *bibl* per le opere (con attributo *@type* con valore "print" o "manuscript" in base alla tipologia di opera). Ognuno di questi elementi TEI è stato accompagnato da un attributo *@ref* contenente il codice identificativo dell'entità in questione.
Per quanto riguarda la codifica del contenuto delle lettere è stato suddiviso in paragrafi corrispondenti a quelli originali e il passaggio a una nuova pagina della lettera è stato segnalato con l'elemento *pb*, arricchito di attributi che indicano sia la numerazione della pagina che il lato (*recto* o *verso*). Inoltre, grazie all'attributo *@facs*, è stato possibile inserire riferimenti alle digitalizzazioni del manoscritto. Queste ultime sono condivise dalla Classense tramite il protocollo IIIF (International Image Interoperability Framework), dal quale sono stati estratti i riferimenti alle immagini.[9]
Particolare attenzione è stata dedicata alla traduzione in TEI-XML delle lettere accompagnate da allegati, per le quali la Biblioteca Classense ha scelto, in fase di digitalizzazione, di rappresentarli come degli oggetti separati dalla lettera, stabilendo tuttavia una relazione semantica tra di essi. Nella codifica in TEI-XML è stato scelto, invece, di mantenerli uniti grazie all'utilizzo di *group*, un elemento che consente di raggruppare più elementi *text* "distinti [...] considerati come unità per determinati scopi" (TEI Consortium 2023). Gli allegati hanno il medesimo *@xml-id* della lettera a cui sono acclusi a eccezione del suffisso nel formato "AXX", ovvero una lettera "A" seguita da due cifre per distinguere e rappresentare diversi allegati in modo univoco (l'elemento *text* della lettera principale ha invece come suffisso la lettera "L"). La scelta di riunirli deriva dalla volontà di voler mantenere un flusso lineare di lettura fedele alle lettere originali.
Elementi come *opener*, *closer* e *postscript*, invece, hanno permesso di marcare le classiche aperture, chiusure e poscritti, che includono saluti, datazioni e firme, marcati con gli elementi *salute*, *dateline* e *signed* (Figura 1).
Nel caso di porzioni di testo illeggibili o danneggiate, i classici tre punti di sospensione dentro le parentesi quadre sono stati marcati dall'elemento *gap* insieme a un attributo che spiega i motivi dell'interruzione del testo. Invece, i testi barrati per correzione autografe sono stati segnalati con l'elemento TEI *del* accompagnato dall'attributo *@rend* con valore *overstrike* e quelli sottolineati con l'elemento *hi* corredato dall'attributo *@rend* con valore *underline*.

```
<closer>
   <salute>Resto con caramente abbracciarla.</salute>
   <salute>DVPMR.</salute>
   <dateline><placeName ref="#DLCL_CF_L0028">Perugia</placeName> <date when-iso="1718-02-12">12 Febbraio 1718</date></dateline>
   <salute>Cordialissimo Servitore e Padre Affettuosissimo nel Signore</salute>
   <signed><persName ref="#DLCL_CF_PC0001">Don Pietro Canneti</persName></signed>
</closer>

<pb xml:id="DLCL_CF_E10071_L_C2V" n="2 - Verso" facs="https://classense.unibo.it/iiif/2/25094/full/625,800/0/default.jpg"/>

<postscript>
   <p>
      Al Molto Reverendo Padre mio Osservandissimo il <persName ref="#DLCL_CF_PC0002">Padre Don Mariangelo Fiacchi</persName> Monaco Camaldolese
      <placeName ref="#DLCL_CF_L0020">Classe</placeName> <placeName ref="#DLCL_CF_L0002">Ravenna</placeName>
   </p>
   <p>
      [<foreign xml:lang="la">Januar. 2. | Februar. 3. | Martius. 3. | Aprilis 3. | Propileum 1. | Maius 6 7. | Junus 7. | 26</foreign>]
   </p>
</postscript>
```

Figura 1: Esempio marcatura chiusure e poscritto lettera.

2.3 Verso la pubblicazione web con EVT

Per coadiuvare il processo di sviluppo del supporto web dell'edizione digitale del carteggio si è scelto di utilizzare il tool EVT (Edition Visualization Technology) (Rosselli Del Turco 2019). Si tratta di un software concepito per l'edizione digitale del Codex Vercellensis,[10] e condiviso in open-source per facilitare l'elaborazione del supporto web di progetti analoghi di altre istituzioni (Del Grosso et al. 2019).
Poiché la pubblicazione della prima busta del carteggio Canneti-Fiacchi è ancora un *work in progress*,[11] si è scelto, in questa sede, solamente di accennare all'utilizzo della versione di EVT 2.0 basata su AngularJS,[12] per la quale si sono resi necessari alcuni adattamenti fondamentali per il progetto in questione: è stato

[9] Digitalizzazione Carteggio Canneti-Fiacchi.
[10] Edizione Digitale del Libro di Vercelli.
[11] Una prima versione dell'edizione digitale della Busta 10 può essere consultata a questo indirizzo https://esdcarteggiocannetifiacchi.unibo.it/.
[12] EVT 2.0.

esteso il supporto alle identificazioni di entità marcate con l'elemento *bibl* (rappresentanti le opere) e con l'elemento *rs* (le personalità nominate indirettamente); è stato integrato il supporto a progetti che partono da un elemento *teiCollection*, e di documenti che partono da un elemento radice *TEI* ma che utilizzano l'elemento *group* per racchiudere più *text*; infine è stato implementato il parsing completo degli elementi all'interno di *teiHeader*. Tramite l'utilizzo di EVT all'utente sarà, quindi, possibile visualizzare in modo chiaro e accessibile le digitalizzazioni delle lettere insieme alle trascrizioni di queste popolate di entità interattive (Figura 2).
Parimenti al progetto TEI del carteggio, anche quello dello sviluppo del sito per la pubblicazione dell'edizione digitale è stato versionato e pubblicato in una repository Github.[13]

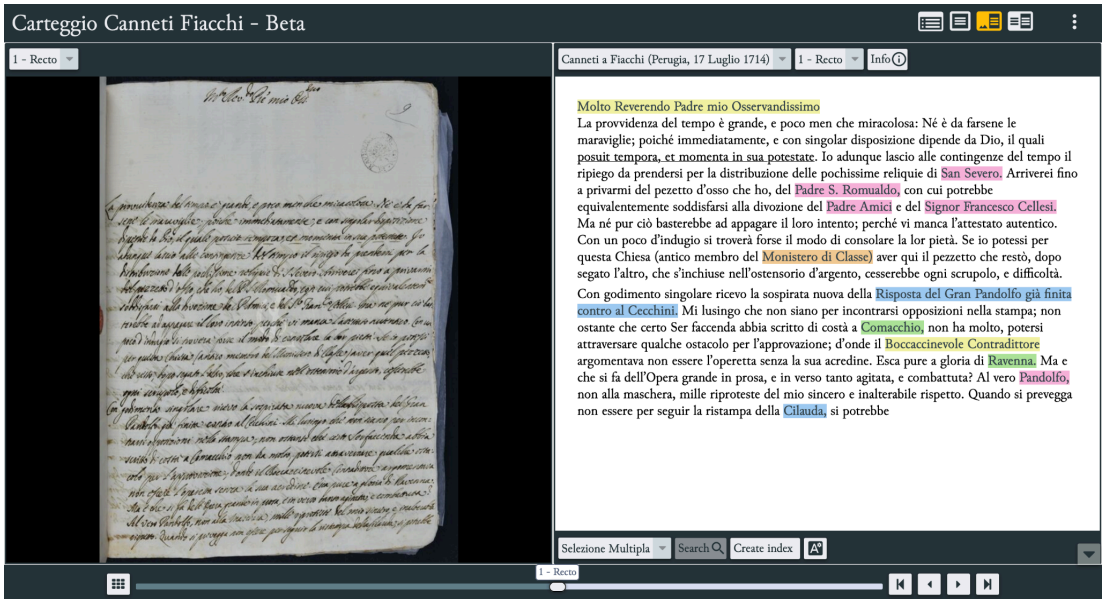


Figura 2: Visualizzazione lettera con entità evidenziate.

3. CONCLUSIONI

L'edizione digitale del carteggio Canneti-Fiacchi si presenta come uno strumento di studio interdisciplinare, che mira a combinare metodologie di ricerca storica e nuove modalità di rappresentazione in ambito digitale. Il progetto pone davanti a molteplici questioni inerenti i procedimenti e i criteri possibili per realizzare una *Digital Scholarly Edition* accessibile e fruibile online da diverse tipologie di utenti.
Con il contributo qui esposto si intende presentare la metodologia utilizzata per la trascrizione e la marcatura delle lettere manoscritte della Busta 10 del carteggio Canneti-Fiacchi, le criticità e le scelte riguardanti la prima fase del lavoro. Il completamento dell'edizione digitale, dotata di trascrizioni interattive, indici di personalità, luoghi e opere citate all'interno delle lettere, permetterà di evidenziare la preziosità del carteggio conservato dalla Biblioteca Classense di Ravenna, dando risalto ad elementi di notevole interesse scientifico, derivati da un attento lavoro di individuazione dei contenuti. Attraverso le scelte sopra descritte, il carteggio si dimostra quindi una fonte imprescindibile per lo studio del contesto storico-culturale del XVIII secolo, della Biblioteca Classense e per nuove ricerche in grado di ampliare ulteriormente il quadro di indagine.

RINGRAZIAMENTI

Il progetto è finanziato dall'Unione Europea - NextGenerationEU a valere sul Piano Nazionale di Ripresa e Resilienza (PNRR) – Missione 4 Istruzione e ricerca – Componente 2 Dalla ricerca all'impresa – Investimento Investimento 1.3, Avviso D.D. 341 del 15/03/2022, dal titolo: Cultural Heritage Active Innovation for Sustainable Society, codice proposta PE0000020.

[13] Repository Github Edizione Digitale Carteggio Canneti-Fiacchi.

BIBLIOGRAFIA

Canneti, P., Ravenna, Istituzione Biblioteca Classense, Carteggio P. Canneti - M. Fiacchi, Lettere, Buste 10-12 (1711-1730).

Del Grosso, A. M. et al. (2019). Bellini's Correspondence: a Digital Scholarly Edition for a Multimedia Museum. Umanistica Digitale, 7, 23-47. DOI: 10.6092/ISSN.2532-8816/9162. https://umanisticadigitale.unibo.it/article/view/9162.

Fiacchi, M., Ravenna, Istituzione Biblioteca Classense, Carteggio M. Fiacchi - P. Canneti, Lettere, Buste 24, fasc. 2 (anni 1714-1730).

Fiacchi, M., Canneti, P. Ravenna, Istituzione Biblioteca Classense, Miscellanea Mob. 3. 5 I^{2} (1724-1726, 1730).

Hankins, G. (2013). Correspondence: Theory, Practice, and Horizons in Literary Studies in the Digital Age. Modern Language Association of America. DOI: 10.1632/lsda.2015.13.

Jodogne, P. (2018). Il momento della trascrizione nel lavoro ecdotico in Epistolari dal due al seicento: modelli, questioni ecdotiche, edizioni, cantieri aperti. Berra C., Borsa P., Comelli M. & Martinelli S. Tempesta (a cura di), Milano, Università degli studi di Milano, Dipartimento di studi letterari, filologici e linguistici, 1-16. DOI: 10.13130/quadernidigargnano-02-03.

TEI Consortium, cur. (2023). TEI P5: Guidelines for Electronic Text Encoding and Interchange. 4.7.0. Last updated on 16th November 2023. TEI Consortium. http://www.tei-c.org/Guidelines/P5/.

Rosselli Del Turco, R. (2019). Designing an advanced software tool for Digital Scholarly Editions. Textual Cultures, 91-111. DOI: 10.14434/ textual.v12i2.27690.

Sabba, F. (2016). L'importanza dei carteggi come fonti per la storia del libro, delle biblioteche e della bibliografia. Biblioteche oggi, 4, 60-64. DOI: 10.3302/0392-8586-201604-060-1.

Stadler, P., Illetschko, M., & Seifert S. (2016). Towards a Model for Encoding Correspondence in the TEI: Developing and Implementing <correspDesc>. Journal of the Text Encoding Initiative Issue, 9. DOI: 10.4000/jtei.1433. http://dx.doi.org/ 10.4000/jtei.1433.

Vetrugno, R. (2018). Una proposta di criteri per l'edizione di carteggi rinascimentali italiani in Epistolari dal due al seicento: modelli, questioni ecdotiche, edizioni, cantieri aperti. Berra C., Borsa P., Comelli M. & Martinelli Tempesta S. (a cura di), Milano, Università degli studi di Milano, Dipartimento di studi letterari, filologici e linguistici, 597–610. DOI: 10.13130/quaderni di gargnano-02-25.

L’edizione critica digitale della 'Scienza nuova' di Giambattista Vico in Scholarly Digital Edition

Alessia Scognamiglio[1], Roberto Evangelista[2], Manuela Sanna[3], Salvatore Prinzi[4], Stefano Veneroni[5], Chiara Aiola[6], Luca de Santis[7]

[1] ISPF - CNR, Italia – alessia.scognamiglio@ispf.cnr.it

[2] ISPF - CNR, Italia – roberto.evangelista@ispf.cnr.it

[3] ISPF - CNR, Italia – manuela.sanna@ispf.cnr.it

[4] ISPF - CNR, Italia – salvatore.prinzi@ispf.cnr.it

[5] ISPF - CNR, Italia – stefano.veneroni@ispf.cnr.it

[6] Net7, Italia – aiola@netseven.it

[7] Net7, Italia – desantis@netseven.it

ABSTRACT (ITALIANO)

L’edizione critica digitale della *Scienza nuova* del filosofo napoletano Giambattista Vico in Scholarly Digital Edition è un’iniziativa che si inserisce nell’ambito del progetto PNRR del Consiglio Nazionale delle Ricerche (CNR) “H2IOSC - Humanities and Heritage Italian Open Science Cloud”, che mira a creare una federazione di Infrastrutture di Ricerca (IR), coinvolgendo i nodi italiani di 4 infrastrutture che fanno parte della roadmap dell’European Strategy Forum on Research Infrastructure nel settore dell’innovazione sociale e culturale: DARIAH-IT, CLARIN-IT, OPERAS.it, E-RIHS.it. Scholarly Digital Edition è una piattaforma digitale per pubblicare e gestire le edizioni digitali – critiche e diplomatiche – di testi prevalentemente filosofici. L’Istituto per la Storia del Pensiero Filosofico e Scientifico Moderno (ISPF) del CNR con il supporto tecnico di Net7 sta lavorando all’edizione critica digitale della *Scienza nuova* del filosofo napoletano Giambattista Vico. La *Scienza nuova* è un’opera che ha tre redazioni diverse: la prima del 1725, quella del 1730 e, infine, quella del 1744. Le ultime due edizioni presentano un impianto piuttosto simile, sebbene con diverse varianti, e sono facilmente assimilabili, peculiarità, questa, che ha permesso a Vico di essere un ottimo caso studio per un’edizione pilota. Inoltre, il “caso Vico” permette di stabilire un criterio di trattamento dei dati e dei metadati considerando le diverse varianti non solo tra manoscritto e testo a stampa, ma anche e soprattutto tra le due edizioni del 1730 e del 1744. La ricostruzione di questo lavoro è lunga e complessa, e il mezzo digitale (se adeguatamente progettato) può permettere di seguirlo e di renderlo visibile. Il caso di studio scelto aiuterà la progettazione della piattaforma a prevedere esigenze complesse di trattamento del testo.

Parole chiave: edizione critica digitale; scienza aperta; Giambattista Vico; Scienza nuova.

ABSTRACT (ENGLISH)

The Critical Digital Edition of Giambattista Vico's *Scienza nuova* in Scholarly Digital Edition. The Scholarly Digital Edition of the *Scienza nuova* of the Neapolitan philosopher Giambattista Vico is an initiative that is part of the PNRR project of the National Research Council (CNR) ‘H2IOSC - Humanities and Heritage Italian Open Science Cloud’, which aims to create a federation of Research Infrastructures (IR) involving the Italian nodes of 4 infrastructures that are part of the roadmap of the European Strategy Forum on Research Infrastructure in the field of social and cultural innovation: DARIAH-IT, CLARIN-IT, OPERAS. it, E-RIHS.it. Scholarly Digital Edition is a digital platform for publishing and managing digital editions - critical and diplomatic - of mainly philosophical texts. The Institute for the History of Modern Philosophical and Scientific Thought (ISPF) of the CNR with technical support by Net7 working on the digital critical edition of the *Scienza nuova* by the Neapolitan philosopher Giambattista Vico. The *Scienza nuova* is a work that has three different editions: the first of 1725, the edition of 1730 and, finally, that of 1744. The last two editions present a rather similar layout, albeit with different variants, and are easily assimilated, a peculiarity that made Vico an excellent case study for a pilot edition. Moreover, the ‘Vico case’ allows us to establish a data and metadata processing criterion considering the different variants not only between manuscript and printed text, but also and especially between the two editions of 1730 and 1744. The reconstruction of this work is long and complex, and the digital medium (if properly designed) can enable it to be followed and made visible. The chosen case study will help the design of the platform to anticipate complex text processing needs.

Keywords: digital critical edition; open science; Giambattista Vico; New Science.

1.IL PROGETTO E GLI OBIETTIVI

L'edizione critica digitale della *Scienza nuova* del filosofo napoletano Giambattista Vico in Scholarly Digital Edition è un'iniziativa dell'Istituto per la Storia del Pensiero Filosofico e Scientifico Moderno (ISPF) che si inserisce nell'ambito del progetto PNRR del Consiglio Nazionale delle Ricerche (CNR) "H2IOSC - Humanities and Heritage Italian Open Science Cloud" (finanziato dall'Unione Europea NextGenerationEU – PNRR M4C2 - Codice progetto IR0000029 - CUP B63C22000730005. 9), che mira a creare una federazione di Infrastrutture di Ricerca (IR), coinvolgendo i nodi italiani di 4 infrastrutture, che fanno parte della roadmap dell'European Strategy Forum on Research Infrastructure nel settore dell'innovazione sociale e culturale: DARIAH-IT, CLARIN-IT, OPERAS.it, E-RIHS.it. Scholarly Digital Edition è una piattaforma digitale per pubblicare e gestire le edizioni digitali – critiche e diplomatiche – di testi prevalentemente filosofici. L'ISPF con il supporto tecnico di Net7 sta lavorando all'edizione critica digitale della *Scienza nuova* del filosofo napoletano Giambattista Vico. Scholarly Digital Edition è una piattaforma digitale per pubblicare e gestire le edizioni digitali di testi prevalentemente filosofici, e per edizioni lavorate in Scholarly Digital Edition si intendono edizioni di testi accertate ed oggetti affidabili. Scholarly Digital Edition include sia le edizioni critiche che le edizioni diplomatiche[1].

L'edizione diplomatica offre un testo su cui si è lavorato filologicamente, ma che cela la propria genesi a favore di una maggiore leggibilità e fruibilità; questo tipo di testo è quello che più spesso viene utilizzato dai lettori, e che nasconde – quando esiste – il lavoro di interpretazione, sempre inevitabile soprattutto quando si predispone anche una traduzione. L'edizione propriamente critica, invece, presuppone un lungo lavoro di riflessione sul testo, una ricostruzione storica del contesto, la scelta dei testimoni e la costruzione dello stemma, sulla cui base si costruisce poi l'edizione. Questo lavoro non sempre è visibile nell'apparato critico, e dipende dalla scelta dei curatori e dell'editore[2].

Dal punto di vista più strettamente tecnico, la distinzione tra edizioni aiuta a definire innanzitutto il workflow per la costruzione delle edizioni critiche digitali, in quanto predispone due diversi flussi di lavoro: uno dedicato alle edizioni diplomatiche e uno dedicato alle edizioni critiche[3]. Si sta cercando di valutare la possibilità di far incrociare i due flussi di lavoro per rendere critica un'edizione diplomatica, sfruttando la facoltà di rendere flessibile e aperto il workflow per far sì che l'edizione critica possa essere arricchita nel tempo. Su questo aspetto occorre fare una precisazione: esiste una differenza tra la costruzione di un'edizione critica e la trascrizione di un'edizione critica. L'apertura e la flessibilità del workflow può valere solo nel primo caso, dal momento che, nel secondo caso, intervenire nuovamente sul testo (se non per correggere eventuali errori di trascrizione) significherebbe tornare indietro, e ciò non è comprensibile, a meno che non si voglia rifare una edizione critica a fronte di nuove esigenze o di ritrovamenti di nuovi testimoni. Bisogna predisporre le necessarie "limitazioni" per rispettare le scelte degli editori. Ciò non toglie che l'edizione critica già stabilita possa prevedere interventi successivi, atti ad elaborare un metatesto "aumentato" di notizie, di contesti e riferimenti esterni al testo stesso. Si tratta però in questo caso di un apparato che rispecchia le risorse offerte dal mezzo digitale e che va oltre il concetto di edizione critica, coprendo un ambito più ampio di meta-digitalizzazione del testo. Individuare queste potenzialità, ordinarle in workflow flessibili, permette sicuramente una maggiore versatilità della piattaforma, tali da riuscire a coprire esigenze di utenza più ampie[4].

Gli altri obiettivi da raggiungere, che sono anche aspetti legati alla progettazione, sono la modellizzazione e gli standard di trattamento dei metadati. Nel caso della modellizzazione si può presupporre che la piattaforma non leghi a vincoli particolari, e che l'editore possa essere libero di decidere con la più ampia libertà le scelte legate al testo che si vuole digitalizzare. Nel caso, invece, del trattamento dei dati e dei metadati bisogna considerare la differenza tra edizioni diplomatiche ed edizioni critiche. Le edizioni di testi sono fondamentalmente fatte di dati: questo vale prima di tutto per le edizioni cartacee tradizionali. Per quanto riguarda le edizioni diplomatiche, invece, il dato può essere facilmente modellato e classificato secondo macro categorie. Il testo, infatti, è di per sé un documento che viene considerato definitivo. Nella modellizzazione dell'edizione questa premessa si traduce in una meta-datazione piuttosto leggera e macroscopica che viene in contro alle esigenze della macchina. L'edizione critica, invece, il documento –

1 Pierazzo, E., & Rosselli Del Turco R. (2023). Critica testuale e nuovi metodi: l'edizione scientifica digitale. Ciotti F. (Eds.). Digital Humanities. Roma: Carocci.
2 Italia, P., & Bonsi C. (2016). Edizioni Critiche Digitali. Digital Critical Editions. Edizioni a confronto. Roma: Sapienza Università Editrice.
3 Mancinelli, T., & Pierazzo, E. (2020). Che cos'è una edizione scientifica digitale. Roma: Carocci.
4 Pierazzo, E. (2015). Digital Scholarly Editing: Theories, Models and Methods. Farnham, Surrey: Ashgate.

cioè il testo finale – lo deve costruire. Questa costruzione si fa o sulla base di altri documenti (cioè i testimoni) o sulla base di dati molto particolari che vanno resi poi nella modellizzazione; ci riferiamo ad esempio agli errori, le postille, le varianti etc. A loro volta, i documenti utilizzati per arrivare al testo definitivo, sono costruiti da dati che devono essere riprodotti. Occorre dunque immaginare workflow che possano gestire una grande mole di dati e che lascino molto spazio al lavoro umano, più che all'intelligenza artificiale. Questo rende il lavoro dell'edizione critica digitale un lavoro che non si differenzia così tanto dal lavoro tradizionale, e che permette sempre di ricostruire il testo. Siamo portati a pensare che questo sia un bene e che permetta alla macchina (in questo caso alla piattaforma che si sta progettando) di velocizzare il lavoro di ecdotica ma non di stravolgerlo. È questo, infatti, uno degli obiettivi che ci si propone.

2.STATUS QUAESTIONIS

L'edizione pilota sulla quale si è scelto di lavorare è La *Scienza nuova* del filosofo napoletano Giambattista Vico. La *Scienza nuova* è un'opera che ha tre redazioni diverse: la prima del 1725, quella del 1730[5] e, infine, quella del 1744[6]. Le ultime due edizioni presentano un impianto piuttosto simile, sebbene con diverse varianti, e sono facilmente assimilabili, peculiarità, questa, che ha permesso a Vico di essere un ottimo caso studio per un'edizione pilota. Oltre a effettuare la trascrizione dei testi si possono infatti provare diverse opzioni di allineamento, e questo rende il progetto piuttosto originale e innovativo rispetto ad edizioni critiche digitali di altri testi. Il "caso Vico" permette di stabilire un criterio di trattamento dei dati e dei metadati, in quanto le diverse varianti non solo tra manoscritto e testo a stampa, ma anche e soprattutto tra le due edizioni della *Scienza nuova*, nello specifico quella del 1730 e quella del 1744. Questo aspetto è particolarmente importante se si tiene presente che nei 14 anni che separano le due redazioni, quella del 1730 e quella del 1744, Vico agisce sul testo in maniera specifica e differente, non solo annotando una serie di correzioni su fogli separati, ma anche intervenendo direttamente su alcune copie del volume. Come si vede, la ricostruzione di questo lavoro è lunga e complessa, e il mezzo digitale, se adeguatamente progettato, permette di seguirlo e di renderlo visibile. Il caso di studio scelto aiuterà la progettazione della piattaforma a prevedere esigenze complesse di trattamento del testo. L'edizione della *Scienza nuova* rappresenta pertanto una grande sfida rispetto alla progettazione di una piattaforma per le edizioni critiche digitali, e si avvale dell'importante lavoro di edizione critica tradizionale, che ha già stabilito il testo per tutte e tre le redazioni della *Scienza nuova*, compiendo di fatto già il lavoro di esame e scelta dei testimoni, e di ricostruzione del testo.

3.METODOLOGIE

L'esigenza principale è progettare una piattaforma che sia in grado di raccogliere materiali eterogenei, fornire strumenti per operare su tali materiali e predisporre uno o più output in base alle necessità del progetto. Per quanto riguarda il caso d'uso utilizzato come pilota, ovvero la pubblicazione della *Scienza nuova* di Vico sia come edizione critica che sinottica, la piattaforma dovrà essere in grado di accogliere i testi delle opere, mostrare una versione diplomatica dei testi presentando il facsimile a fronte (si è scelto di affiancare al testo l'editio princeps) e consentire una visione sinottica delle due edizioni della *Scienza nuova*, quella del 1730 e quella del 1744. Allo stesso tempo la piattaforma dovrà essere tanto flessibile da adattarsi a casi d'uso diversi da quello pilota; si è scelto, pertanto, di concepirla come una dashboard di funzionalità (Figura 1), che ogni gruppo di ricerca può attivare e combinare nel modo più adatto al raggiungimento dei propri obiettivi. Le funzionalità previste per questa prima versione della piattaforma saranno relative alla gestione dei testi, delle immagini e alla configurazione di possibili modelli di output.

5 Cristofolini, P., & Sanna, M. (Eds.). 2004. Giambattista Vico La *Scienza Nuova* 1730. Napoli: Alfredo Guida Editore.
6 Cristofolini, P., & Sanna, M. (Eds.). 2013. Giambattista Vico La *Scienza Nuova* 1744. Roma: Edizioni di Storia e Letteratura.

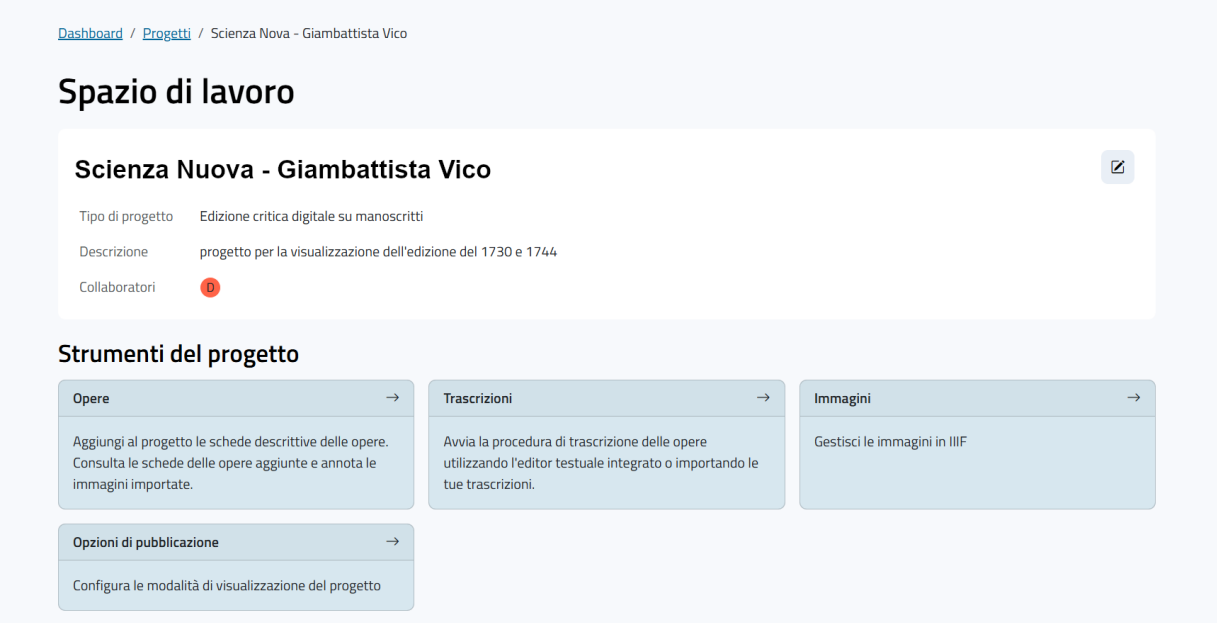


Figura 1. Dashboard degli strumenti di progetto

Alla dashboard, o back office, che permette al ricercatore di lavorare sui materiali di input con una serie di strumenti integrati volti alla costruzione dell'edizione digitale, si affianca un frontend pubblico, che ospiterà e renderà fruibile a tutta la comunità il risultato che il responsabile del progetto vorrà rendere pubblico, attraverso un percorso di navigazione che va dalla presentazione del progetto, con i propri metadati, all'edizione del testo vera e propria (figura 2 e figura 3).

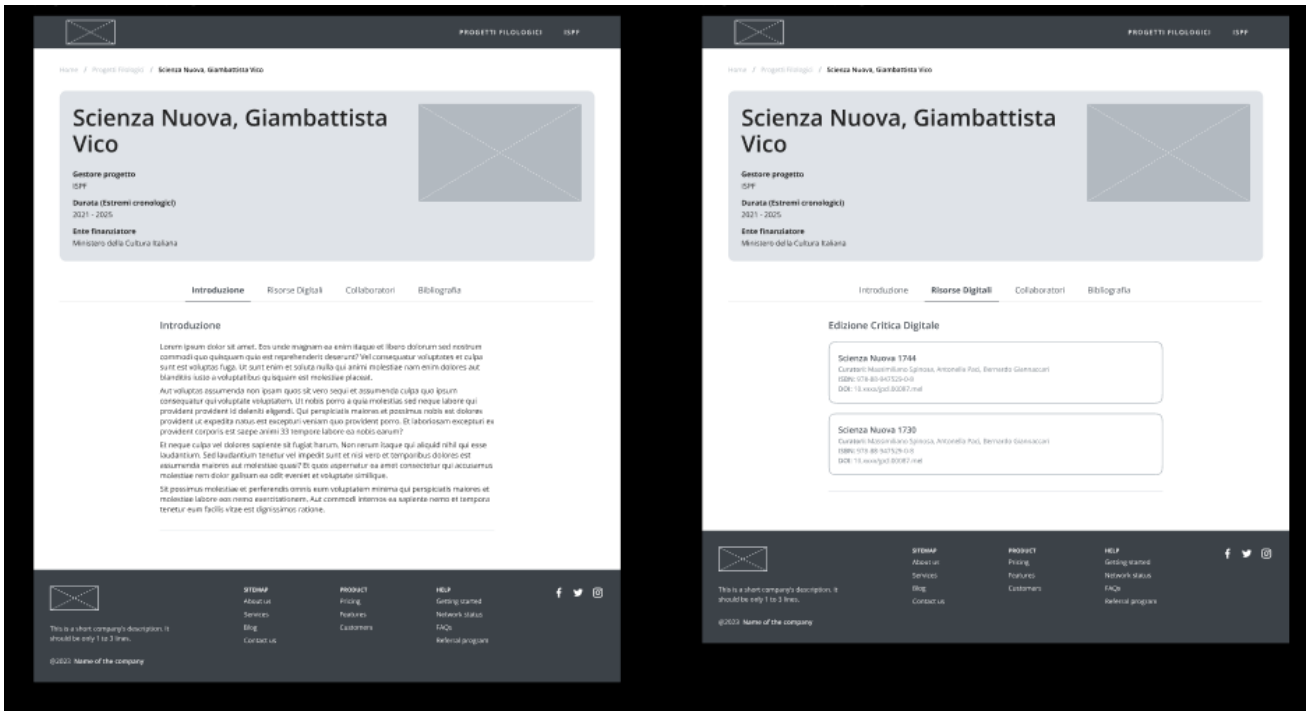


Figura 2. Mockup del frontend, vista progetto e risorse digitali

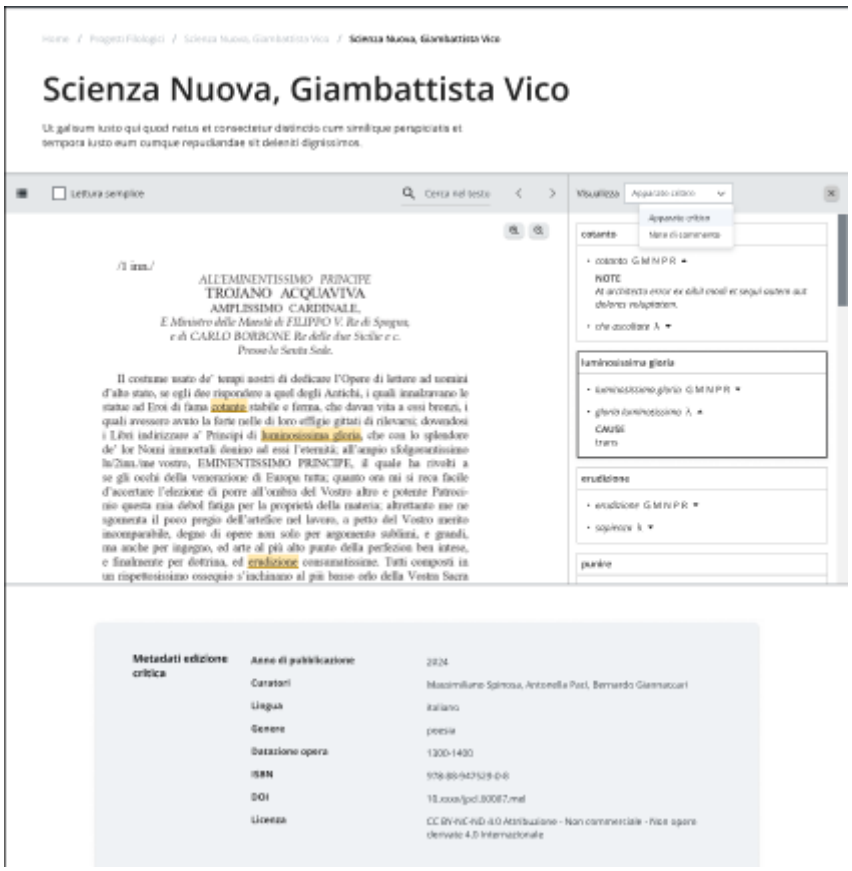


Figura 3. Mockup visualizzazione del testo critico con apparato

4.GESTIONE DEI TESTI E GESTIONE DEL FACSIMILE

La gestione dei testi deve da un lato favorire l'inserimento e la modifica dei testi in modalità semplici per l'utente, ma allo stesso tempo i dati devono essere compatibili con gli standard ormai condivisi in termini di codifica del testo, ovvero XML/TEI. Molti testi (e quello della *Scienza nuova* non è certo un caso isolato) sono già editi e disponibili in vari formati digitali, per quanto non codificati secondo il linguaggio XML. Il primo obiettivo quindi è stato rendere possibile il recupero di questo materiale tramite processi automatici o semiautomatici. Nel caso di Vico i testi, disponibili inizialmente in formato docx, sono stati prima convertiti tramite trasformazioni XSLT in un documento XML con una marcatura TEI volta ad individuare gli elementi principali di struttura (divisioni in paragrafi, titoli etc…). In seguito, tramite una serie di script Python, il testo codificato è stato ripulito da tutto il materiale extra testuale (intestazioni, piè di pagina, segni diacritici), ed è stato reso pronto per eventuali lavorazioni successive, che consentiranno di operare direttamente su un documento fairificato e codificato in XML-TEI, su cui si potrà aggiungere l'apparato critico, individuare named entities o altri elementi di codifica in base alle esigenze dello specifico progetto. Analogamente alla gestione dei testi, la piattaforma consentirà il caricamento delle digitalizzazioni dei testi. Lo standard di riferimento in questo caso sarà IIIF[7]. L'adozione di questo standard consente non solo l'ottimizzazione della fruizione delle immagini e la standardizzazione dei metadati descrittivi, ma anche la possibilità di organizzare in sequenze le singole riproduzioni tramite la struttura del file manifest.json. Tale organizzazione, basata su identificativi univoci, consente anche di creare sistemi di riferimenti tra le pagine codificate e l'immagine corrispondente. La piattaforma dovrà consentire di creare tali riferimenti attraverso interfacce intuitive e integrate.

5.CONFIGURAZIONE DELLE VISUALIZZAZIONI

Questa è la fase più avanzata, ed è quella che consente di impostare le possibili visualizzazioni del materiale fin qui raccolto. Se negli step precedenti è stato fatto un buon lavoro di raccordo tra le varie

7 https://iiif.io/ IIIF (International Image Interoperability Framework) è uno standard aperto che consente la visualizzazione e l'interoperabilità di immagini digitali ad alta risoluzione.

tipologie di materiale, i possibili output del progetto di pubblicazione di edizione potranno essere: facsimile con testo a fronte, testo e apparato critico, più testi paralleli.
Il caso di studio su Vico sarà un importante banco di lavoro per capire quanto gli strumenti progettati saranno adatti allo scopo specifico, pur mantenendosi generici per poter accogliere in futuro altri casi d'uso.

6.PROBLEMI EMERSI

Sono emersi alcuni problemi: il primo è relativo al workflow, il secondo è relativo alla resa del manoscritto e all'allineamento con il testo, il terzo al trattamento dei dati. Per quanto riguarda il primo problema è necessario salvaguardare la visibilità dell'aspetto interpretativo che sta dietro il lavoro di edizione, soprattutto di edizione critica. Non importa, infatti, offrire un testo che sia immediatamente verificabile, ma offrire un testo stabilito secondo scelte motivate, che possono essere rese visibili su richiesta dello studioso e dell'editore; non è necessario mostrare queste scelte, ma deve essere possibile farlo. L'altro problema è relativo alla resa del manoscritto e all'allineamento con il testo, e anche in questo caso dovranno pesare le scelte dell'editore. L'ultimo ordine di problema, forse il più spinoso, è quello relativo alla gestione dei dati e alla loro metadatazione. Le edizioni digitali, soprattutto quelle critiche, hanno un sistema complesso di metadatazione, a meno di non voler costruire edizioni povere che perderebbero di efficacia. Da questo punto di vista è necessario coordinarsi con l'IR OPERAS per trovare un sistema che sia omogeneo e flessibile allo stesso tempo e che salvaguardi le specificità di ogni oggetto.

7.RISULTATI ATTESI

Il risultato atteso è una edizione critica digitale delle redazioni della *Scienza nuova* di Giambattista Vico del 1730 e del 1744, corredata da un apparato che descriva e analizzi ciò che è contenuto nei diversi testimoni, per mostrare il processo della genesi e dell'evoluzione di questa importante opera attraverso la rappresentazione e l'interpretazione delle varianti.
In questa fase del progetto il risultato atteso è un minimum viable product (MVP), la piattaforma Scholarly Digital Edition, che fornisca un back-office per l'ISPF che consenta tanto di gestire e organizzare i dati del progetto secondo gli standard condivisi con il Common semantic framework di H2IOSC (secondo il WP4 del progetto), quanto di gestire la pubblicazione di edizioni digitali di testi a stampa. Il primo pilota sarà quindi l'edizione della *Scienza nuova* di Vico. Nella fase iniziale si lavorerà sulle edizioni del 1730 e del 1744, per le quali esiste già la versione tradizione dell'edizione critica. La piattaforma consentirà sia la consultazione delle edizioni della *Scienza nuova* del 1730 e del 1744 in maniera separata, ognuna accompagnata dal proprio apparato critico e dal facsimile della editio princeps, che la consultazione sinottica delle due redazioni, per consentire di fruire l'evoluzione dell'opera, oltre che lo svolgimento del pensiero di Vico attraverso le redazioni di essa.

RINGRAZIAMENTI

Progetto H2IOSC - Humanities and cultural Heritage Italian Open Science Cloud finanziato dall'Unione europea NextGenerationEU - Piano Nazionale di Ripresa e Resilienza (PNRR) - Missione 4 "Istruzione e Ricerca" Componente 2 "Dalla ricerca all'impresa" Linea di Investimento 3.1 "Fondo per la realizzazione di un sistema integrato di infrastrutture di ricerca e innovazione" Azione 3.1.1 "Creazione di nuove IR o potenziamento di quelle esistenti che concorrono agli obiettivi di Eccellenza Scientifica di Horizon Europe e costituzione di reti" - Codice progetto IR0000029 - CUP B63C22000730005. Soggetto attuatore CNR.

BIBLIOGRAFIA

Cristofolini, P., & Sanna, M. (Eds.). 2004. Giambattista Vico. La Scienza Nuova 1730. Napoli: Alfredo Guida Editore.

Cristofolini, P., & Sanna, M. (Eds.). 2013. Giambattista Vico. La Scienza Nuova 1744. Roma: Edizioni di Storia e Letteratura.

Italia P., & Bonsi C., (2016). Edizioni Critiche Digitali. Digital critical editions. Edizioni a confronto. Roma: Sapienza Università Editrice.

Mancinelli, T., & Pierazzo, E. (2020). Che cos'è una edizione scientifica digitale. Roma: Carocci.

Pierazzo, E. (2015). Digital Scholarly Editing: Theories, Models and Methods. Farnham, Surrey: Ashgate.

Pierazzo, E., & Rosselli Del Turco R. (2023). Critica testuale e nuovi metodi: l'edizione scientifica digitale. Ciotti F. (Eds.). Digital Humanities. Roma: Carocci.

.

L'edizione digitale del papiro P.Tor.Choach. 12 in collaborazione con il Museo Egizio di Torino

Chiara Senatore
Università degli Studi di Roma "La Sapienza", Italia, chiara.senatore@uniroma1.it

ABSTRACT (ITALIANO)

Il papiro *P.Tor.Choach. 12* rappresenta il documento papiraceo più importante del fondo greco del *Museo Egizio* di Torino e quello più importante del cosiddetto "archivio dei Choachiti" (P. Tor. Choachiti). Si tratta di un papiro documentario del 117 d.C., scritto in greco, contenente l'ultimo atto del processo tra Hermias e i Choachiti. Tale papiro rappresenta uno dei casi di studio del progetto "Risorse digitali per i manoscritti", il cui scopo è quello di offrire alle istituzioni culturali nazionali edizioni scientifiche digitali di documenti che custodiscono, con lo scopo di ampliarne la fruizione presso un pubblico non accademico. L'edizione, realizzata in stretta collaborazione con il *Museo Egizio*, è in corso di realizzazione utilizzando il software EVT (Edition Visualization Technology).
Lo scopo del presente contributo è quello di presentare il progetto, le metodologie utilizzate e i risultati attesi.

Parole chiave: edizione scientifica digitale; papirologia; GLAM; EVT; annotazione semantica**.**

ABSTRACT (ENGLISH)

Digital edition of the papyrus P.Tor.Choach. 12 in collaboration with the Museo Egizio in Turin.
The papyrus *P.Tor.Choach. 12* is the most important papyrus of the Greek collection of the *Museo Egizio* in Turin and the most important of the so-called "Choachiti's archive" (P. Tor. Choachiti). It is a documentary papyrus from 117 AD, written in Greek, which contains the last act of the process between Hermias and the Choachites. This papyrus is one of the case studies of the "Digital Resources for Manuscripts" project, which aims to provide national cultural institutions with digital scholarly editions of documents they hold, in order to extend their use to a non-academic public. The edition, prepared in close collaboration with the *Museo Egizio*, is produced using EVT (Edition Visualization Technology) software.
The aim of this paper is to present the project, the methods used and the expected results.

Keywords: digital scholarly editing; papyrology; GLAM; EVT; semantic annotation.

1. INTRODUZIONE

Il progetto "Risorse Digitali per i Manoscritti", nato nel contesto del dottorato nazionale in *Heritage Science* (38° ciclo) finanziato dal PNRR e coordinato dall'Università "La Sapienza" di Roma, ha lo scopo di offrire alle istituzioni culturali nazionali -pubbliche e private- edizioni digitali di documenti che queste custodiscono, per ampliare i fruitori del patrimonio culturale digitalizzato. Infatti, uno dei problemi principali conseguito in seguito alla massiccia campagna di digitalizzazione del patrimonio culturale attuata dal MiC tra il 2016 e il 2020 è la fruizione estremamente settoriale di alcune risorse digitali. Ad esempio, la maggior parte delle edizioni scientifiche digitali e dei facsimili digitali dei manoscritti sono noti solo ad un pubblico di specialisti -e spesso solo da questi utilizzati. L'obiettivo del progetto è quello di cercare di avvicinare anche un pubblico meno settoriale al patrimonio culturale digitalizzato tramite la creazione di edizioni digitali in collaborazione con le istituzioni culturali pubbliche e private che si occuperanno di farle circolare sui propri portali web e di inserirle -ove possibile- nei percorsi museali. Le istituzioni culturali attualmente coinvolte nel progetto sono la *Biblioteca Reale* di Torino, il *Museo Egizio* di Torino, l'*Officina Museo della Scrittura* di Torino e la *Biblioteca Estense Universitaria* di Modena.
Per questo scopo è stato scelto il software EVT (Edition Visualization Technology, http://evt.labcd.unipi.it) e la sua applicazione è stata estesa a casi di studio che non appartengono all'ambiente accademico.
EVT è un software *open source* che permette di creare edizioni digitali consentendo la visualizzazione del facsimile digitale con la rispettiva trascrizione diplomatica e/o interpretativa, permettendo anche l'inserimento di un collegamento testo-immagine. L'edizione digitale è creata partendo dal testo codificato

in formato XML-TEI ed EVT supporta come facsimili digitali anche risorse IIIF. È possibile creare liste di *named entity*, nelle quali possono essere inseriti riferimenti a risorse esterne. Il software utilizza le tecnologie standard del web (HTML 5, CSS, JavaScript) per garantire l'accessibilità a lungo termine dell'edizione (per un approfondimento sul software cfr. Rosselli Del Turco 2019).
Lo scopo del presente contributo è quello di presentare l'edizione digitale del caso di studio offerto dal *Museo Egizio* di Torino: il papiro *P.Tor.Choach. 12*, un papiro documentario del II secolo d.C. scritto in greco, contenente una controversia giudiziaria. È in corso di realizzazione un'edizione del papiro che comprende la trascrizione e la traduzione del papiro, corredate da una -almeno- duplice lista di *named entity*. La destinazione finale dell'edizione è ancora in corso di definizione: sicuramente essa sarà inserita nella piattaforma *Turin Papyrus Online Platform* (TPOP) del *Museo Egizio* (https://collezionepapiri.museoegizio.it) e si sta valutando la fattibilità per inserire l'edizione anche su di un dispositivo *touchscreen* nel percorso di visita museale.

2. STATO DELL'ARTE

Il progetto si inserisce in due contesti importanti: quello dell'edizione digitale di testi papiracei e quello dell'inserimento di edizioni digitali in percorsi museali.
Numerosi sono state le iniziative volte alla creazione di repertori di testi papiracei ed edizioni di papiri di rilievo scientifico. Alcuni progetti intendono offrire una collezione di diversi documento, come nel caso del progetto *Papyri.info* (https://papyri.info), che funge da aggregatore di risorse digitali per la papirologia, offrendo edizioni testuali di documenti e dei relativi materiali, traduzioni, commenti, metadati, cataloghi istituzionali, bibliografie e facsimili digitali. Altri progetti, al contrario, intendono focalizzarsi sull'edizione digitale di un singolo documento: è il caso, ad esempio, del progetto *The Greek Philosophical Schools* (https://greekschools.eu/project/), che intende fornire una nuova edizione critica del trattato *L'ordinamento dei filosofi* di Filodemo di Gadara, che rappresenta la più antica storia della filosofia dell'antichità e che è custodito esclusivamente in papiri di Ercolano.
Per quanto riguarda l'inserimento di edizioni digitali in contesti museali, nell'ultimo biennio EVT è stato scelto in due casi come software di visualizzazione a tale scopo: si tratta del caso del *Museo Civico Bellini* di Catania e del *LETS* di Trieste. Per quanto riguarda il primo caso, il museo ha inserito nel proprio programma di valorizzazione BellinInRete, il cui scopo è quello di usare il digitale al servizio del patrimonio museale, il progetto *Bellini Digital Corrispondence* (https://bellinicorrespondence.cnr.it). Con tale progetto si mirava a creare un'edizione scientifica digitale delle lettere autografe del compositore catanese Vincenzo Bellini e di inserirne un estratto su dispositivi multimediali nel percorso museale. Il progetto è stato inserito nel percorso museale nel mese di marzo 2023. Per quanto riguarda il secondo caso, nella sezione del percorso museale dedicata a Saba, è stata inserita su un dispositivo multimediale l'edizione scientifica digitale del prezioso manoscritto contenente il *Canzoniere* di Saba (https://lets.trieste.it/museo-saba/canzoniere/): quello databile nel 1919 e noto come il primo manoscritto autografo dell'opera. Il museo è stato inaugurato nel mese di settembre 2024.

3. METODOLOGIA

Nella prima parte del lavoro, la metodologia prediletta è stata quella puramente filologica. Riguardo la trascrizione, si è scelto di utilizzare quella di Pestman (Pestman, 1992). Una volta ricevuti i facsimili digitali, ho iniziato a lavorare alla traduzione del testo: l'unica traduzione italiana disponibile è quella ormai un po' datata di Pestman (Pestman, 1992) mentre è disponibile una traduzione inglese più recente curata da Bagnall (Bagnall, 2004: 218-225). Si valuterà la possibilità di inserire anche una traduzione in inglese per ampliare ulteriormente l'utenza dell'edizione. Lavorando alla traduzione del documento, è stato necessario definire fin dall'inizio alcuni criteri da adottare in collaborazione con il museo: si è optato per la traslitterazione dei nomi propri degli attori nel processo e non per la rispettiva forma italianizzata, così come sono stati lasciati traslitterati e non tradotti i titoli onorifici. Infatti, benché a prima impressione la traslitterazione potrebbe sembrare una scelta meno adatta della traduzione se si intende arrivare ad un pubblico non specialista, è anche vero che la traduzione di alcuni titoli onorifici avrebbe portato spesso ad una ambiguità che avrebbe creato ancora maggior confusione: è il caso, per citarne uno, di τῶν φίλων, lettalmente traducibile come "gli amici" che però nei papiri documentari di età tolemaica costituisce un titolo onorifico.

Una volta trascritti e tradotti i documenti, è stato necessario scegliere il software di visualizzazione più adatto allo scopo. La scelta è ricaduta su EVT che, essendo EVT uno strumento molto flessibile, è apparso fin da subito il software di visualizzazione più adatto per questo progetto: l'annotazione semantica e i collegamenti ipertestuali permettono di creare approfondimenti su alcune entità marcate nel testo (per esempio personaggi storici rilevanti, luoghi, associazioni, date e così via); essendo uno strumento configurabile e personalizzabile, esso permette di far decidere all'editore quali funzionalità rendere visibili. Un altro aspetto tenuto in considerazione è stata la destinazione ipotizzata per le edizioni: si è subito pensato all'inserimento delle singole edizioni sui portali e/o su dispositivi *touchscreen* all'interno dei percorsi di visita delle istituzioni culturali che ne custodiscono il documento analogico. Anche a questo scopo, è stata riconfermata la scelta di EVT: il software produce l'edizione automaticamente una volta copiata la cartella sul web o sul dispositivo di destinazione e l'utilizzo di tecnologie standard del Web (HTML, CSS, JavaScript) garantisce l'accessibilità a lungo termine delle edizioni.
Il *Museo Egizio* ha selezionato il papiro in questione come caso di studio da fornire per il progetto in considerazione del fatto che esso -come precedentemente anticipato- rappresenta il documento più importante del fondo greco del *Museo*. Tuttavia, si tratta di un argomento difficile da rendere accessibile: il testo risulta particolarmente esteso per essere un papiro documentario, essendo distribuito su ben dieci colonne, e l'argomento non è tra i più accattivanti per un pubblico ampio trattandosi di un argomento giuridico. La criticità principale emersa nel corso della traduzione è stata quella di fornire una versione tradotta del testo che fosse scientificamente valida e allo stesso tempo riuscisse a semplificare un campo abbastanza settoriale come quello del diritto tolemaico. Attualmente l'edizione è stata realizzata con EVT2, che non presenta un'apposita sezione per la traduzione del testo. Per ovviare a ciò, la traduzione è stata momentaneamente inserita nella sezione <front> (vd. Fig. 1).

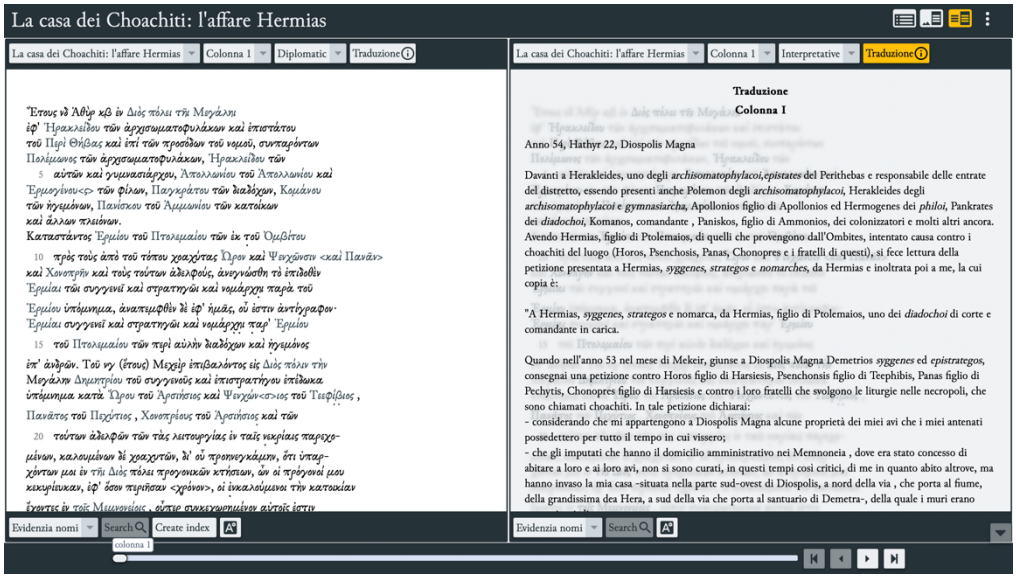


Figura 1. Traduzione nella tab <front> in EVT2

Di pari passo, si è proceduto con la codifica della trascrizione e della traduzione. Il testo e la sua rispettiva traduzione sono stati codificati in XML-TEI, seguendo le attuali linee guida (http://www.tei-c.org/Guidelines/P5/) per l'annotazione semantica. Per quanto riguarda la trascrizione, è stato necessario utilizzare particolari regole CSS per poter rispettare fedelmente la convenzione di Leida (il sistema standard di segni diacritici per l'edizione di testi antichi) per i segni diacritici presenti nella trascrizione di Pestman. Anche in questo caso si riporta un esempio a titolo esemplificativo: sono numerose le lettere danneggiate o la cui lettura comunque non risulta chiara all'interno del papiro, individuate dal punto sottoscritto. Per poter ottenere il punto sottoscritto, è stata aggiunta una regola CSS per l'elemento <unclear>, che è la seguente:

```
.unclear {
text-emphasis-style: dot;
   text-emphasis-position: under ;
```

}

Il testo è stato annotato semanticamente e sono state seguite le linee guida EpiDoc al fine di evidenziarne le scelte editoriali di Pestman: sono state marcate le lettere incerte, quelle restaurate, quelle cancellate e quelle integrate. Utilizzando l'attributo "reason", queste ultime sono state suddivise in lettere integrate dall'editore perché omesse nel testo e lettere integrate dall'editore perché perse nel testo. Sono stati inoltre marcati toponimi e antroponimi. Ogni toponimo è visualizzabile *infratesto* su mappa (vd. Fig. 2).



Figura 2. Toponimo in EVT2

Considerato lo scopo divulgativo presso un pubblico non specializzato, uno degli intenti da realizzare sarebbe quello di riuscire ad inserire una lista anche per le numerose cariche politiche e amministrative citate nel testo. L'elemento TEI da scegliere in questo caso per la marcatura dovrebbe essere <roleName>, da utilizzare all'interno dell'elemento <label> accompagnato da una breve descrizione della carica all'interno dell'elemento <item>. Tuttavia, il problema che ne deriva è che EVT2 non permette di visualizzare per l'elemento <list> le occorrenze e le informazioni aggiuntive a differenza di quanto avviene per <listPerson> e <listPlace>.
I metadati descrittivi del papiro sono stati inseriti nella sezione "MS Desc". Infine, per rendere l'interfaccia utente adatta ad un documento papiraceo, sono state rinominate alcune voci del file it.json (MS è stato rinominato descrizione del papiro, MS_CONTENTS contenuto del papiro, MS_IDENTIFIER identificatore del papiro e così via). Per cercare di rendere più intuitiva la sezione dedicata alle liste di *named entity* si è optatato per la dicitura "Evidenzia Nomi/Nessuna Evidenziazione" per la rispettiva tab di selezione.

4. RISULTATI ATTESI

Il risultato auspicato è la realizzazione di un'edizione digitale adatta sia per un pubblico di specialisti sia per un pubblico più ampio e meno specializzato come può essere quello dei visitatori del museo. L'obiettivo è quello di riuscire a inserire un collegamento ipertestuale all'edizione nella scheda del relativo papiro sulla piattaforma TPOP del *Museo Egizio* e contemporaneamente di inserire l'edizione su un dispositivo *touchscreen* nel percorso di visita museale. Nel caso in cui quest'ultimo obiettivo non dovesse risultare fattibile per il museo, in accordo con la curatrice del museo, la dottoressa Töpfer, si è pensato di inserire sulla teca che custodisce il papiro un QR code scansionabile che rimandi all'edizione digitale.
Essendo attualmente in corso di sviluppo la versione beta di EVT3, per la quale si sta lavorando ad una versione tripartita dell'interfaccia utente (documento-trascrizione-traduzione), l'obiettivo è quello di utilizzare la versione nascente di EVT3 per ottenere questo scopo, potendo così visualizzare la traduzione in una sua apposita sezione accanto al facsimile digitale e alla trascrizione.
Inoltre, per poter riconoscere una validità scientifica all'edizione, sarebbe auspicabile riuscire ad attribuire ad essa un DOI/ISBN una volta completato il lavoro (sul punto cfr. Pierazzo-Rosselli Del Turco 2023: 94).

RINGRAZIAMENTI

Si ringrazia il prof. Rosselli Del Turco per la revisione alla codifica, la prof.ssa Rosa Maria Piccione per la revisione alla traduzione e la dott.ssa Susanne Töpfer in rappresentanza del *Museo Egizio* di Torino per aver fornito il caso di studio e i relativi facsimili digitali.

BIBLIOGRAFIA

Bagnall, R. S. (2004). *The Hellenistic period: historical sources in translation*, Malden, MA : Blackwell.

Bellini Digital Correspondence, n.d. Accessed April 1, 2025. https://bellinicorrespondence.cnr.it

EVT- Edition Visualization Technology, n.d. Accessed April 1, 2025. http://evt.labcd.unipi.it

Papyri.info, n.d. Accessed April 1, 2025. https://papyri.info

Pestman, P. W. (1992), *Il processo di Hermias e altri documenti dell'archivio dei choachiti: (P. Tor. Choachiti) ; papiri greci e demotici conservati a Torino e in altre collezioni d'Italia Serie 1 vol. 6*, Torino: Ministero per i Beni Culturali e Ambientali, Soprintendenza al Museo delle Antichità Egizie.

Pierazzo E., & Rosselli Del Turco, R. (2023). *Critica testuale e nuovi metodi: l'edizione scientifica digitale* in Ciotti F. (a cura di), *Digital Humanities. Metodi, strumenti, saperi*. Roma: Carocci, 80-96.

Progetto Saba1919, n.d. Accessed April 1, 2025. https://lets.trieste.it/museo-saba/canzoniere/

Rosselli Del Turco, R. (2019). *Designing an Advanced Software Tool for Digital Scholarly Editions: The Inception and Development of EVT (Edition Visualization Technology)*. *Textual Cultures* 12/2, 91–111. https://doi.org/10.14434/textual.v12i2.27690 .

TEI Consortium, eds. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. P5 Version 4.7.0. Last updated November 16, 2023. TEI Consortium. http://www.tei-c.org/Guidelines/P5/ .

TPOP- *Turin Papyrus Online Platform*, n.d. Accessed April 1, 2025. https://collezionepapiri.museoegizio.it

Dal palcoscenico al digitale: modelli di data visualization per la valorizzazione dell'Archivio Teatro delle Marionette di Gianni e Cosetta Colla

Elena Radaelli
Università degli Studi di Milano, Italia – elena.radaelli@unimi.it

ABSTRACT (ITALIANO)

Il presente contributo illustra una sezione del progetto "Marionette 3.0", realizzato nell'ambito del WP4 dello Spoke 2 del PNRR CHANGES, con l'obiettivo di comunicare e valorizzare il patrimonio teatrale conservato nell'Archivio Teatro delle Marionette di Gianni e Cosetta Colla, custodito presso il Centro Apice dell'Università degli Studi di Milano. Le marionette costituiscono il patrimonio materiale dell'archivio, mentre la storia delle loro performance, ricostruita attraverso i documenti conservati nell'archivio, ne costituisce il patrimonio immateriale. Le testimonianze della storia della Compagnia hanno orientato il progetto verso la realizzazione di prodotti digitali rivolti a target differenziati, come eBook per bambini, storytelling per adulti, modellazioni 3D delle marionette, e data visualization del materiale d'archivio, che costituiscono il nucleo centrale di questo contributo. I metadati che descrivono la collezione sono stati elaborati attraverso una data analysis con il linguaggio di programmazione R e rappresentati visivamente tramite data visualization – tavole grafiche digitali, interattive o statiche, user-friendly che consentano un approccio intuitivo all'esplorazione del materiale e che suggeriscano spunti per nuove prospettive di ricerca.
Parole chiave: data visualization; data storytelling; patrimonio culturale immateriale; teatro italiano

ABSTRACT (ENGLISH)

From stage to digital: data visualization models for enhancing the Gianni and Cosetta Colla Puppet Theater Archive.

This paper illustrates a section of the "Marionette 3.0" project, achieved within the framework of WP4 of Spoke 2 of PNRR CHANGES, with the aim of communicating and enhancing the intangible theatrical heritage preserved in the Archivio Teatro delle Marionette di Gianni e Cosetta Colla, stored at the Apice Center of the University of Milan. The historical accounts of the Company, embodied in the archival material, have oriented the project towards the creation of digital products aimed at different target groups, such as eBooks for children, storytelling for adults, 3D modeling of puppets, and data visualization of the archive material, which represent the primary focus of this contribution. The metadata describing the collection were processed through a data analysis with the programming language R and visualized by means of data visualization techniques – digital, interactive or static graphic tables, which are user-friendly and allow an intuitive approach to material exploration and suggest new directions for research.
Keywords: data visualization; data storytelling; intangible cultural heritage; Italian theater

1. INTRODUZIONE E STATO DELL'ARTE

Gli archivi, che tradizionalmente svolgono il ruolo di custodi di manufatti, sono oggi chiamati a gestire anche ingenti quantità di dati. A fronte di ciò, le istituzioni si trovano ad affrontare problemi quali il sovraccarico informativo e la progettazione di sistemi di rappresentazione che consentano un accesso immediato e una navigazione efficace (Wallace, 2023). Lo sviluppo di nuove modalità di rappresentazione del materiale d'archivio contribuisce ad ampliare le possibilità di fruizione, passando a nuove forme di accesso. Le interfacce basate esclusivamente sulla ricerca testuale nelle collezioni digitali del patrimonio culturale sono state definite «avare» (Whitelaw, 2015), in quanto celano l'abbondanza e la complessità dei materiali, richiedendo all'utente di formulare query specifiche e limitando così le possibilità di esplorazione. Al contrario, si auspica lo sviluppo di interfacce «generose» (Whitelaw, 2015), capaci di offrire rappresentazioni ricche e stratificate delle collezioni. Tali interfacce invitano alla scoperta, supportano la navigazione attraverso panoramiche orientanti e contesti narrativi, e permettono l'accesso a diversi livelli di dettaglio (York, 2017). Queste modalità non solo si rivelano strumenti preziosi per i ricercatori, ma rendono anche l'archivio più accessibile e coinvolgente per un pubblico ampio e diversificato (Miller, 2019). Tuttavia, progettare un'interfaccia ricca e realmente esplorabile implica confrontarsi con le caratteristiche stesse del contenuto archivistico, spesso eterogeneo, incompleto o fortemente strutturato. I dati raccolti nelle collezioni del patrimonio culturale pongono delle sfide legate ai metodi di visualizzazione e interazione, a causa dell'eterogeneità dei metadati associati a oggetti molto diversi tra loro (libri, testi,

musica, film, manufatti), che richiedono anche una contestualizzazione storica e socioculturale. Questa complessità riflette la natura intrinsecamente «parziale, contingente e suggestiva» dei dati umanistici, spesso mediati da archivi istituzionali e statali, e soggetti a omissioni, volontarie o accidentali (York, 2017).

L'impiego della rappresentazione grafica come strumento euristico del patrimonio culturale rappresenta un campo di indagine recente e in fieri: la prima pubblicazione risale al 2004, ma è solo a partire dal 2010, con la diffusione di grandi repository digitali come Europeana e DPLA, che si registra un interesse crescente per l'integrazione della visualizzazione nei progetti culturali. Una recente mappatura di 186 progetti di Digital Humanities che impiegano metodi di visualizzazione (Battisti, 2024) fornisce una panoramica dello stato dell'arte: la visualizzazione è particolarmente presente nei repository legati a linguaggio e letteratura (74), storia e archeologia (65), mentre è meno utilizzata nei contesti dell'arte (20), delle arti performative (6) e della musica (2). Nel contesto teatrale spicca il *progetto Comédie-Française Registers Project Analytics Tool,* sviluppato da Christopher York, che propone uno strumento online per l'analisi dei registri teatrali digitalizzati della Comédie-Française, dal 1680 al periodo post-rivoluzionario. Progettato per un uso sia esperto che amatoriale, questo strumento permette, ad esempio, a uno studente universitario di esplorare i registri della Comédie-Française mentre scrive un articolo di storia del teatro. La visualizzazione si propone non solo di facilitare la consultazione, ma anche di restituire agli utenti la complessità delle collezioni (Whitelaw, 2015) offrendo una panoramica generale e, allo stesso tempo, la possibilità di approfondire singoli aspetti, superando i limiti della sola ricerca testuale (Miller, 2019) e favorendo l'emergere di nuove domande di ricerca (York, 2019).

Tendenzialmente gli utenti possono essere suddivisi in due gruppi principali: casuali ed esperti (Windhager et al., 2019); anche se l'impressione è che gli archivi raccolgano un pubblico composto principalmente da studiosi specializzati su uno specifico campo o con vocazione interdisciplinare. Nonostante le esigenze differenti, entrambi i gruppi mostrano interesse per le attività di esplorazione e ricercano una panoramica completa dei fondi archivistici. Le interfacce destinate agli utenti esperti tendono a offrire strumenti avanzati per l'analisi dei metadati e per la visualizzazione analitica, mentre quelle pensate per un pubblico più generalista privilegiano immagini e tecniche più immediate, come le word cloud. In generale, le interfacce digitali che integrano metodi di visualizzazione per rappresentare le collezioni del patrimonio culturale sostengono sei principali attività informative: la panoramica e l'orientamento, la ricerca degli oggetti, l'accesso ai dettagli, l'esplorazione verticale o orizzontale, nonché percorsi curati (Windhager et al., 2019). È ampiamente diffusa la presenza di viste multiple per compensare i limiti di singole tecniche di rappresentazione visiva, con una predilezione per visualizzazioni che includono griglie, mappe, network, grafici a barre.

2. METODOLOGIA: ANALISI E VISUALIZZAZIONE DEI DATI

Nel 2008, con l'inserimento dei Pupi siciliani nella Lista Rappresentativa del Patrimonio Culturale Immateriale UNESCO, si assiste al riconoscimento ufficiale del ruolo sociale, identitario e culturale delle marionette (Maselli, 2024). Questo riconoscimento include non solo le arti performative, ma anche «tradizioni ed espressioni orali, ivi compreso il linguaggio, in quanto veicolo del patrimonio culturale immateriale; pratiche sociali, eventi rituali e festivi; artigianato tradizionale»[1]. In questo contesto di rinnovata attenzione verso il teatro di figura, l'Archivio del Teatro delle Marionette di Gianni e Cosetta Colla rappresenta una risorsa fondamentale per la ricostruzione e la valorizzazione della memoria storica di una delle principali compagnie italiane del settore. Costituisce una testimonianza preziosa della storia della Compagnia milanese, raccogliendo copioni, spartiti, bozzetti di marionette e scenografie, corrispondenza, fotografie di prove e spettacoli, locandine, rassegna stampa, programmi di sala, volantini, alcune VHS e una selezione di trenta marionette, tra cui spiccano opere realizzate da Luigi Veronesi[2]. Per rendere più accessibile e trasparente questo patrimonio agli utenti è stato adottato un approccio data-driven, finalizzato sia a restituire una visione d'insieme del materiale conservato, sia a permettere un'esplorazione mirata di aspetti specifici. Che tipo di materiali sono presenti in archivio? In quale quantità? Quali spettacoli hanno segnato la storia della Compagnia? Esistono periodi di maggiore o minore attività? A questi interrogativi si può rispondere tramite un'analisi esplorativa, che favorisce una forma di «ricerca partecipativa» (Miller, 2019), consentendo all'utente di orientarsi in autonomia all'interno dell'archivio,

[1] UNESCO riconosce i Pupi siciliani come Patrimonio Culturale Immateriale: https://unesco.cultura.gov.it/projects/opera-dei-pupi-sicilian-puppet-theatre/ (cons. 05/01/2025)

[2] Archivio Teatro delle Marionette di Gianni e Cosetta Colla, Centro Apice: https://www.apice.unimi.it/collezioni/archivio-teatro-delle-marionette-di-gianni-e-cosetta-colla/

approfondendo liberamente i temi di suo interesse. L'approccio adottato riprende il motto dello studio di design Stamen Studio, «*show everything*» (Hinton & Whitelaw, 2010), che mira a visualizzare la struttura alla base dei dati e a far emergere le informazioni. Sebbene la digitalizzazione degli archivi sia talvolta criticata per il rischio di decontestualizzazione e semplificazione (Sternfeld, 2011), l'impiego della data visualization può contribuire a una lettura più articolata e stratificata dell'archivio stesso (Whitelaw, 2009). Il lavoro ha preso avvio con l'estrazione e organizzazione dei metadati descrittivi dal sistema gestionale del centro APICE. I dati sono stati esportati in formato HTML tabellare e rielaborati in un file Excel strutturato secondo i livelli previsti dal modello ISAD(G), lo standard internazionale di descrizione archivistica, comprendente i livelli archivistici e le informazioni descrittive. Ogni unità archivistica o documento iconografico fa parte di una serie, talvolta di una sottoserie, occasionalmente di un fascicolo o un sottofascicolo, e per ciascuna di esse viene indicato il contenuto e la datazione relativa all'anno di redazione del documento, qualora presente. L'analisi preliminare è stata condotta con il linguaggio di programmazione R, sfruttandone la versatilità nella combinazione di analisi esplorativa e prototipazione di visualizzazioni. Successivamente, i risultati sono stati trasportati in ambienti web-oriented, sviluppando un sito web tramite Sveltekit[3], un framework JavaScript open source altamente performante, in combinazione con D3.js, una libreria JavaScript specializzata nella creazione di visualizzazioni dati interattive, e la piattaforma online Flourish per integrazioni rapide. Questo approccio consente di creare interfacce data-rich, interattive e al contempo garantire una navigazione fluida. Inizialmente è stato fondamentale dare una forma visiva all'intero archivio e la forma prescelta è stata un *zoomable circle packing* (vd. Fig. 1), una tecnica di visualizzazione che organizza i dati in maniera gerarchica mediante cerchi concentrici di diverse dimensioni, esplorabili in maniera interattiva dall'utente. Seguendo il principio «*overview first, zoom and filter, then details-on-demand*» (Shneiderman, 1996), questa visualizzazione consente agli utenti di passare dal generale al particolare, individuando ogni elemento e la sua relazione all'interno della struttura archivistica. Si tratta di un'alternativa visiva all'alberatura testuale multilivello, spesso dispersiva e disorientante per i non addetti ai lavori, che restituisce un'immagine coerente e leggibile dell'intero archivio.

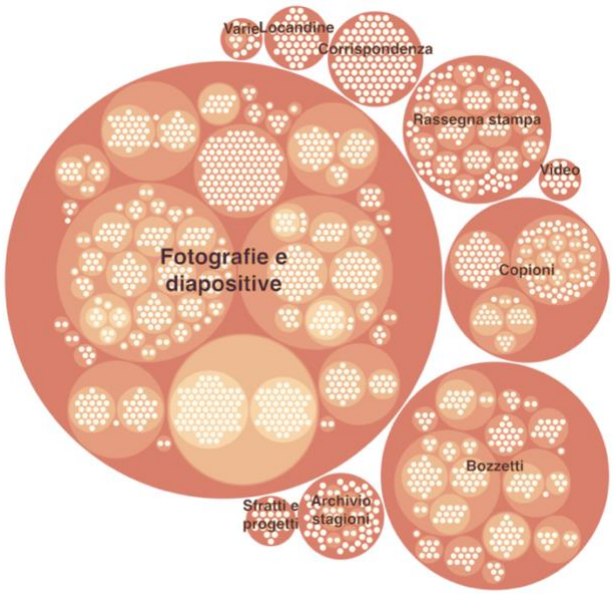


Figura 1. Zoomable circle packing archivio Colla

Successivamente, sono stati realizzati dei grafici a barre per rendere conto dell'analisi quantitativa delle entità conservate, dalla quale è emersa una predominanza assoluta di fotografie, diapositive e bozzetti, in netto contrasto con la scarsità di materiale video. In seguito, è stata integrata la dimensione cronologica, che ha richiesto un'attenzione particolare, data la frequente assenza o incertezza delle date: molte unità archivistiche risultano infatti prive di datazione (s.d.) o accompagnate da stime generiche (ad es. decenni). Tale ambiguità ha rappresentato una criticità sia nella standardizzazione dei dati che nella scelta delle metafore visive, rendendo necessario un approccio cauto nella loro raffigurazione. Per rappresentare

[3] Sveltekit: https://svelte.dev/

efficacemente questa ambiguità, si è optato per un grafico a barre in pila organizzato per decenni e segmentato per serie documentarie (vd. Fig. 2), analogo al *Decade Summary* del progetto Australian Prints and Printmaking (Whitelaw, 2015). I documenti privi di data sono stati inclusi in una categoria separata, evitando l'esclusione silenziosa di una parte significativa dell'archivio. Inoltre, grafici a torta mostrano la proporzione tra materiali datati e non (vd. Fig. 3), rendendo immediatamente visibile l'affidabilità parziale delle conclusioni ricavabili dai dati. Ad esempio, l'analisi mostra una forte concentrazione documentale negli anni Ottanta del Novecento, suggerendo un'intensa attività della Compagnia Colla in quel periodo. Pur trattandosi di un'inferenza plausibile e coerente con la storia della compagnia, è fondamentale accompagnare ogni deduzione con una riflessione critica: i dati visualizzati devono stimolare ipotesi, non sostituire il confronto diretto con le fonti primarie.

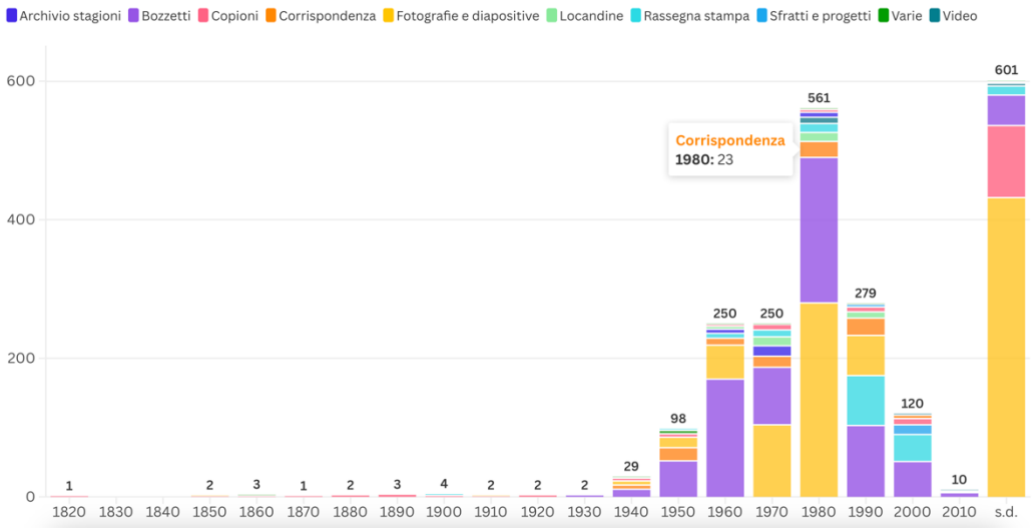


Figura 2. I documenti dell'archivio Colla in ordine cronologico

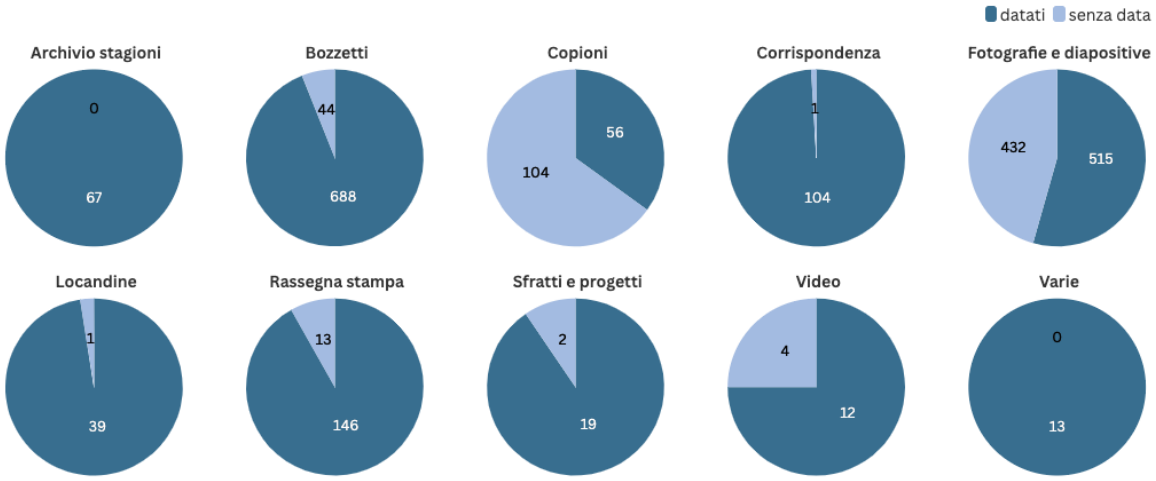


Figura 3. I documenti datati e non dell'archivio Colla

Per accompagnare l'utente nell'esplorazione dell'archivio si è deciso di adottare un approccio coinvolgente e narrativo, attraverso data story tematiche. Sebbene la data visualization sia efficace nel trasmettere grandi quantità di informazioni complesse in modo intuitivo, essa può risultare ancora più potente se affiancata da narrazioni che contestualizzino i dati e guidino l'utente verso nuove scoperte (Kosara & Mackinlay, 2013). Tuttavia, questo approccio sembra essere ancora poco diffuso: secondo la mappatura di 186 progetti di Digital Humanities (Battisti, 2024), solo 43 adottano una componente narrativa, prevalentemente in ambito storico. Quando si combina narrazione e visualizzazione è essenziale trovare un equilibrio tra elementi «*driven by the author*» e quelli «*driven by the reader*» (Segel & Heer, 2010): una

narrazione efficace non impone un'unica lettura, ma offre un percorso esplorativo aperto. Dal punto di vista grafico è fondamentale integrare il contenuto narrativo con il contenuto visivo senza che nessuno dei due sovrasti l'altro. Un approccio particolarmente efficace è stato individuato nello «*scrollytelling*» (Seyser & Zeiller, 2018) – dall'unione di *storytelling* e *scrolling* – che combina al racconto visuale il gesto familiare dello scroll per navigare all'interno di una pagina web. Man mano che l'utente scorre la pagina verso il basso, gli elementi visivi e testuali si rivelano gradualmente e si trasformano, garantendo un'esperienza narrativa fluida e accessibile (vd. Fig. 4). Questo approccio consente di semplificare la complessità, mantenendo alta la densità informativa e migliorando l'accessibilità generale dei contenuti digitali.

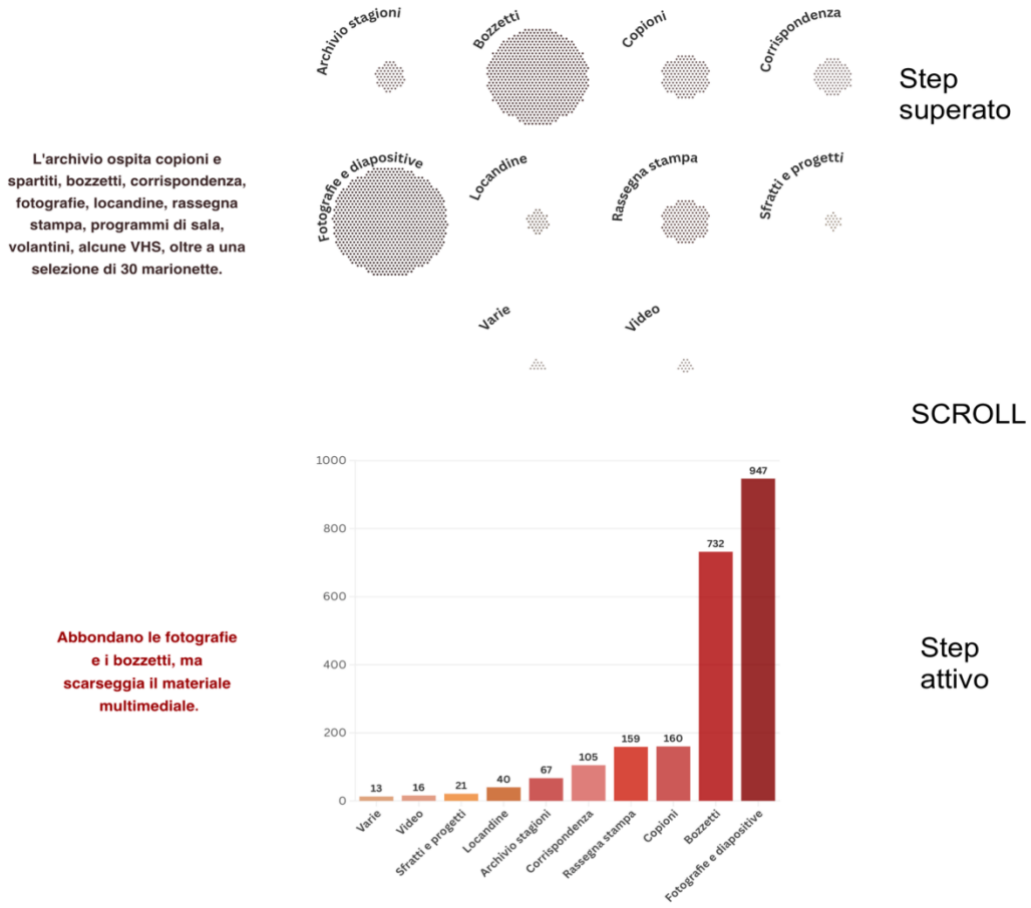


Figura 4. Segmento di data story in formato scrollytelling

3. CRITICITÀ E PROSPETTIVE FUTURE

Lo studio condotto sull'Archivio del Teatro delle Marionette di Gianni e Cosetta Colla ha messo in luce come le collezioni archivistiche relative al patrimonio culturale immateriale rappresentino uno spazio di ricerca particolarmente stimolante, ma anche complesso, per la data visualization. Una delle principali criticità emerse è legata alla natura stessa dei dati archivistici: estremamente eterogenei, con descrittori numerosi ma non sempre coerenti o applicabili a tutte le unità, e con terminologie talvolta inconsistenti. La carenza o incompletezza delle informazioni, in particolare per quanto riguarda la datazione, ha reso difficoltosa l'automazione dei processi di normalizzazione e ha richiesto interventi manuali, oltre a una selezione accurata delle tecniche di visualizzazione. L'obiettivo era quello di evitare che le lacune nei dati restituissero una visione parziale o, peggio, potenzialmente fuorviante dell'archivio. D'altra parte, il potenziale per sviluppare nuove rappresentazioni resta ampio: in futuro sarebbe auspicabile approfondire l'analisi grafica dei generi teatrali, delle connessioni tra personaggi, nonché delle figure coinvolte nella Compagnia, attraverso la costruzione di network relazionali tra persone, ruoli e concetti. Questo tipo di esplorazione presuppone tuttavia un ampliamento e un arricchimento dei metadati attualmente disponibili. Ulteriori sviluppi potrebbero includere analisi computazionali delle diverse edizioni dei copioni, con l'obiettivo di indagare le evoluzioni nel tempo dei testi, e potenzialmente anche delle musiche. In questa

prospettiva la data visualization applicata al patrimonio culturale non solo facilita l'esplorazione e la scoperta delle collezioni (Whitelaw, 2015), ma consente di rilevare tendenze, ricorrenze e anomalie che difficilmente emergerebbero da una semplice consultazione tabellare (York, 2017). Promuove la curiosità, incoraggia la formulazione di nuove domande (Miller, 2019), e si configura come un vero e proprio oggetto epistemico, capace di suggerire inedite interpretazioni dei materiali, contribuendo all'avanzamento della ricerca e della conoscenza. Resta tuttavia essenziale mantenere uno sguardo critico: le visualizzazioni non devono essere intese come strumenti conclusivi, bensì come supporti per formulare ipotesi, da verificare e approfondire attraverso un confronto diretto con le fonti. Solo così sarà possibile promuovere approcci di ricerca sempre più consapevoli e interdisciplinari.

RINGRAZIAMENTI

Il progetto di valorizzazione e ricerca "Marionette 3.0" è finanziato dal progetto 0000020 CHANGES - CUP G53C22000430006, Piano Nazionale di Ripresa e Resilienza (PNRR), Missione 4 (Istruzione e ricerca), Componente 2 (Dalla Ricerca all'Impresa), Investimento 1.3 (Partenariati Estesi), Tematica 5 (Cultura umanistica e patrimonio culturale come laboratori di innovazione e creatività). Si ringraziano i coordinatori del WP4 dello Spoke 2 del PNRR CHANGES Francesco Tissoni e Maddalena Mazzocut-Mis; Vincenzo Pernice, responsabile del progetto "Marionette 3.0"; il Centro Apice e il Teatro Colla.

BIBLIOGRAFIA

Battisti, T. (2024). Classification of web-based Digital Humanities projects leveraging information visualisation techniques. Zenodo. DOI:10.5281/zenodo.14192758

Hinton, S. L., & Whitelaw, M. (2010). Exploring the digital commons: An approach to the visualisation of large heritage datasets. EVA London. DOI:10.14236/ewic/EVA2010.10.

Kosara, R., & Mackinlay, J. (2013). Storytelling: The next step for visualization. Computing Research and Education, 46(5), 44–50. DOI:10.1109/MC.2013.36.

Maselli, V. (2024). Puppets' Tales. New Design Perspectives for a Multimedia Archive of a Humanity's Intangible Heritage. Zanella, F., et al. Multidisciplinary Aspects of Design. Design! OPEN 2022. Springer Series in Design and Innovation, 37, Springer, Cham. DOI:10.1007/978-3-031-49811-4_7.

Miller, A. (2019). Data Visualization as Participatory Research: A Model for Digital Collections to Inspire User-Driven Research. Journal of Web Librarianship, 13(2), 127–155. DOI:10.1080/19322909.2019.1586617

Segel, E., & Heer, J. (2010). Narrative visualization: Telling stories with data. IEEE Transactions on Visualization and Computer Graphics, 16(6), 1139–1148. DOI:10.1109/TVCG.2010.179.

Seyser, D., & Zeiller, M. (2018). Scrollytelling: An analysis of visual storytelling in online journalism. Advances in Data Science and Information Engineering, 401–406. DOI:10.1109/iV.2018.00075.

Shneiderman, B. (1996). The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. Proceedings of the 1996 IEEE Symposium on Visual Languages, 336–343. DOI: 10.1109/VL.1996.545307

Sternfeld, J. (2011). Archival Theory and Digital Historiography: Selection, Search, and Metadata as Archival Processes for Assessing Historical Contextualization. The American Archivist, 74(2), 544–575. DOI:10.17723/aarc.74.2.644851p6gmg432h0.

Wallace, D. A. (2023). SNAFU: Archives are everywhere and vulnerable. Conference session, University of Michigan & University of Witwatersrand - Mellon Foundation Workshop, Magaliesburg, South Africa.

Whitelaw, M. (2015). Generous Interfaces for Digital Cultural Collections. Digital Humanities Quarterly, 9(1).

Whitelaw, M. (2009). Visualising archival collections: The Visible Archive project. Archives & Manuscripts, 37, 22–40.

Windhager, F., Federico, P., Schreder, G., Glinka, K., Dörk, M., Miksch, S., & Mayr, E. (2019). Visualization of cultural heritage collection data: State of the art and future challenges. IEEE Transactions on Visualization and Computer Graphics, 25, 2311–2330. DOI:10.1109/TVCG.2018.2830759

York, C. (2017). Exploratory Data Analysis for the Digital Humanities: The Comédie-Française Registers Project Analytics Tool. English Studies, 98(5), 459–482. DOI: 10.1080/0013838X.2017.1332024

Metodologie computazionali per l'organizzazione di archivi nati digitalmente

Mariangela Giglio
[1] Università di Bologna, Italia – mariangela.giglio2@unibo.it

ABSTRACT (ITALIANO)
Il progetto si colloca nell'ambito della digital forensics e dell'archivistica, affrontando la complessità dell'organizzazione e attraversamento di archivi born digital attraverso un caso di studio specifico: l'analisi comparativa di 58 copie forensi derivate da floppy disk d'autore provenienti dall'Archivio Franco Fortini. L'obiettivo primario è l'elaborazione di una metodologia replicabile di organizzazione archivistica, allo scopo di fornire una classificazione deterministica e verificabile di materiali nativi digitali. A tal fine si propone un approccio sistematico basato sull'hashing dei materiali testuali contenuti nei floppy come indizio di vicinanza tra i dati. Tali hash diventano la base di una matrice binaria che permette di evidenziare le relazioni gerarchiche tra le diverse cartelle e, di conseguenza, di sistematizzare l'archivio e la definizione dei rapporti tra i diversi materiali.
Parole chiave: archivi born digital; clusterizzazione; preservazione digitale; organizzazione archivistica

ABSTRACT (ENGLISH)
Computational Methodologies for the Organization of Born-Digital Archives.
This project lies at the intersection of digital forensics and archival science, addressing the complexity of organizing and analyzing born digital archives through a specific case study: the comparative analysis of 58 forensic copies derived from the floppy disks of the Franco Fortini Archive. The primary objective is the development of a replicable methodology for archival organization, aiming to provide a deterministic and verifiable classification of born digital materials. The project adopts a systematic approach based on hashing the textual materials within the archive as an indicator of proximity or distance between various files. From this, a binary matrix was developed to highlight hierarchical relationships between different folders and, consequently, to systematize the archive and define the relationships between its various materials**.**
Keywords: born digital archives; clustering; digital preservation; archival organization

1. INTRODUZIONE

Il progetto si propone di esplorare e sistematizzare i materiali digitali dell'Archivio Franco Fortini, un archivio ibrido che contiene, in aggiunta a documenti cartacei, anche floppy disk oggi conservati presso la Biblioteca di Area Umanistica dell'Università di Siena. Di questi supporti, molte sono copie di backup desunte dall'ultimo hard disk dell'autore, copie trasposte successivamente in immagini forensi. Si tratta di un caso esemplare di complessità archivistica: sono presenti, infatti, almeno tre gruppi di copie di backup, realizzati in momenti diversi da operatori distinti, unitamente a materiali non d'autore e a una serie di copie parziali effettuate più recentemente da studiosi del Centro Fortini (Marrucci & Tinacci, 2005). L'obiettivo di questo studio è duplice: sviluppare una metodologia replicabile per la gestione e la classificazione di archivi born digital complessi e, al contempo, applicare tale metodologia al caso specifico dell'Archivio Fortini, particolarmente emblematico per la sua complessità e per la stratificazione dei materiali conservati.
La gestione di archivi born digital, come è noto, presenta diverse complessità (Carbé, 2023; Carroll et al., 2011; Digital Preservation Coalition, 2015; Guercio, 2013; Marrucci & Tinacci, 2005), ulteriormente aggravate dalla limitata attenzione che si è data in passato alla corretta preservazione di questa tipologia di materiale, destinata a rapida obsolescenza. Il caso dei floppy conservati nel fondo Fortini non è privo di problematiche: manca una numerazione d'archivio e un ordine sistematico dei materiali, i gruppi di floppy non sono chiaramente distinti e le etichettature sono spesso incoerenti o del tutto assenti. L'archivio inoltre è caratterizzato da ridondanze e difformità strutturali di vario genere. Si è reso dunque necessario stabilire

in via preliminare un ordinamento adeguato, che non può prescindere dall'individuazione dei singoli gruppi di backup.
In questa direzione si è elaborato uno script che, basandosi sull'hashing dei contenuti testuali, possa garantire una classificazione dei materiali deterministica e replicabile. Questo approccio mira ad affrontare in maniera strutturata la complessità intrinseca dell'archivio, automatizzando le operazioni necessarie alla codifica delle relazioni tra i file. I risultati preliminari, sebbene per ora applicati a un singolo archivio, appaiono promettenti e lasciano intravedere una possibile estensione della metodologia ad altri archivi.[1] In questo senso il lavoro vorrebbe proporre una traccia metodologica che fornisca un modello replicabile per la classificazione e l'organizzazione di archivi born digital, offrendo uno strumento di facile uso per contesti archivistici analoghi.

2. METODOLOGIA

Nella gestione degli archivi born digital, le sfide derivate dalla loro complessità intrinseca e dalla varietà dei formati dei dati richiedono un approccio metodico e sistematico. A tale scopo, è stato sviluppato uno script Python dotato di un'interfaccia grafica intuitiva, concepito per facilitare l'adozione dello strumento anche da parte di utenti non specialisti.[2] Il design dello script privilegia la modularità, offrendo agli utenti la possibilità di adattare l'utilizzo dello strumento alle esigenze specifiche del proprio ambito archivistico. Questa struttura modulare permette di isolare e gestire individualmente ciascuna fase del processo, migliorando così la manutenibilità e la scalabilità del sistema.
L'architettura dello script si articola in sei fasi principali, ognuna corrispondente a una funzione chiave, progettata per trattare le problematiche legate alla complessità e alla ridondanza dei materiali. Ogni fase è concepita come un passaggio deterministico nella classificazione e nell'analisi comparativa dei contenuti, assicurando trasparenza e replicabilità dell'intero processo.

3. LO SCRIPT: FUNZIONI CHIAVE

Fase preliminare: preparazione dell'ambiente di lavoro

Il lavoro preliminare consiste nella configurazione di un ambiente di lavoro separato dai file originali, un passaggio indispensabile per assicurare una manipolazione sicura e controllata dei dati. In questa fase, i documenti originali vengono trasferiti in una nuova directory di lavoro denominata Working Directory (WD), che replica esattamente la struttura originaria dei dati, mantenendo inalterati i metadati. Nel contesto dell'archivistica digitale è imperativo lavorare su copie dei file originali per mitigare il rischio di danneggiamenti o alterazioni accidentali. Questa metodologia, ampiamente riconosciuta e attestata (Chassanoff et al., 2016; Ries, 2022), è diventata una prassi per salvaguardare l'integrità del materiale originale durante l'analisi e la manipolazione dei dati. L'utilizzo di copie assicura infatti che eventuali errori di elaborazione o fenomeni di corruzione del materiale incidano unicamente sulle versioni duplicate, preservando così l'integrità del documento originario (Carroll et al., 2011).

Fase 1: Estrazione di file legacy

La fase iniziale dello script è preminentemente volta all'estrazione dei contenuti testuali dai documenti del corpus, operazione imprescindibile per archivi contenenti file in formati legacy e/o materiali non rilevanti ai fini di analisi, quali file di sistema, residui di backup e metadati vari. La presenza di file legacy con estensioni non standard rappresenta una sfida comune negli archivi ibridi o born digital (Light, 2010). Per affrontare questa problematica – limitatamente ai file testuali – è stato adottato l'uso di LibreOffice, software che incorpora avanzate librerie dedicate alla gestione di file testuali obsoleti.[3]
Al fine di garantire un'esclusione mirata e accurata, sono stati automaticamente omessi elementi considerati non essenziali, quali file temporanei e nascosti (e.g., .DS_Store e desktop.ini), nonché directory di sistema (e.g., resource.frk e aree non allocate), consentendo inoltre all'utente l'aggiunta di filtri personalizzati per l'esclusione. I file ritenuti pertinenti sono stati successivamente convertiti in formato

[1] Si pensi, a titolo d'esempio, al fondo di Francesco Pecoraro conservato al Centro Manoscritti di Pavia, al computer del filosofo e matematico Imre Toth conservato presso la Biblioteca di Area Umanistica di Siena o, ancora, ai materiali digitali conservati all'Archivio Bonsanti del Gabinetto Vieusseux (Carbé, 2023, p. 41,92).

[2] Lo script (in forma anonimizzata) è disponibile su Github.

[3] Per ulteriori informazioni sulle librerie incluse in LibreOffice è possibile fare riferimento alla pagina Licensing and Legal information del software (cons. 01/01/2025).

.odt mediante l'interfaccia a riga di comando (Command Line Interface - CLI) di LibreOffice. Durante questa fase ci si mantenuta inalterata la struttura gerarchica originale dell'archivio, una pratica che assicura che ciascun elemento del corpus possa essere precisamente ricollocato nella sua posizione originaria, facilitando l'analisi successiva e conservando l'integrità contestuale dei dati estratti. L'estrazione dei materiali testuali in formati standard garantisce che la fase di hashing successiva possa avvenire senza essere influenzata da fattori non significativi quali i metadati dei file o contenuti non rilevanti.

Fase 2: Hashing

La prosecuzione dello script introduce una fase di hashing che costituisce un momento cruciale della clusterizzazione, influenzando direttamente l'efficacia delle fasi processuali successive. L'hashing è un processo computazionale che trasforma un insieme di dati di dimensioni variabili in una rappresentazione univoca, denominata hash, una sorta di "impronta digitale" di ciascun file. Questa operazione permettere confronti rapidi tra file, ed è utile sia per la verifica dell'integrità dei dati sia per la loro autenticazione (Chi & Zhu, 2018; Roussev, 2009).

Al fine di ottenere una rappresentazione fedele dei documenti, la tecnica implementata esclude metadati, percorsi dei file e altre variabili esterne che potrebbero alterare l'hash generato, focalizzandosi sul solo contenuto testuale estratto dai file.

Utilizzando l'algoritmo SHA-2564[4] ogni documento in formato .odt è stato processato per ottenere un hash univoco, garantendo così una precisa identificazione dei file anche in presenza di minime variazioni. Questi hash sono stati successivamente catalogati in un dizionario in cui ogni chiave rappresenta una cartella, mentre i valori sono gli elenchi degli hash dei file contenuti nella directory stessa e nelle sue sottodirectory. Questo metodo consente di ridurre la complessità del dataset, focalizzandosi sulla distribuzione dei contenuti testuali e facilitando la comparazione dei file all'interno dell'archivio.

Fase 3: Creazione di una matrice binaria hash/directory

La terza fase del processo consiste in una conversione del dizionario degli hash in una matrice binaria che stabilisce una correlazione tra ogni hash univoco alle directory che lo contengono, offrendo così una rappresentazione formale della presenza o assenza di ciascun hash unico nelle diverse directory. Tale struttura costituisce una prima schematizzazione che delinea le relazioni tra i contenuti e le loro posizioni all'interno dell'archivio.

Nella matrice, ogni riga è associata a un hash distinto, mentre ciascuna colonna corrisponde a una determinata cartella. Le celle sono riempite con valori booleani, zero o uno, indicanti rispettivamente l'assenza o la presenza dell'hash specifico nella directory corrispondente. Questa rappresentazione strutturale facilita l'identificazione rapida delle relazioni tra i file distribuiti nei vari supporti di archiviazione, permettendo di evidenziare le copie identiche attraverso colonne perfettamente allineate e di rilevare le divergenze, parziali o totali, mediante colonne con pattern dissonanti.

Il numero di righe nella matrice è pari alla quantità di hash unici prodotti, mentre il numero di colonne riflette il numero di directory di livello superiore esaminate.

	Cartella α	***Cartella β***	***Cartella γ***	***Cartella δ***	***Cartella ε***
Hash 1	1	0	1	1	0
Hash 2	0	0	0	1	1
Hash 3	1	1	0	0	0
Hash 4	1	0	1	1	1
Hash 5	1	0	1	1	0

Tab.1 Esempio di output in matrice binaria

Fase 4: Matrice di similarità

Nella fase successiva del processo, basandosi sulla matrice binaria sviluppata precedentemente, si procede alla configurazione di una matrice di similarità che rappresenti formalmente il grado di affinità tra le varie

[4] SHA-256 è una funzione di hash crittografica ampiamente impiegata per garantire l'integrità, l'autenticità e la sicurezza dei dati in varie applicazioni, incluse le firme digitali, i codici di autenticazione dei messaggi e la tecnologia blockchain (Devi & Jayasri, 2023; Gilbert & Handschuh, 2004).

directory. La matrice, di natura simmetrica, attribuisce un valore diagonale costante pari a 1, segnalando l'identità di ciascuna directory rispetto a sé stessa. Le presenze o le assenze di hash univoci, codificati precedentemente nella matrice binaria, vengono così trasformati in valori numerici che riflettono la similarità tra i contenuti delle directory. All'interno di questa matrice, ogni riga corrisponde a un hash distinto, mentre ogni colonna rappresenta una specifica cartella. Le celle, popolate da valori booleani (0 o 1), indicano rispettivamente l'assenza o la presenza di un determinato hash nella cartella corrispondente. Questa configurazione strutturale agevola l'identificazione immediata delle relazioni tra i supporti di memorizzazione, permettendo di evidenziare sia le copie esatte – attraverso colonne identiche – sia le divergenze parziali o totali manifestate attraverso colonne con pattern divergenti.
Il principio su cui si basa la matrice di similarità deriva dal confronto tra le colonne della matrice binaria, dove ogni colonna simbolizza una directory e ogni riga un hash unico. La misura di similarità tra due colonne viene calcolata attraverso l'identificazione delle discrepanze binarie, ovvero le righe in cui un hash appare in una colonna ma è assente nell'altra. Questa analisi produce risultati che variano su una scala di similarità: un valore di 1 indica una sovrapposizione totale, mentre un valore di 0 denota una divergenza completa.
La formula utilizzata per calcolare la similarità è:

$$S = 1 - \frac{d}{nmax}$$

Form.1 Formula per calcolare la similarità

Dove S rappresenta il valore di similarità, d indica il numero di differenze binarie tra due colonne, e $nmax$ è il numero massimo di hash presenti in una singola cartella.
Questa formula viene implementata nella funzione per attribuire un punteggio di similarità variabile da 0 a 1 a ogni coppia di cartelle esaminata.
Consideriamo adesso il caso di due cartelle, indicate come *α* e *β*, per illustrare l'applicazione pratica della formula di similarità.

$$S\alpha\beta = 1 - \frac{d\alpha\beta}{nmax}$$

Form.2 Formula per calcolare la similarità tra le cartelle α e β

Il valore di similarità tra α e β può dare diversi risultati a seconda della effettiva somiglianza tra le due cartelle. A titolo di esempio:

- $S\alpha\beta=1$ indica che α e β sono cartelle identiche
- $S\alpha\beta=0$ indica che α e β sono cartelle totalmente diverse (non condividono nessun file)
- $0.5< S\alpha\beta<1$ indica che α e β presentano delle sovrapposizioni parziali
- $S\alpha\beta=0.01$ indica che α e β sono profondamente diversi ma condividono alcuni file
- $S\alpha\beta=0.8$ indica la presenza di numerose sovrapposizioni

Nel contesto dell'Archivio Franco Fortini, questa distinzione è particolarmente rilevante: ogni gruppo di backup è composto da floppy che contengono dati differenti, ma la presenza di file identici tra gruppi diversi produce corrispondenze puntuali all'interno della matrice. Di conseguenza, punteggi pari a 1 non indicano necessariamente appartenenza a uno stesso gruppo di backup, bensì replicazione esatta tra floppy gemelli. Al contrario, la struttura globale della matrice di distanza, con cluster di somiglianza parziale, consente di identificare i gruppi di backup e comprendere le relazioni tra essi. Questa fase rappresenta il punto di arrivo del processo analitico, traducendo i dati grezzi in una visione strutturata e quantitativa delle relazioni tra i materiali.

Fase 5: Visualizzazione dei risultati
La fase conclusiva dello script si avvale della matrice di similarità elaborata nelle fasi antecedenti per generare visualizzazioni che delineano chiaramente le relazioni tra i documenti. Per implementare il clustering gerarchico, essenziale per queste visualizzazioni, la matrice di similarità è trasformata in una matrice di distanza, definendo la distanza come $1-S$.[5] Questa trasformazione è indispensabile perché

[5] Similarità, vedi Formula 1.

numerose tecniche di clustering operano più efficacemente con le metriche di distanza piuttosto che di similarità. Si è tuttavia scelto di conservare la matrice di similarità come output principale, in quanto offre una rappresentazione intuitiva e facilmente accessibile, sia per l'interpretazione umana sia per l'elaborazione informatica, e può essere convertita all'occorrenza per analisi ulteriori.
I dendrogrammi vengono generati impiegando la funzione "linkage" di SciPy, partendo dalla matrice di distanza condensata, derivata dalla matrice di distanza quadrata[6] elaborata in precedenza. Metodo di linkage scelto per l'analisi è quello di Ward, noto per la sua efficacia nel minimizzare la varianza all'interno dei cluster. Questo approccio assicura un raggruppamento dei dati coerente e informativo, come evidenziato da Großwendt et al. (2019) e Strauss & Von Maltitz (2017). Per consentire una maggiore flessibilità è stata comunque lasciata all'utente la possibilità di personalizzare metodo di linkage e distanza.

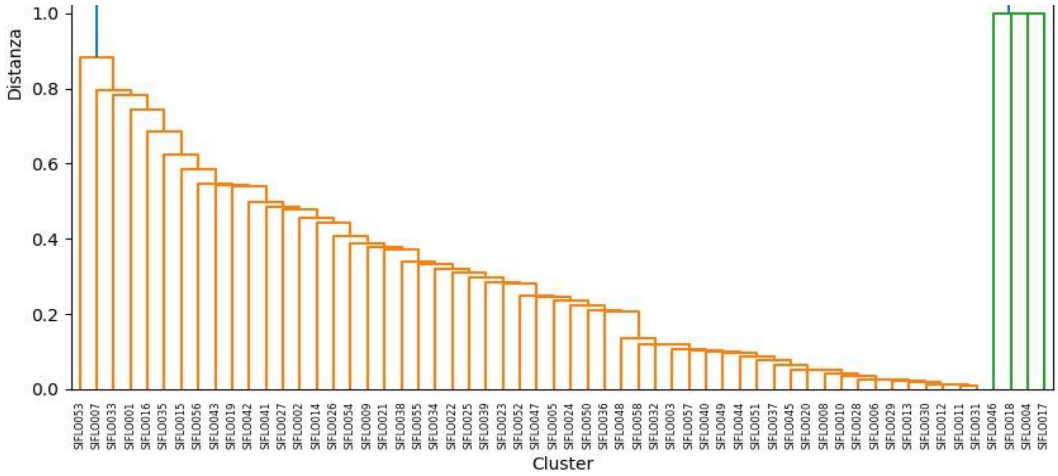


Fig.1 Il dendrogramma rappresentante il raggruppamento dei materiali

4. VALIDAZIONE DEI RISULTATI

Per valutare l'attendibilità dei risultati ottenuti tramite la procedura di clusterizzazione automatizzata, è stata condotta un'operazione di validazione manuale che ha assunto il ruolo di *ground truth* dell'intero processo. Tale attività ha previsto un'analisi approfondita dei contenuti testuali dei file, della struttura delle directory presenti nei supporti e dei metadati espliciti (come etichette, nomi di cartelle e formati di file), con l'obiettivo di individuare somiglianze ricorrenti e relazioni funzionali all'interno del corpus. A questo si è affiancata un'analisi forense condotta direttamente sulle immagini dei floppy disk, con particolare attenzione ai *timestamp* di creazione e modifica, ai nomi utente e ad altre tracce residue conservate nei file di sistema. Tali metadati nascosti, spesso non rilevabili tramite semplice ispezione documentale, hanno permesso di formulare ipotesi di raggruppamento indipendenti dal contenuto, fornendo ulteriori indizi di prossimità tra i supporti.
Il confronto tra la clusterizzazione *human validated* e i risultati prodotti dallo script ha evidenziato una sostanziale coerenza tra i due modelli. I principali gruppi di backup sono stati riconosciuti correttamente dall'algoritmo, e i dendrogrammi generati riflettono con buona approssimazione le relazioni gerarchiche ipotizzate manualmente. Le copie identiche e le somiglianze parziali rilevate automaticamente trovano corrispondenza nella struttura fisica e logica degli archivi digitali, a conferma della solidità dell'approccio computazionale. Le discrepanze riscontrate si concentrano su casi ambigui o liminali, in cui la collocazione archivistica risulta influenzata da fattori extratestuali, come la frammentazione di una directory su più supporti o operazioni successive di duplicazione.
Per consolidare ulteriormente la validità del metodo, è stata condotta una validazione quantitativa sulla porzione di corpus clusterizzata manualmente. Le performance della clusterizzazione sono state misurate

[6] La matrice quadrata, usata per rappresentare le distanze tra coppie di punti in una tabella bidimensionale, è caratterizzata dalla simmetria (ad esempio, la distanza tra i punti a e b è uguale a quella tra b e a) e da elementi diagonali nulli. Una versione più compatta è la matrice condensed, che si presenta come un array monodimensionale. Quest'ultima include solo le distanze sopra o sotto la diagonale principale, eliminando ridondanze e valori zero della diagonale, il che riduce significativamente l'uso della memoria e aumenta l'efficienza computazionale, soprattutto per matrici di grandi dimensioni.

attraverso un insieme di metriche standard per la valutazione del clustering non supervisionato, tra cui l'Adjusted Rand Index (ARI), la Jaccard macro-average, e le misure derivate dalla mutual information: Homogeneity, Completeness e V-Measure.[7] I risultati confermano una buona coerenza strutturale tra i due modelli (ARI ≈ 0.69), anche se la frammentazione di alcuni gruppi e la parziale sovrapposizione evidenziata dallo Jaccard score (≈ 0.46) indicano margini di miglioramento.
Le limitazioni principali emerse riguardano la generalizzabilità del metodo a contesti archivistici diversi. L'efficacia della procedura dipende infatti in larga misura dalla qualità e coerenza dei dati disponibili: archivi con file non testuali, strutture di file system danneggiate o contenuti altamente eterogenei possono ridurre l'accuratezza dei risultati. Inoltre, l'uso di tecniche di hashing e tokenizzazione privilegia una logica di somiglianza contenutistica, che può non cogliere appieno l'organizzazione semantica o funzionale di alcuni archivi. Ciò risulta particolarmente evidente in casi in cui directory affini sul piano concettuale differiscano sensibilmente nella forma o nel contenuto. La strategia adottata si dimostra quindi solida in contesti coerenti, ma richiede un affinamento metodologico per gestire anomalie locali, supporti parziali o materiali liminali, e per una eventuale estensione del modello a dataset meno strutturati.

5. CONCLUSIONI

Le clusterizzazioni effettuate dallo script sono state messe a confronto con le ipotesi preliminari di raggruppamento basate sull'osservazione diretta dei file, evidenziando risultati non ancora perfetti ma promettenti e significativi per ulteriori indagini. Le precedenti analisi dei materiali digitali dell'archivio non avevano permesso di identificare tutte le relazioni significative, come ad esempio le somiglianze parziali tra diverse cartelle, che sono invece emerse chiaramente nel raggruppamento elaborato con lo script. L'esame del dendrogramma ha rivelato la presenza di ulteriori copie correlate ai tre gruppi principali di backup (i floppy 046, 018, 004, 017 evidenziati in verde nella Figura 1). Questi raggruppamenti, sebbene individuabili attraverso un'analisi tradizionale, potrebbero non essere riconosciuti senza un esame completo dell'archivio: risalta l'efficacia del raggruppamento in insiemi di tre, suggerito dall'algoritmo, che pare riflettere accuratamente la suddivisione originaria in tre gruppi di backup. Permane, tuttavia, la presenza di alcune clusterizzazioni problematiche o errori di riconoscimento che potrebbe essere emendata con ulteriori affinamenti dell'algoritmo di clustering. Nel caso specifico del fondo Fortini questa interrelazione potrebbe, tuttavia, essere un residuo del processo di copia sequenziale in più floppy di dimensioni diverse, ipotesi che necessita di un'analisi più approfondita per essere supportata.
Mediante un approccio modulare e automatizzato, si è affrontata la complessità di un corpus caratterizzato da ridondanze, formati eterogenei e strutture non uniformi. L'adozione di tecniche computazionali, come l'hashing dei contenuti testuali e la costruzione di matrici binarie e di distanza, ha permesso di tradurre la complessità intrinseca dei dati in una rappresentazione strutturata e quantificabile. Le varie fasi del progetto sono state armonizzate all'interno di una pipeline che ambisce a fornire replicabilità e trasparenza, componenti essenziali per l'applicazione di tali metodologie in contesti archivistici analoghi. Le visualizzazioni grafiche, infine, hanno supportato l'interpretazione dei dati, rendendo leggibili le relazioni interne ai materiali. Il lavoro evidenzia come la combinazione di analisi statistiche automatiche e revisioni manuali possa contribuire alla migliore gestione di archivi born digital complessi. Sebbene il focus sia stato l'Archivio Franco Fortini, la metodologia proposta potrebbe essere estesa ad altre realtà archivistiche, ponendo le basi per un approccio più generalizzabile alla gestione di materiali nativi digitali.

RINGRAZIAMENTI

Desidero esprimere la mia sincera gratitudine alla professoressa Emmanuela Carbé per avermi indicato una preziosa traccia metodologica da seguire.
Ringrazio inoltre Francesco Garosi, esperto di tecnologie digitali, per il supporto e la competenza messi a disposizione nella realizzazione della componente informatica del lavoro.

[7] Le metriche di validazione includono: Adjusted Rand Index (che misura la concordanza tra assegnazioni indipendentemente dai nomi dei cluster), Jaccard macro-average (che valuta la sovrapposizione tra le classi), Homogeneity (quanto ciascun cluster contiene elementi di una sola classe), Completeness (quanto tutti gli elementi di una classe sono nello stesso cluster) e V-Measure (media armonica tra Homogeneity e Completeness).

BIBLIOGRAFIA

Carbé, E. (2023). *Digitale d'autore: Macchine, archivi e letterature*. Firenze University Press. https://doi.org/10.36253/979-12-215-0023-3.

Carroll, L., Farr, E., Hornsby, P., & Ranker, B. (2011). A Comprehensive Approach to Born-Digital Archives. *Archivaria*, *72*, 61–92.

Chassanoff, A., Woods, K., & Lee, C. A. (2016). Digital Preservation Metadata Practice for Disk Image Access. In A. Dappert, R. S. Guenther, & S. Peyrard (A c. Di), *Digital Preservation Metadata for Practitioners* (pp. 99–109). Springer International Publishing. https://doi.org/10.1007/978-3-31943763-7_8

Chi, L., & Zhu, X. (2018). Hashing Techniques: A Survey and Taxonomy. *ACM Computing Surveys*, *50*(1), 1–36. https://doi.org/10.1145/3047307

Devi, S. S., & Jayasri, K. (2023). Esha-256: An Enhanced Secure Cryptographic Hash Algorithm for Information Security. *2023 2nd International Conference on Automation, Computing and Renewable Systems (ICACRS)*, 952–958. https://doi.org/10.1109/ICACRS58579.2023.10405239

Digital Preservation Coalition. (2015). *Digital Preservation Handbook. 2nd edition*. https://www.dpconline.org/handbook

Gilbert, H., & Handschuh, H. (2004). Security Analysis of SHA-256 and Sisters. In M. Matsui & R. J. Zuccherato (A c. Di), *Selected Areas in Cryptography* (Vol. 3006, pp. 175–193). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-24654-1_13

Großwendt, A., Röglin, H., & Schmidt, M. (2019). *Analysis of Ward's Method*. 2939–2957. https://doi.org/10.1137/1.9781611975482.182

Guercio, M. R. (2013). *Conservare il digitale: Principi, metodi e procedure per la conservazione a lungo termine di documenti digitali* (1. ed). Laterza.

Light, M. (2010). *Designing a Born-Digital Archive*. 1.

Marrucci, M., & Tinacci, V. (2005). L'edizione di uno scritto a testimonianza plurima, cartacea e informatica: Un giorno o l'altro di Franco Fortini. *Filologia italiana 2*.

Ries, T. (2022). Digital history and born-digital archives: The importance of forensic methods. *Journal of the British Academy*, *10*, 157–185. https://doi.org/10.5871/jba/010.157

Roussev, V. (2009). Hashing and Data Fingerprinting in Digital Forensics. *IEEE Security & Privacy Magazine*, *7*(2), 49–55. https://doi.org/10.1109/MSP.2009.40

Strauss, T., & Von Maltitz, M. J. (2017). Generalising Ward's Method for Use with Manhattan Distances. *PLOS ONE*, *12*(1), e0168288. https://doi.org/10.1371/journal.pone.0168288

Archivi digitali per la conservazione e valorizzazione del patrimonio culturale: il caso del Santuario della Madonna di Carufo

Caterina Ciccotti[1]
[1] Università degli Studi di Macerata, Italy, c.ciccotti@unimc.it

ABSTRACT
Il presente contributo illustra un progetto in corso volto alla creazione di un archivio digitale multi-tipologico dedicato al Santuario della Madonna di Garufo (Camporotondo di Fiastrone, MC), danneggiato dal sisma del 2016 e oggetto di un complesso intervento di restauro. L'archivio nasce con l'obiettivo di documentare, valorizzare e rendere accessibile una pluralità di materiali eterogenei – storici, tecnici, multimediali – prodotti nel corso del cantiere e della ricerca storico-artistica, secondo principi di interoperabilità, sostenibilità e replicabilità. La metodologia adottata combina l'uso di standard archivistici (ISAD(G), Dublin Core, RiC-CM), linee guida professionali (ANAI, 2023) e tecnologie semantiche (RDF, IIIF, SPARQL), applicate all'interno della piattaforma open source Omeka S. Il progetto propone un modello operativo per la gestione di archivi digitali in contesti post-sismici, integrando documentazione tecnica e memoria culturale e restituendo alla comunità un patrimonio fragile in forma accessibile, strutturata e interrogabile.
Parole chiave: Digital Humanities; archivi digitali; Linked Open Data; preservazione digitale; Web Semantico.

ABSTRACT
Digital Archives for Cultural Heritage Preservation and Valorization: The Case of the Sanctuary of Madonna di Carufo.
This paper presents an ongoing project aimed at creating a multi-type digital archive dedicated to the Sanctuary of Madonna di Garufo (Camporotondo di Fiastrone, Italy), which was severely damaged by the 2016 earthquake and is currently undergoing restoration. The archive is designed to document, preserve, and provide access to a diverse array of historical, technical, and multimedia materials generated during the restoration process and related research. The methodological framework integrates archival standards (ISAD(G), Dublin Core, RiC-CM), professional guidelines (ANAI, 2023), and semantic web technologies (RDF, IIIF, SPARQL), implemented within the open-source platform Omeka S. The project proposes a sustainable and replicable model for managing digital archives in post-disaster contexts, combining technical documentation and cultural memory to make a vulnerable heritage resource available in a structured, accessible, and queryable form.
Keywords: Digital Humanities; digital archives; Linked Open Data; digital preservation; Semantic Web.

1. INTRODUZIONE

Il Santuario della Madonna di Garufo, situato a Camporotondo di Fiastrone (MC), rappresenta un caso emblematico di patrimonio culturale colpito dagli eventi sismici del 2016-2017 e oggetto di un processo complesso di restauro. In questo contesto, il progetto di ricerca attualmente in corso si propone di costruire un archivio digitale multi-tipologico che accompagni e documenti tutte le fasi dell'intervento, restituendo centralità alla memoria del luogo attraverso strumenti propri delle Digital Humanities (Metin & Rognoni, 2024).
Il lavoro nasce nell'ambito della ricerca dottorale presso l'Università di Macerata, in collaborazione con l'associazione Identità Sibillina ETS, e si fonda sull'adozione di standard internazionali per la descrizione archivistica (Dublin Core, ISAD(G), RiC-CM) e su tecnologie semantiche interoperabili (RDF, SPARQL). L'archivio si configura quindi non solo come strumento di conservazione, ma come piattaforma partecipativa di conoscenza e valorizzazione, capace di coinvolgere una pluralità di attori e di rendere accessibili risorse documentarie frammentarie o disperse (Fiorani & Romeo, 2020; Grimoldi, 2017; Salarelli, 2016) in un'ottica di apertura, condivisione e rigenerazione culturale (Balzani, Raco & Suppa, 2019; UNESCO, 2003; Tufarelli, 2019).
La struttura archivistica, che si fonda sulle Linee Guida ANAI 2023 per la gestione di archivi digitali e ibridi negli studi di progettazione (ANAI, 2023), integra strumenti e standard descrittivi per garantire l'interoperabilità (come Dublin Core, ISAD(G), RiC-CM e RDF), tenendo conto delle esigenze di pubblici diversificati.

L’archivio è progettato per ospitare e descrivere risorse eterogenee – fotografie di cantiere, relazioni tecniche, disegni CAD, rilievi 3D, permessi amministrativi e documentazione storica – valorizzandone la funzione conoscitiva e contribuendo a una più consapevole gestione del patrimonio (Fiorani & Romeo, 2020; Grimoldi, 2017). La documentazione viene trattata non solo come supporto operativo, ma come elemento chiave di un patrimonio informativo da rendere accessibile, anche in ottica di rigenerazione sociale e culturale delle comunità colpite (Salarelli, 2016; Balzani, Raco & Suppa, 2019).
Nella fase di analisi e progettazione della piattaforma, è stato effettuato un confronto tra soluzioni open source come Omeka S e CollectiveAccess. La scelta è ricaduta su Omeka S per la sua flessibilità nel modellare vocabolari personalizzati, la possibile integrazione con IIIF e il supporto a ontologie RDF, elementi fondamentali per una valorizzazione semantica delle risorse.
Il caso del Santuario di Garufo permette così di riflettere sul ruolo degli archivi digitali non solo come strumenti di conservazione, ma come veri e propri hub di conoscenza, capaci di attivare nuove pratiche di accesso, riuso e partecipazione (UNESCO, 2003; Tufarelli, 2019), laddove l’integrazione tra metodologie tradizionali e tecnologie digitali dimostra la possibilità di restituire alle comunità un patrimonio documentario fragile e disperso, trasformandolo in risorsa condivisa e attivando un circolo virtuoso di memoria, consapevolezza e innovazione.

2. IL SANTUARIO DELLA MADONNA DI GARUFO

Quello di Garufo rappresenta il caso di studio scelto per l’implementazione di un archivio digitale dedicato alla documentazione, conservazione e valorizzazione del patrimonio culturale danneggiato dal sisma del 2016. L’edificio, risalente al XIII secolo, è stato gravemente compromesso dagli eventi sismici, evidenziando la fragilità del patrimonio materiale e l’urgenza di attivare strategie che affianchino la ricostruzione fisica alla salvaguardia della memoria storica e identitaria.
Oltre al valore storico-artistico, il Santuario è profondamente radicato nella memoria religiosa della comunità locale. Le prime attestazioni documentarie del culto mariano nel sito risalgono almeno al Duecento, come ricordato anche da Angelo Bittarelli, che sottolinea la presenza di una "Chiesa di S. Maria di Carrufa" nelle pertinenze storiche dell’antico centro di Pievefavera, da cui il toponimo avrebbe poi evoluto nelle forme “Grufo”, “Carufo”, “Garufo” (Bittarelli, 2004). Il luogo era frequentato da fedeli non solo del comune di Camporotondo, ma anche delle comunità limitrofe, configurandosi come riferimento condiviso per le pratiche religiose, le processioni e le devozioni popolari, specie durante le festività mariane.



Figura 1 - SANTUARIO DELLA MADONNA DI GARUFO (SANTUARIO) (INTERNO) Fonte: ICCD, Condizioni d’uso: Beni Culturali Standard (BCS).

In un saggio postumo, lo storico e letterario Febo Allevi riflette poi sul valore simbolico del sito, collocandolo in una prospettiva di continuità storica e antropologica del sacro.
Secondo Allevi, il Santuario è un esempio di come i luoghi sacri conservino la loro funzione anche attraverso il mutare delle epoche, trasformando antiche pratiche e simboli in nuove forme religiose, in un

processo di mediazione fra paesaggio fisico ed esperienza spirituale: «la sacralità si manifesta in continuità storica, raccogliendo e trasformando pratiche religiose e simboliche arcaiche in forme cristiane» (Allevi, 2000, p. 5). Tale visione rafforza l'idea di una devozione che si radica nei luoghi, generando identità e memoria collettiva.
La centralità del Santuario, dunque, non si esaurisce nel suo valore architettonico, ma si estende alla sfera culturale e antropologica. Il progetto di restauro, avviato in seguito al terremoto, si è rapidamente trasformato in un laboratorio multidisciplinare, in cui la raccolta e la sistematizzazione della documentazione prodotta durante le fasi di intervento si è rivelata essenziale per una riflessione più ampia sul ruolo della documentazione nei processi di rigenerazione.
È in questo contesto che nasce l'idea di sviluppare un archivio digitale multi-tipologico, in grado di ospitare e descrivere materiali eterogenei secondo criteri archivistici, tecnici e narrativi. Il modello di archivio adottato si inserisce nella categoria degli archivi "di progetto" (Ghezzi, 2023), ovvero sistemi dinamici che accompagnano l'intero ciclo di vita di un bene culturale, integrando fonti storiche, documentazione tecnico-amministrativa, materiali multimediali e risorse divulgative. L'obiettivo è valorizzare il potenziale informativo e partecipativo della documentazione, non solo come supporto tecnico, ma come infrastruttura di memoria condivisa e strumento di rigenerazione culturale.
Nel caso specifico di Garufo, l'archivio è articolato in una struttura logica multilivello (collezione, fascicolo, documento) e comprende documenti storici, fotografie di cantiere, disegni tecnici, rilievi tridimensionali, permessi amministrativi, modelli digitali e materiali educativi. Questa articolazione riflette l'interdisciplinarietà del progetto e la necessità di garantire una descrizione coerente delle risorse, mediante l'utilizzo di standard archivistici (Dublin Core, ISAD(G), RiC-CM) e strumenti digitali interoperabili.
Attraverso la digitalizzazione e la pubblicazione online di questi contenuti, il progetto mira a rendere accessibile il patrimonio documentale e a restituire alle comunità locali la memoria di un luogo centrale nella loro identità culturale e spirituale, ma vuole anche fornire un modello replicabile per altri contesti post-sismici, in cui l'archivio possa fungere da leva per l'accessibilità, la trasparenza e la rigenerazione culturale, riaffermando il valore della documentazione come infrastruttura di memoria condivisa (UNESCO, 2003; Tufarelli, 2019).

3. METODOLOGIA

Il progetto di digitalizzazione e archiviazione del Santuario della Madonna di Garufo si inserisce all'interno di un più ampio quadro di riflessione sull'importanza della documentazione come infrastruttura per la conoscenza (Balzani, Raco & Suppa, 2019; Salarelli, 2016), la conservazione e la valorizzazione del patrimonio culturale in contesti di crisi. L'idea di strutturare un archivio digitale multi-tipologico nasce da un'analisi preliminare del contesto territoriale e del quadro di vulnerabilità in cui si colloca il bene: analizzando l'elaborazione Openpolis dei dati Ispra(Openpolis, 2023), data la vulnerabilità dei beni culturali posti in zona sismica 1 e 2, appartenenti al cratere del 2016, si rende evidente la necessità di progettare sistemi informativi e modelli di gestione documentale che accompagnino e supportino le fasi di restauro, integrando conoscenza tecnica e memoria storica.
La prima fase del lavoro ha previsto la raccolta e selezione del materiale documentario da inserire nell'archivio. Tale corpus include da un lato la documentazione prodotta nel corso del cantiere di restauro del Santuario, dall'altro fonti storiche conservate presso archivi locali e contributi di carattere storico-artistico e devozionale. Per quanto riguarda il primo gruppo, si è seguito il framework proposto dalle Linee Guida ANAI (2023) per la gestione e conservazione degli archivi digitali e ibridi negli studi di progettazione, che fornisce criteri metodologici per organizzare e descrivere la documentazione tecnica (rilievi, relazioni, elaborati grafici, fotografie di cantiere, permessi amministrativi). Per il secondo gruppo, è stato condotto un lavoro di ricerca presso gli archivi locali e sulla letteratura prodotta da studiosi locali come Febo Allevi e Angelo Bittarelli (Allevi, 2000; Bittarelli, 2004), che hanno ricostruito la storia e la devozione legata al Santuario.
Una terza fase riguarda la progettazione dell'archivio e la selezione della piattaforma digitale. L'obiettivo era costruire un modello replicabile e sostenibile, adatto a contesti con risorse limitate e gestibile anche da enti culturali privi di competenze tecniche avanzate in base a criteri di interoperabilità, flessibilità semantica, sostenibilità, accessibilità e facilità di configurazione (Popović, Škorić & Rujević, 2020; Ravikumar & Hazarika, 2020; Alcaraz-Martínez, 2022). Come approfondito nella sezione successiva, la scelta è ricaduta su Omeka S per la sua capacità di coniugare semplicità d'uso e compatibilità con modelli descrittivi RDF-based e standard come Dublin Core, ISAD(G) e Records in Contexts.

Infine, è stata definita la struttura concettuale e archivistica del progetto, articolata secondo una logica multi-tipologica e multilivello. L'intero progetto si configura come archivio digitale, al cui interno ogni fondo (es. "Archivio storico Comune Camporotondo") è modellato come item, mentre le serie documentarie sono rappresentate da item set (es. "Riformanze"), articolati in ulteriori sottoserie o rubriche (es. "Rubrica 1 – Acque e Strade"), che contengono i singoli documenti (item) descritti e digitalizzati. La struttura rispecchia modelli avanzati di archivi ibridi (Fiorani & Romeo, 2020), capaci di integrare componenti storiche, tecniche e multimediali, con una particolare attenzione all'interoperabilità semantica e alla pubblicazione sul Web attraverso tecnologie RDF e IIIF.
Attraverso questa impostazione metodologica, il progetto mira non solo alla conservazione e accessibilità dei contenuti, ma anche a promuovere una riflessione più ampia sul ruolo degli archivi come dispositivi di conoscenza, partecipazione e rigenerazione territoriale, in linea con i paradigmi contemporanei delle Digital Humanities (Tammaro & Casarosa, 2022; UNESCO, 2003; Balzani, Raco & Suppa, 2019).

4. CONFRONTO E SCELTA DELLA PIATTAFORMA

La selezione della piattaforma archivistica ha rappresentato un nodo metodologico centrale all'interno del progetto, in quanto da essa dipendono la sostenibilità, la scalabilità e l'efficacia del modello proposto. Non si è trattato semplicemente di individuare una soluzione tecnologicamente valida, ma di identificare uno strumento in grado di gestire l'eterogeneità documentaria generata in contesti post-sismici, assicurando al contempo la fruizione pubblica, l'interoperabilità semantica e la replicabilità del modello da parte di enti e comunità con risorse limitate (Salarelli, 2016; Balzani, Raco & Suppa, 2019).
La seguente tabella (vd.Tabella1) sintetizza i risultati del confronto:

Tabella 1 - confronto tra OmekaS e CollectiveAccess

Obiettivo primario	Narrazione interattiva, pubblicazione e accessibilità online.	Gestione complessa e dettagliata di dati archivistici.
Interfaccia	Intuitiva e semplice; adatta a utenti non esperti (Alcaraz-Martínez, 2022).	Complessa; richiede competenze tecniche avanzate (Ravikumar & Hazarika, 2020).
Standard archivistici	Supporta Dublin Core, ISAD(G), RiC tramite plugin (Piotrowski, 2024).	Supporta ISAD(G), ISAAR(CPF), CIDOC CRM (Ravikumar & Hazarika, 2020).
Gestione multimediale	Ottimale per immagini, video, mostre virtuali (Alcaraz-Martínez, 2022).	Perfetta per modelli 3D, file CAD, video (Popović, Škorić, & Rujević, 2020).
Web Semantico	Integrazione tramite RDF Import e SPARQL Endpoint (Bruneau et al., 2020).	Supporto nativo per LOD e RDF (Ravikumar & Hazarika, 2020).
Personalizzazione	Moderata; configurabile con moduli aggiuntivi (Alcaraz-Martínez, 2022).	Massima flessibilità per progetti complessi (Ravikumar & Hazarika, 2020).

Alla luce di questi elementi, Omeka S è stato selezionato come strumento di riferimento per la fase attuale del progetto. La scelta è motivata dalla sua struttura modulare, dalla relativa facilità di installazione e manutenzione, dalla capacità di rappresentare relazioni complesse tra entità archivistiche, e dalla compatibilità con i principali standard descrittivi e semantici.
L'ambiente di installazione è stato inizialmente configurato in locale, su una macchina virtuale (VM) con sistema operativo Debian GNU/Linux, in ambiente LAMP (Linux, Apache, MariaDB, PHP). Questa soluzione ha permesso di effettuare le prime fasi di sviluppo e test in un contesto controllato, garantendo flessibilità nella personalizzazione e facilità di manutenzione.
L'installazione di Omeka S è avvenuta con database MariaDB e PHP 8.1, e con l'attivazione dei principali moduli per l'importazione di dati (CSV Import), la definizione di vocabolari personalizzati (Custom Vocab, Value Suggest), la compatibilità semantica (RDF Import) e l'integrazione con il framework IIIF (IIIF Server).
Il progetto prevede in una fase successiva il trasferimento dell'intero archivio digitale sulla piattaforma web ufficiale dell'associazione Identità Sibillina ETS, che ospiterà i contenuti in modo permanente, garantendo accessibilità pubblica, integrazione con le attività territoriali e aggiornamenti futuri.
L'adozione iniziale di Omeka S non preclude comunque, in una fase successiva, l'estensione del sistema verso piattaforme a maggiore granularità come CollectiveAccess o ICA-AtoM, qualora il progetto si espanda verso un livello inter-istituzionale o territoriale più ampio. Al momento, però, Omeka S garantisce

un equilibrio tra usabilità, potenziale semantico e sostenibilità, in linea con le più recenti riflessioni sulla natura trasformativa degli archivi digitali (Fiorani & Romeo, 2020; Grimoldi, 2017).

5. LINEE GUIDA E STANDARD INTERNAZIONALI

La costruzione dell'archivio combina, come già accennato, l'adozione delle Linee Guida ANAI 2023 per la documentazione tecnica e progettuale con l'applicazione degli standard archivistici internazionali per i documenti storici e amministrativi. Tale approccio ibrido si rivela necessario per garantire una corretta descrizione, conservazione e accessibilità di un corpus documentario multiforme, generato in contesti e con finalità differenti.

Le Linee Guida ANAI forniscono un framework orientato alla gestione dei documenti tecnici e degli archivi di progetto, prodotti durante le fasi di restauro e ricostruzione. Esse favoriscono l'adozione di buone pratiche documentarie in ambito architettonico e ingegneristico, ponendo l'accento su elementi come la tracciabilità delle fonti, la conservazione digitale e l'interoperabilità (ANAI, 2023).

Parallelamente, i documenti storici e d'archivio, recuperati presso gli archivi, vengono trattati secondo gli standard ISAD(G) (General International Standard Archival Description) e Records in Contexts (RiC), per organizzare logicamente le risorse in fondi, serie, sottoserie e unità documentarie, preservando la relazione gerarchica e il principio di provenienza (International Council on Archives, 2000) e in prospettiva di una modellazione semantica più avanzata, con la prospettiva di introdurre entità e relazioni che riflettono la complessità delle connessioni tra documenti, soggetti, enti produttori ed eventi (ICA, 2016).

A completare il quadro, lo standard Dublin Core è utilizzato in modo trasversale per la descrizione minimale e interoperabile delle risorse. La sua flessibilità e l'adozione diffusa lo rendono ideale per ambienti digitali orientati al Web Semantico (Weibel et al., 1998).

Per quanto riguarda la struttura dell'archivio, l'intero progetto è concepito come un archivio digitale multi-tipologico, in cui ogni corpus documentario è rappresentato come un fondo (item). All'interno di ciascun fondo sono organizzate le serie archivistiche (item set), a loro volta articolate in rubriche tematiche (altri item set) che contengono i singoli documenti (item).

Le figure 2-3 chiariscono il modello applicato e forniscono l'esempio della "Rubrica 1- Acque e Strade" (1816) dell'Archivio Comunale di Camporotondo di Fiastrone, conservato presso l'Archivio di Stato di Macerata - sezione di Camerino.

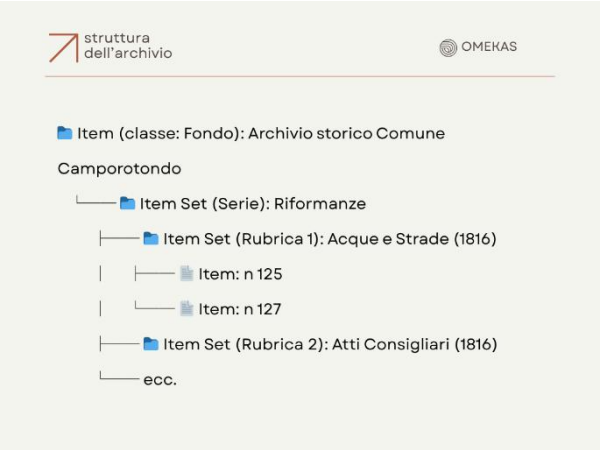


Figura 2 - Struttura gerarchica dell'archivio digitale in Omeka S.

Questa struttura consente di mantenere la coerenza archivistica e al contempo valorizzare la descrizione documentaria in ambiente digitale. Ogni risorsa è descritta con metadati Dublin Core estesi, e sono in fase di implementazione vocabolari personalizzati coerenti con ISAD(G) e RiC, tramite l'uso dei moduli Custom Vocab e Value Suggest di Omeka S. L'adozione dell'RDF framework consente la rappresentazione semantica dei dati e ne abilita l'interrogazione via SPARQL Endpoint.

L'integrazione del modulo IIIF Server consente invece di servire immagini ad alta risoluzione attraverso standard IIIF facilitando l'accesso, la comparazione e l'inclusione in viewer esterni come Universal Viewer o Mirador: ogni immagine associata a un documento è dotata di manifest.json generato automaticamente e può essere richiamata da altri ambienti digitali, favorendo l'integrazione con progetti di ricerca e piattaforme culturali.

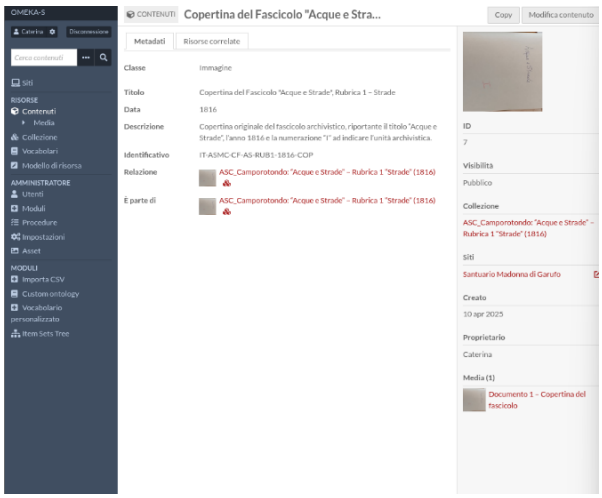


Figura 3 - Scheda descrittiva dell'item: "Copertina del Fascicolo Acque e Strade"

In prospettiva, è prevista l'integrazione con Archivematica, al fine di garantire la preservazione a lungo termine tramite l'adozione dei metadati PREMIS e la creazione di pacchetti informativi (AIP/DIP).
L'obiettivo è rendere l'archivio compatibile con i più avanzati sistemi di digital preservation e facilitare il riuso dei dati in progetti futuri.
La progettazione mira, dunque, a proporre un modello sostenibile, replicabile e accessibile, capace di coniugare esigenze archivistiche, documentazione tecnica e valorizzazione partecipata del patrimonio culturale.
In conclusione, l'intero impianto dell'archivio è concepito per essere replicabile in altri contesti post-sismici o di emergenza culturale, proponendo un modello di archiviazione digitale accessibile, sostenibile e scalabile, in linea con i più aggiornati paradigmi delle Digital Humanities e della gestione partecipata del patrimonio culturale (Fiorani & Romeo, 2020; UNESCO, 2003; Balzani et al., 2019).

6. CONCLUSIONI

Il progetto dell'archivio digitale del Santuario della Madonna di Garufo rappresenta un laboratorio aperto per la sperimentazione di pratiche archivistiche e digitali applicate al patrimonio culturale in contesti post-sismici. In un'epoca in cui la fragilità dei beni culturali si intreccia con l'urgenza di nuove modalità di tutela e valorizzazione, l'integrazione tra approcci tradizionali e strumenti delle Digital Humanities si configura come una strada necessaria per rispondere in modo consapevole e sostenibile alle sfide contemporanee.
La scelta di adottare una piattaforma open source come Omeka S risponde all'esigenza di costruire un sistema accessibile, progressivamente interoperabile e potenzialmente replicabile, adatto a contesti in cui le risorse – tecniche, economiche e umane – sono limitate. Pur essendo ancora in fase di implementazione, il progetto prevede l'integrazione di standard archivistici come ISAD(G) e Records in Contexts (RiC-CM) attraverso vocabolari personalizzati, l'adozione di tecnologie del Web Semantico (RDF, SPARQL, IIIF) e l'allineamento completo alle buone pratiche per la documentazione tecnica e progettuale delineate dalle Linee Guida ANAI 2023.
L'obiettivo non è solo quello di conservare e rendere accessibili i materiali prodotti nel corso del restauro, ma anche quello di creare un modello dinamico di archivio che funzioni come hub di conoscenza, capace di attivare processi di partecipazione, riuso, educazione e memoria condivisa attraverso l'accesso e la correlazione di risorse, storie, luoghi e fonti. L'esperienza del Santuario di Garufo ci porta a riflettere su come l'archivio possa contribuire, se adeguatamente progettato, alla rigenerazione culturale dei territori colpiti da eventi traumatici, rafforzando il legame tra comunità, patrimonio e tecnologia.
In prospettiva, l'ulteriore sviluppo della piattaforma – attualmente ospitata su server locale, ma destinata a essere trasferita sull'infrastruttura digitale dell'associazione Identità Sibillina ETS – consentirà di testare e perfezionare le soluzioni adottate, ampliando l'archivio a nuovi fondi, nuove tipologie documentarie e nuove forme di fruizione. La costruzione di questo ecosistema documentario rappresenta, dunque, non solo un'opportunità scientifica, ma anche un atto di responsabilità verso la memoria culturale e identitaria delle comunità locali.

BIBLIOGRAFIA

Alcaraz-Martínez, R. (2022). Omeka S como alternativa para el desarrollo de colecciones digitales y proyectos de humanidades digitales. *BiD: textos universitaris de biblioteconomia i documentació, 48*, 1–20. https://doi.org/10.1344/BiD2022.48.06

Allevi, F. (2000). *Il Santuario di S. Maria di Garufo*. In Accademia Filelfica (a cura di), *Quaderni di ricerca storica (1997–2000)* (pp. 3–22). Tolentino: Biblioteca Egidiana.

ANAI. (2023). *Linee guida per la formazione, gestione e conservazione degli archivi digitali e ibridi negli studi di progettazione.* Roma: Associazione Nazionale Archivistica Italiana.

Artefactual Systems. (2023). *Archivematica Documentation.* Retrieved from https://www.archivematica.org

Balzani, M., Raco, F., & Suppa, M. (2019). *Protocolli di acquisizione e gestione dati per la documentazione, rappresentazione e conservazione del patrimonio culturale danneggiato dal sisma.* Ferrara: Università degli Studi di Ferrara.

Bittarelli, A. (2004). *Pievefavera romana e medievale*. Pievefavera: Associazione "Quelli che… Pievefavera".

Bruneau, O., Lasolle, N., Lieber, J., Nauer, E., & Pavlova, S. (2020). Applying and developing semantic web technologies for exploiting a corpus in history of science: The case study of the Henri Poincaré correspondence. *Semantic Web.*

Brunetti, D. (2018). L'archivio multitipologico: definizione, descrizione e identità. *La Gazette des archives, 249*, 39–49. https://doi.org/10.3406/gazar.2018.5590

Digital Preservation Coalition. (2022). *Digital Preservation Handbook.* Retrieved from https://www.dpconline.org

Fiorani, D., & Romeo, E. (2020). Realizzazione degli interventi, valorizzazione, gestione, prevenzione: una panoramica introduttiva. In S. F. Musso & M. Pretelli (Eds.), *Restauro: Conoscenza, Progetto, Cantiere, Gestione* (pp. 519–525). Roma: Edizioni Quasar.

Grimoldi, A. (2017). Storia della costruzione, storia materiale del costruito, tutela e conservazione del patrimonio architettonico. In D. Fiorani (Coord.), *RICerca/REStauro. Conoscenza dell'edificio: metodo e contenuti* (pp. 481–495). Roma: Edizioni Quasar di S. Tognon.

Ingram-Monteiro, N., & McKernan, R. (2022). An Omeka S Repository for Place- and Land-Based Teaching and Learning. *Information Technology and Libraries, 41*(3). https://doi.org/10.6017/ital.v41i3.15123

International Council on Archives (ICA). (2016). *Records in Contexts: A Conceptual Model for Archival Description.* Paris: ICA.

Library of Congress. (2023). *PREMIS Data Dictionary for Preservation Metadata.* Retrieved from https://www.loc.gov/standards/premis

Metin, A., & Rognoni, F. (2024). Storia dell'architettura e Digital Humanities per la conoscenza e la conservazione del patrimonio costruito: un'esperienza italiana. *magazén, 5*(1), 1–28. https://doi.org/10.30687/mag/2724-3923/2024-01-001

Openpolis. (2023). Nelle Marche più beni culturali nelle zone a maggior rischio sismico. Accesso da: https://www.openpolis.it/numeri/nelle-marche-piu-beni-culturali-nelle-zone-a-maggior-rischio-sismico/ Consultato 10 Aprile 2025

Piotrowski, D. M. (2024). Omeka S: otwarte studium bibliometryczne. *Biuletyn EBIB, 3*(214), 3–15.

Popović, P., Škorić, M., & Rujević, B. (2020). The use of the Omeka Semantic platform for the development of the University of Belgrade, Faculty of Mining and Geology digital repository. *Infotheca, 20*(1–2), 136–148.

Ravikumar, S., & Hazarika, H. J. (2020). Comparative Study between Omeka, DSpace and E-Print: A Special Reference with Plugin Features. *Conference Paper,* 1–12.

Rueff, B. (2023). Dealing with post-excavation data: the Omeka S TiMMA web-database. *Archaeological Data Management Journal,* 12–34.

Salarelli, A. (2016). Gestire piccole collezioni digitali con Omeka: l'esperienza di MoRE (A Museum of Refused and Unrealised Art Projects). *Bibliothecae.it, 5*(2), 177–200. https://doi.org/10.6092/issn.2283-9364/6393

Tufarelli, M. (2019). *Futuro del patrimonio, patrimonio del futuro.* Firenze: Università degli Studi di Firenze.

UNESCO. (2003). *Convention for the Safeguarding of the Intangible Cultural Heritage.* Paris: UNESCO.

Weibel, S., Kunze, J., Lagoze, C., & Wolf, M. (1998). Dublin Core Metadata for Resource Discovery. *International Journal on Digital Libraries, 1*(2), 108–114. https://doi.org/10.1007/s007990050002

Il progetto *ArPeR*. Per un *Archivio dei periodici romaneschi*

Martina Ludovisi
[1] Università di Perugia, Italia, martina.ludovisi@unipg.it

ABSTRACT (ITALIANO)

Scopo del progetto di cui qui si riferisce è realizzare, in un ambiente web integrato e open access, un archivio dei periodici dialettali romani pubblicati tra l'Unità e la Prima Guerra Mondiale (1870-1920), in forma digitale e liberamente consultabile in linea. Non ci si limita però alla sola operazione di digitalizzazione: la possibilità di interrogare il *corpus* secondo diversi parametri (per forme, autori, componimenti, riviste ecc.), obiettivo primario del progetto, consentirà sia di riportare alla luce figure oggi neglette, contribuendo ad ampliare l'attuale conoscenza del panorama dialettale capitolino di fine Ottocento-primo Novecento, sia di porre le basi per ricerche future. Il *corpus* è digitalizzato mediante procedure di OCR e immesso in una piattaforma realizzata per il progetto. I principali risultati attesi sono: costituire una Sala di lettura open access che contenga le riproduzioni digitali dei periodici scelti per la costituzione del *corpus*; realizzare una biblioteca digitale che consenta di interrogare l'intero *database* tramite una ricerca personalizzata; offrire una prima descrizione, specie lessicografica, del romanesco postunitario.
Parole chiave: Periodici romaneschi; Archivio digitale; digitalizzazione; romanesco.

ABSTRACT (ENGLISH)

Paper Title for AIUCD2025. The *ArPeR* Project. For an Archive of Romanesco Periodicals

This paper presents the purposes of the SNSF-Project ArPeR.Archive of Romanesco Periodicals. The project seeks to establish a digital archive of Roman dialect periodicals published between the Unification of Italy and World War I (1870–1920), which will be freely accessible online. By digitizing these dialect texts, the project makes this important material available to the public for the first time. It also provides an opportunity to highlight previously overlooked figures, contributing to a richer understanding of the Roman dialect landscape in the late 19th and early 20th centuries, while laying a foundation for future research. The corpus is being digitized using OCR procedures and entered into a platform created specifically for the project. The main expected results are: the creation of an open-access Reading Room containing digital reproductions of the periodicals selected for the corpus; the creation of a digital library allowing users to query the entire database through personalized searches; and the provision of an initial description, particularly lexicographic, of post-unification Romanesco.

Keywords: Romanesco Periodicals; Digital Archive; Digitization; Romanesco

1. INTRODUZIONE

Fornite le linee generali del progetto, si potrebbe subito osservare che una riproduzione fotografica di alcuni numeri di testate giornalistiche dialettali è già prodotta da piattaforme come Internet culturale o dalle emeroteche digitali delle singole biblioteche; ma i loro archivi, oltre a riprodurre solo pochi fascicoli, non sono agilmente interrogabili, e non consentono alla comunità scientifica di sfruttare appieno il materiale che conservano. L'impossibilità di indagare agilmente tale patrimonio, considerato il cospicuo numero di fascicoli dei diversi giornali pubblicati in quegli anni, è un fattore che scoraggia la ricerca.[1] Questa lacuna tuttavia può essere colmata: grazie alle attuali risorse informatiche è possibile offrire al fruitore un archivio digitale con un *corpus* testuale praticamente inedito, finora solo parzialmente indagato (cfr. il §2), e facilmente interrogabile dall'utente (cfr. il §3): il progetto ArPeR si muove infatti su un doppio binario: oltre a digitalizzare un materiale di valore storico-linguistico-letterario, si propone di fornire all'utente una lista di testi e relativi autori che, come vedremo (vd. il §4), ha già prodotto l'acquisizione di nuovi dati e l'aggiornamento della bibliografia primaria di alcuni scrittori.

[1] All'interno del progetto è prevista comunque l'acquisizione del materiale già digitalizzato e disponibile in rete, come nel caso della rivista «Fornarina», i cui 21 numeri dell'anno 1883 sono liberamente accessibili (vd. https://books.google.it/books?id=bCklnr0C6M0C&printsec=frontcover&hl=it&source=gbs_ge_summary_r&cad=0#v=onepage&q&f=false). Uno degli obiettivi, infatti, è quello di stabilire connessioni con risorse esterne, già digitalizzate e disponibili online, così da valorizzare e riutilizzare il materiale già esistente.

2. STATO DELLA QUESTIONE

Le prime ricerche filologiche e linguistiche sul dialetto dell'Urbe risalgono al periodo compreso tra la fine del secolo XIX e l'inizio del XX, quando studiosi quali Ernest Bovet (1898), Tito Morino (1899), Giuseppe Senes (1900) e i più celebri Francesco Sabatini (1890) e Graziadio Isaia Ascoli (1900) si interessarono al romanesco coevo e alla lingua del Belli, attenzione suscitata dalla pubblicazione dell'edizione Morandi (1886-1889) dei *Sonetti*. Posto l'interesse per il romanesco belliano, i successivi lavori scientifici, come quelli di Merlo (1929) e Migliorini (1932),[2] si concentrarono quasi esclusivamente sul romanesco di prima fase, studi che col tempo non hanno fatto altro che moltiplicarsi, complici il dibattito su tempi e modi della toscanizzazione del dialetto capitolino, che ha interessato diversi studiosi (una sintesi si trova in Trifone, 2008) e il rinnovato interesse per le ricerche d'archivio. Dunque, a fronte di un quadro assai mutato per il romanesco di prima fase, lo studio di quello ottocentesco è invece ancora oggi fin troppo incentrato sulla figura del Belli e sul dialetto documentato dai *Sonetti*, benché le vicende postunitarie abbiano avuto importanti ripercussioni sulla varietà capitolina, che proprio in questi decenni comincia a conoscere una serie di trasformazioni strutturali a tutti i livelli di analisi (cfr. *infra* e in particolare Faraoni, 2021). Data la centralità di questo periodo storico anche da un punto di vista linguistico, è evidente che per approfondire la conoscenza del dialetto postunitario la soluzione più efficace è quella di allargare la base documentaria: l'universo delle riviste dialettali dell'Otto e del Novecento è certo un buon punto di partenza. Del resto, già il lavoro di Picchiorri (2019: 478) sui testi romaneschi contenuti nei giornali nati tra il 1848 e il 1849 ha dimostrato come questi «merit[i]no attenzione dal punto di vista linguistico come spaccato realistico [...] del romanesco di metà Ottocento, soprattutto perché presentano un ricco patrimonio lessicale, spesso scarsamente o niente affatto documentato dai repertori, e testimon[i]no l'emersione di alcuni tratti morfosintattici incipienti nel dialetto dell'epoca». Simili risultati sono stati prodotti esaminando solo pochi fogli scritti interamente in dialetto, giacché la stampa periodica romanesca avrà il suo periodo di maggiore splendore solo più tardi, proprio in quei decenni postunitari che il progetto mira a indagare; è certo, quindi, che dai giornali di quest'epoca si possano ricavare dati utili per la ricerca scientifica. D'altra parte, anche Faraoni (2021) ha evidenziato l'esistenza di due direttrici per l'evoluzione del "romanesco postunitario": una che muove in direzione dell'italiano, con il progressivo avvicinamento del dialetto alla lingua standard, un'altra che muove in direzione anti-italiana e che ha consentito la sopravvivenza del romanesco. Di quest'ultima direttrice, che Stefinlongo (1985, 26) identifica con la meridionalizzazione otto-novecentesca, il materiale che si intende raccogliere tramite il presente progetto è testimone d'eccezione, specie dal punto di vista lessicale, con l'intrusione da sud di quel «diuturno flusso di napoletanismi che segna tutta la storia preunitaria e [in particolare] postunitaria del romanesco» (De Mauro, 1989: XXVII). Malgrado ciò, la visione che si ha della varietà capitolina tra Belli (metà Ottocento) e Trilussa (particolarmente attivo nei primi tre decenni del Novecento) è allo stato attuale piuttosto lacunosa: non molti gli studi sistematici sugli autori dialettali dell'ultimo trentennio del sec. XIX, le cui opere risentono spesso del modello belliano che, in parte, àncora il romanesco a una fase pregressa (ne sono un fulgido esempio i debiti belliani contratti da Zanazzo, su cui cfr. Costa, 2011: 35). Quest'ombra però non si allunga sulla pubblicistica dialettale, dove la scrittura in prosa risulta, com'è ovvio, più spontanea e più ricca di nuove voci importate da sud o dalle varietà contermini, dai flussi migratori d'epoca postunitaria e perciò non registrate da Belli. Sul versante storico-letterario, la pubblicistica d'età umbertina interessa per le tematiche affrontate nei diversi periodici e per le relative posizioni politiche: Puglisi (2011), ad esempio, per quanto riguarda il «Rugantino», segnala quelle collegate alla polemica anticlericale, all'impegno politico di Crispi, all'erezione del monumento a Giordano Bruno e, più in generale, ai fatti di cronaca. Un'attenta disamina dei materiali contenuti nei giornali costituirà quindi anche un ausilio per lo studio della situazione politico-culturale e sociale della Roma postunitaria.

La mancanza di uno strumento di consultazione che consenta di estrapolare dai testi giornalistici romaneschi di fine Ottocento informazioni mirate e allo stesso tempo permetta la visione del suddetto materiale è stata alla base dell'ideazione di *ArPeR*.

3. METODOLOGIE E ARCHITETTURA DELL'OPERA

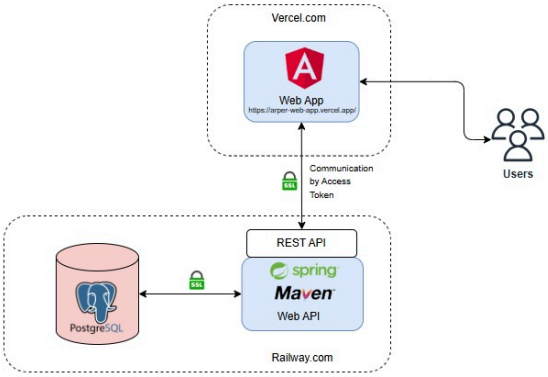
Per iniziare, è utile esaminare la struttura dell'opera. Il *corpus* è costituito dai fascicoli di alcuni dei più importanti giornali dialettali d'epoca postunitaria, elencati qui di séguito in ordine cronologico: 50 fascicoli di «Don Pirloncino» (30 lug. 1871 - 2 gen. 1887), 279 di «Capitan Fracassa» (25 mag. 1880 – 10 ott. 1905), 35 di «Fornarina» (7 gen. 1882 - 3 giu. 1883), 654 fascicoli del «Rugantino» (18 set. 1887 - 28

[2] È più tardo il fondamentale lavoro di Ernst, 1970.

dic. 1915), 20 di «Casandrino in dialetto romanesco» (8 agosto – 14 ottobre 1897) e 299 di «La frusta» (gen. 1873 - 31 dic. 1873).[3] Considerata la consistenza documentaria, al momento della selezione si sono privilegiati i periodici conteneti un numero maggiore di testi in dialetto; nonostante ciò, il *database* è destinato a un continuo ampliamento e, in una fase successiva, sarà possibile aggiungere testi e riproduzioni di altri giornali. Una volta definite le specifiche tecniche per i formati dei file digitali – selezionate sulla base di standard internazionali – il progetto si compone di tre fasi operative principali:

1) Strutturazione del lavoro: inserimento dei dati in un apposito file Excel per facilitare il processo di organizzazione dei numeri; identificazione dei problemi di indicizzazione (file mancanti); individuazione dei numeri assenti nelle biblioteche.
2) Acquisizione del materiale: è la fase in cui le varie testate vengono scannerizzate. Successivamente, si valuta la qualità delle immagini digitali e, se non conformi alle specifiche approvate, si procede con una nuova acquisizione.
3) Trascrizione (e correzione) dei testi del *corpus*: il materiale digitalizzato viene trascritto tramite metodologie e tecnologie di OCR open source e commerciali (come Tesseract,[4] ABBYY FineReader[5] ed eScriptorium[6]) scegliendo il formato plain text, che permette di preparare i testi a una successiva fase di revisione. Tra le soluzioni impiegate, ABBYY FineReader, il più impiegato, è stato oggetto di un addestramento mirato, basato sulla correzione manuale di un campione selezionato di pagine particolarmente rappresentative del *corpus*. Questo processo ha consentito al software di apprendere le specificità grafiche e tipografiche dei periodici romaneschi ottocenteschi, migliorando progressivamente la precisione del riconoscimento. L'addestramento ha riguardato in particolare l'identificazione di caratteri irregolari, l'interpretazione di impaginazioni non standard e la gestione delle forme dialettali che, in assenza di norme ortografiche condivise, si presentano spesso con elevata variabilità.
4) Inserimento: i documenti in formato testo vengono salvati nella piattaforma creata per il progetto, un database relazionale PostgreSQL, insieme alla relativa riproduzione fotografica.[7]

I testi delle riviste, correttamente trascritti e accompagnati dal documento riprodotto nella sua veste originaria, sono resi accessibili e interrogabili dagli utenti mediante un'interfaccia web.
L'applicazione web è composta da due fondamentali componenti: un'interfaccia utente implementata utilizzando Angular (front-end) e un'API web (back-end) costruita tramite il framework Spring Boot. L'interazione fra i due sistemi avviene in modalità sicura, attraverso l'uso di un meccanismo di autenticazione basato su token di accesso (login dell'utente, limitato al pannello di controllo), che assicura l'integrità e la privacy dei dati scambiati. Dal punto di vista della distribuzione, l'applicazione front-end è ospitata sulla piattaforma Vercel, mentre l'API back-end è distribuita tramite Railway: si tratta, per entrambe, di soluzioni cloud che offrono infrastrutture scalabili e ad alte prestazioni; una scelta che permette da un lato di ottimizzare la gestione delle risorse, dall'altro di garantire una risposta efficiente alle richieste dei fruitori. Di séguito, un grafico che illustra i rapporti descritti:



[3] Si tratta di numeri approssimativi, giacché si annoverano all'interno del *corpus* esclusivamente i fascicoli che contengono testi dialettali. Per il «Rugantino» e il «Casandrino» vuol dire considerarli tutti, essendo periodici redatti esclusivamente in dialetto; diversamente, per «Fornarina» ci sono state esclusioni.
[4] https://github.com/tesseract-ocr/tesseract
[5] https://pdf.abbyy.com/it/
[6] https://gitlab.com/scripta/escriptorium
[7] Attraverso Apache PDFBox (https://pdfbox.apache.org/), libreria java open source, è inoltre possibile inserire nella piattaforma ArPeR l'eventuale pdf a nostra disposizione ed estrarre, visualizzare e correggere il testo.

Figura 1. Grafico dei rapporti

4. PRIMI RISULTATI E PROSPETTIVE ATTESE

La piattaforma per l'interrogazione della banca dati testuale è tuttora in fase di controllo e di implementazione; di séguito un esempio del prototipo (figura 2):



Figura 2. Esempio di una ricerca per forme

Il motore di ricerca consente l'impiego dei caratteri jolly e la possibilità di effettuare una ricerca espansa, ovvero senza tener conto degli accenti: la ricerca di una forma dà inoltre accesso ai dati a essa relativi (ad es. autori che la impiegano, riviste e fascicoli dove compare) con la possibilità di visualizzare sia contesti immediati, di lunghezza prestabilita, sia contesti allargati.[8] Sarà possibile, poi, effettuare indagini incrociando più parametri: per esempio, basterà selezionare un autore per trovare le occorrenze di quella forma nei testi a lui attribuiti o attribuibili, raffinando così la propria ricerca.[9] Peraltro, il collegamento testo-autore, effettuato dall'operatore nel pannello di controllo di ArPeR, è un'operazione con risvolti non trascurabili: infatti, sebbene la funzione primaria della presente banca dati sia quella di fornire un *corpus* di testi dialettali che serva di base per ricerche linguistiche e lessicografiche, la piattaforma ArPeR restituisce per ogni autore testi che a volte non erano stati considerati parte della loro bibliografia primaria. Del resto, l'indagine sugli autori e sui relativi pseudonimi, imprescindibile per la creazione delle schede "Autori",[10] ha già integrato la bibliografia sinora nota per Zanazzo (vd. figura 3): i componimenti *Bbone Feste* e *Er parlà de l'Africani*, benché pubblicati nel «Rugantino», non sono stati inclusi nelle antologie complete postume più importanti, quella del figlio Alfredo (Zanazzo, 1921) e quella di Orioli (1968).

[8] Si è scelto di non inserire una ricerca di tipo fuzzy, in quanto tale modalità, se utile per ricerche di tipo lessicografico, rischia di restituire risultati troppo ampi e non controllabili, inadatti a indagini linguistiche mirate che richiedono corrispondenze esatte, come nel caso dell'analisi di tratti fonologici o morfologici specifici. Si è pertanto optato per una ricerca esatta seguendo i criteri adottati nella costruzione del *corpus OVI*, ormai consolidati nella prassi degli studi linguistici e filologici, dove è consuetudine operare su corrispondenze formali precise per garantire l'affidabilità dell'analisi.

[9] Riguardo agli scritti finora anonimi, l'attribuzione avverrà su base stilistica e si ricorrerà alla formula [*nome dell'autore* (attrib.)] per registrare tale attribuzione nel *database*. Sarà fondamentale, in quest'ultima fase del progetto, l'uso del software Stylo impiegato presso l'Università Jagellonica di Cracovia per le ricerche stilometriche (cfr. Eder, Kestemont & Rybicki, 2016).

[10] Ciò è possibile specie grazie alle schede di Majolo Molinari (1963) sulla stampa periodica romana dell'Ottocento e sui suoi protagonisti.

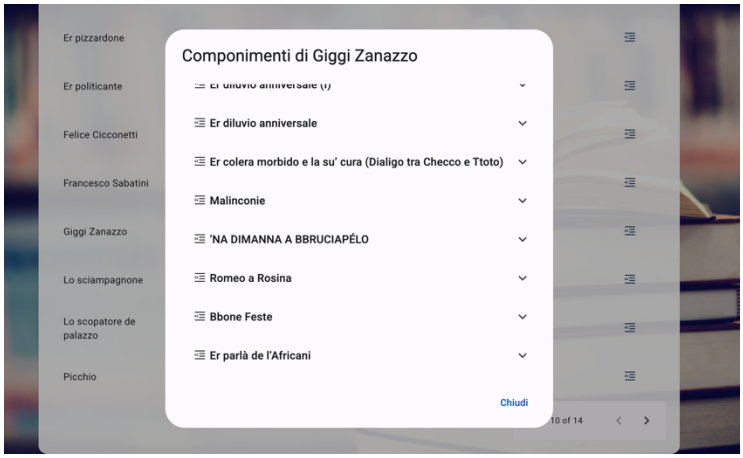


Figura 3. Lista testi di un autore

Se è possibile aggiornare la bibliografia primaria uno dei maggiori autori dialettali di quest'epoca, si può ben immaginare il margine di miglioramento della situazione dei poeti "minori" (per es. Giaquinto, Ilari e Francati; sul «Rugantino», inoltre, scrive il giovane Trilussa).
Non va dimenticato, in conclusione, che la prospettiva *open source* permette interventi di aggiornamento e di perfezionamento di fatto infiniti: si pensi alla possibilità di inserire una ricerca per lemmi[11] o all'apertura dell'orizzonte cronologico verso i decenni precedenti e successivi, così da creare, col tempo, un *Archivio unico dei periodici romaneschi*.

RINGRAZIAMENTI

Il progetto (n. P500PH_222332) è finanziato dal Fondo Nazionale Svizzero per la Ricerca Scientifica per il periodo 2024-2026, nell'ambito della borsa Postdoc Mobility. Si ringrazia Michele Loporcaro per il sostegno dato nel corso dell'avviamento del progetto e in particolare Giulio Vaccaro per l'insostituibile guida nell'ideazione della struttura della piattaforma. L'intera realizzazione dell'output digitale si deve a Fabrizio Nicolò.

BIBLIOGRAFIA

Ascoli, G. I. (1900). Varia (Del romanesco ancora), *Archivio Glottologico Italiano*, 15, 323–325.

Bovet, E. (1898). *Le peuple de Rome vers 1840 d'après les sonnets en dialecte trastévérin de G. G. Belli*. Attinger.

Costa, C. (2011). «Nun se venne, è sincera, forte, onesta». La poesia romanesca di Giggi Zanazzo. In Onorati F. e Scalessa G. (Eds.), *Le voci di Roma. Omaggio a Giggi Zanazzo, Atti del convegno di studi* (pp. 33–45*)*, Roma: Il cubo. ISBN: 9788897431022.

De Mauro, T. (1989). Per una storia linguistica della città di Roma. in De Mauro, T. (Ed.). *Il romanesco ieri e oggi* (pp. 13-37). Roma: Bulzoni. ISBN: 978-88-7119-021-1.

Eder, M., Kestemont, M. e Rybicki, J. (2016). Stylometry with R: A package for Computational Text Analysis, in *The R Journal* 8(1), 107-121. DOI: 10.32614/RJ-2016-007.

Ernst, G. (1970). *Die Toskanisierung des romischen Dialekts im 15. und 16. Jahrhundert*. Niemeyer. https://doi.org/10.1515/9783111328492

Faraoni, V. (20-21 maggio 2021). *Il romanesco prima e dopo Porta Pia*. [Conference presentation] Dalla Roma pontificia alla Roma italiana. Le istituzioni culturali e la città, Roma, Università La Sapienza.

Majolo Molinari, O. (1963). *La stampa periodica romana dell'Ottocento*. Istituto Nazionale di Studi Romani. ISBN: 9788873114727.

Merlo, C. (1929). Vicende storiche della lingua di Roma I. Dalle origini al sec. XV. *L'Italia dialettale*, 5, 172-201. (Reprinted from *Saggi linguistici*, pp. 33–62, by C. Merlo, Ed., 1959, Pisa: Pacini-Mariotti).

[11] Allo stato attuale non è prevista una ricerca per lemmi data l'elevata richiesta di tempo e risorse, sia in modalità automatica (che richiederebbe addestramento e revisione) sia manuale. L'implementazione potrà eventualmente essere considerata, come detto, in una fase futura del progetto.

Migliorini, B. (1932). Dialetto e lingua nazionale a Roma. *Capitolium*, 10, 350–356. (Reprinted from *Lingua e cultura*, pp. 109–123, by B. Migliorini, Ed., 1948, Roma, Tumminelli).

Morandi L. (Ed.). 1886-1889. I sonetti, Città di Castello: Lapi.

Morino, T. (1899). *La grammatica del dialetto romanesco secondo i sonetti di G. G. Belli. I. Suoni - Forme*. Colitti.

Orioli. G. (ed.). 1968. *Poesie romanesche di G. Zanazzo*. Roma: Avanzini e Torraca.

Picchiorri, E. (2019) *«Nun vorrebbia che fusse na cianchetta der nemico». Il romanesco nei giornali della repubblica romana».* in Vaccaro, G. (Ed.). *Marcello 7.0. Studi in onore di Marcello Teodonio* (pp. 477–487), Roma: Il cubo. ISBN: 9788897431091.

Puglisi, P. (2011). «Ggente mia, garbata e bbella. Zanazzo giornalista». In Onorati F. e Scalessa G. (Eds.), *Le voci di Roma. Omaggio a Giggi Zanazzo, Atti del convegno di studi* (pp. 33–45*)*, Roma: Il cubo. ISBN: 9788897431022.

Sabatini, F. (1890). *Il volgo di Roma*. Loescher.

Senes, G. (1900). *Anatomia filologica del dialetto romanesco fatta sulla Scoperta dell'America di Cesare Pascarella*. Florenz.

Stefinlongo, A. (1985). Note sulla situazione sociolinguistica romana. Preliminari per una ricerca. *Rivista italiana di dialettologia*, 9, 43–67.

Trifone, P. (2008). *Storia linguistica di Roma*. Carocci. ISBN: 9788843044597.

Zanazzo, A. (ed.) 1921-1922. *Versi romaneschi editi ed inediti*, di G. Zanazzo. M. Carra e C. di Luigi Bellini.

Un modello integrato per il *Roman d'Alexandre* del codice Correr 1493: annotazione linguistica e edizione critica digitale

Giacomo Costa[1], Simone Zenzaro[2]
[1] Università degli Studi di Padova, Italia – giacomo.costa.4@phd.unipd.it
[2] Cnr-Istituto di Linguistica Computazionale "Antonio Zampolli", Italia - simone.zenzaro@cnr.it

ABSTRACT (ITALIANO)
L'obiettivo di questa proposta è di presentare alcune riflessioni, a livello sia pratico sia teorico, su di un nuovo modello di edizione critica digitale in cui i metadati derivati dall'annotazione linguistica di un testo possano essere assimilati e consultati. Lo scopo è di superare la divisione che spesso intercorre tra il piano ecdotico-filologico e quello dei corpora linguistici, proponendo, pertanto, lo sviluppo di un sistema integrato in cui il metadato linguistico possa essere interrogato in tutte le sue parti direttamente dall'interfaccia dedicata all'edizione critica digitale del testo oggetto di annotazione. L'oggetto di indagine di questo progetto sono testi che presentano particolarità linguistiche, in particolar modo legati a fenomeni di ibridismo tra diverse varietà e code-mixing. La *proof of concept* è in fase di sviluppo su di un testo medioevale appartenente al corpus del cosiddetto francese d'Italia, ovvero l'insieme di testi composti o trascritti in francese da autori della Penisola, tra il XIII e il XV secolo, nei quali i volgari medioevali italiani hanno trovato punti di contatto e di fusione con le varietà del francese antico.
Parole chiave: edizioni critiche digitali; annotazione linguistica automatica; corpora linguistici; Francese d'Italia (XIII-XV secolo)

ABSTRACT (ENGLISH)
An Integrated Model for the Roman d'Alexandre *of the manuscript Correr 1493: Linguistic Annotation and Scholarly Digital Edition*
The aim of this proposal is to present some reflections, both practical and theoretical, on a new model of scholarly digital editing in which metadata derived from the linguistic annotation of a text can be integrated and consulted. The goal is to overcome the division that often exists between the philological-ecdotical level and that of linguistic corpora, proposing the development of an integrated system where linguistic metadata can be queried in all its parts directly from the interface dedicated to the digital edition of the annotated text. The object of investigation for this project is texts that present linguistic peculiarities, related to phenomena of hybridity between different varieties and code-mixing. The proof of concept is under development on a medieval text from the corpus of what is called "The French of Italy", that is texts composed or transcribed by Italian authors between the 13th and 15th centuries, in which the medieval Italian vernaculars encountered points of contact and fusion with varieties of Old French.
Keywords: digital scholarly editions; automatic linguistic annotation; linguistic corpora; the French of Italy (13th -15th century)

1. INTRODUZIONE

Lo sviluppo di un'edizione critica digitale integrata alle buone pratiche dell'annotazione linguistica nasce dall'esigenza di rappresentare e strutturare in modo integrale e ordinato una serie di informazioni ecdotiche, linguistiche e letterarie insite nei testi appartenenti al corpus medievale del cosiddetto francese d'Italia (Gambino & Beretta, 2023). Per questo motivo proponiamo un prototipo di modello di edizione per il *Roman d'Alexandre* del codice Venezia, Biblioteca del Museo Correr, Correr 1493 (per l'edizione si veda La Du, 1937; per il facsimile del codice Benedetti, 1998).
Questo prototipo di edizione critica digitale integrato all'insieme dei dati derivati dal lavoro di annotazione linguistica ha l'obiettivo di configurarsi come modello unitario e generale per la banca data *RIALFrI* (*Repertorio Informatizzato Antica Letteratura Franco-Italiana*)[1] che raccoglie in linea il corpus dei testi del francese d'Italia. Al momento, difatti, le edizioni dei testi presenti nel repertorio si potrebbero configurare

[1] https://www.rialfri.eu/.

come edizioni «digitalizzate» poiché costituiscono digitalizzazioni di edizioni a stampa già esistenti. Tuttavia, con l'avviarsi di un progetto PRIN (*FrIngE - The French in/of Italy: Code-MixiNG in Medieval Europe* [2023-2025, PRIN 2022XRFSXR]), che ha portato all'annotazione linguistica completa di un certo numero di testi franco-italiani, è sorta anche la necessità di ripensare al modello di edizioni, per le quali il digitale non fosse più solo un medium di rappresentazione, ma anche e soprattutto la base di partenza per il lavoro, con un valore trasformativo su quest'ultimo (Pierazzo & Mancinelli, 2020: 9).

2. IL PROCESSO DI EDIZIONE

Il prototipo del modello di edizione è stato sviluppato a partire dal *Roman d'Alexandre* del codice Venezia, Biblioteca del Museo Correr, Correr 1493. La versione del *Roman* contenuta in questo manoscritto, databile al secondo quarto del XIII secolo, presenta una serie di caratteristiche linguistiche (tratti di ibridismo) ed ecdotico-materiali tali per cui risulta essere un campo di prova ideale per lo sviluppo di un modello di edizione digitale integrato.

Dopo una prima fase di lavoro filologico, durante la quale il testo è stato accuratamente trascritto e editato, quest'ultimo è stato sottoposto al lavoro di annotazione linguistica. Il testo è stato annotato linguisticamente tramite l'applicazione web *Pyrrha*, sviluppata dall' *École nationale des chartes* per l'analisi del francese antico (Clérice et al., 2024). L'utilizzo di uno strumento allenato sulle le varianti del francese antico su di un testo franco-italiano ha posto, fin dal principio, alcuni problemi legati alla scelta del *référentielle* per tutte quelle forme non appartenenti allo standard oitanico e alla necessità di allenare ulteriormente il modello affinché potesse essere efficiente anche per i testi del francese d'Italia (su questo aspetto si noti il lavoro già citato del progetto PRIN: *FrIngE - The French in/of Italy: Code-MixiNG in Medieval Europe* [2023-2025, PRIN 2022XRFSXR]; per i primi risultati cfr. Gambino et al., 2024).

Tramite questo strumento di annotazione e post-correzione, ad ogni forma è stato associato il lemma, la *part of speech* e il *morph tag*[2]. Inoltre, sono state applicate delle specifiche etichette all'analisi morfologica che segnalano la presenza di una forma ibrida o italianismo (SPEC=it), di un hapax causato dalla posizione in rima (SPEC=rim), di latinismi (SPEC=lat), la presenza di occitanismi (SPEC=occ) e, infine, di eventuali forme filologicamente complesse o di difficile attribuzione lemmatica (SPEC=probl). In particolar modo, l'utilizzo dell'etichetta «SPEC=it» permette di individuare tratti italianizzanti per certi aspetti devianti rispetto allo "standard" oitanico, consentendo, dunque, di svolgere delle indagini sulla distribuzione e sulla tipologia degli italianismi. Si riporta in Figura 1 un esempio di lavoro di annotazione. Le stringhe bianche sono state analizzate correttamente dal software, mentre il colore blu indica un intervento di correzione del risultato da parte dell'annotatore umano:

Id	Form	Lemma	POS	Morph	Context	Similar	Save	+
801	27rb	croire	OUT	MORPH=empty	PONfbl c . **27rb** De Eumenedus que	0	Save	
802	De	de	PRE	MORPH=empty	c . 27rb **De** Eumenedus que pasma	63	Save	
803	Eumenedus	Eumenedus	NOMpro	NOMB.=s|GENRE=m|CAS=r	. 27rb De **Eumenedus** que pasma por	8	Save	
804	que	que2	PROrel	SPEC=it|NOMB.=s|GENRE=m|CAS=n	27rb De Eumenedus **que** pasma por lo	144	Save	
805	pasma	pasmer	VERcjg	MODE=ind|TEMPS=psp|PERS.=3|NOMB.=s	De Eumenedus que **pasma** por lo sanc	0	Save	
806	por	por2	PRE	MORPH=empty	Eumenedus que pasma **por** lo sanc que	20	Save	
807	lo	le	DETdef	NOMB.=s|GENRE=m|CAS=r	que pasma por **lo** sanc que il	4	Save	
808	sanc	sanc	NOMcom	NOMB.=s|GENRE=m|CAS=r	pasma por lo **sanc** que il avoit	1	Save	

Figura 1. Annotazione del testo

L'integrazione di tutte queste informazioni linguistiche all'interno di un'edizione critica trova nel mezzo digitale, e in particolar modo nella codifica in XML/TEI, una sua congenita collocazione. Il supporto digitale, difatti, risulta l'unico efficace per questo obiettivo, sia per quanto riguarda la possibilità di conservazione a lungo termine dei dati sia per rendere disponibili a tutti una serie di informazioni altrimenti difficilmente rappresentabili tramite il supporto cartaceo.

Il software di annotazione *Pyrrha* permette di esportare i dati dell'annotazione linguistica direttamente in

[2] Il sistema di etichette per l'annotazione linguistica segue lo schema di Cattex (http://bfm.ens-lyon.fr/spip.php?article176).

formato XML/TEI grezzo, di cui la Figura 2 ne è un esempio, che costituisce la base di partenza per lo sviluppo dell'edizione digitale.

```
<w xml:id="t1" n="1" lemma="conte2" type="POS=NOMcom|NOMB.=s|GENRE=m|CAS=r" >Conte</w>
<w xml:id="t2" n="2" lemma="voloir" type="POS=VERcjg|MODE=ind|TEMPS=psp|PERS.=1|NOMB.=s" >voil</w>
<w xml:id="t3" n="3" lemma="dire" type="POS=VERinf|MORPH=empty" >dire</w>
<w xml:id="t4" n="4" lemma="par" type="POS=PRE|SPEC=it|MORPH=empty" >per</w>
<w xml:id="t5" n="5" lemma="rime4" type="POS=NOMcom|NOMB.=s|GENRE=f|CAS=r" >rime</w>
<w xml:id="t6" n="6" lemma="et" type="POS=CONcoo|MORPH=empty" >e</w>
<w xml:id="t7" n="7" lemma="par" type="POS=PRE|MORPH=empty" >per</w>
<w xml:id="t8" n="8" lemma="lëonime" type="POS=NOMcom|NOMB.=s|GENRE=f|CAS=r" >leonine</w>
<w xml:id="t9" n="9" lemma="de" type="POS=PRE|MORPH=empty" >De</w>
<w xml:id="t398" n="10" lemma="le" type="POS=DETdef|NOMB.=s|GENRE=m|CAS=r" >#l</w>
<w xml:id="t10" n="11" lemma="fil2" type="POS=NOMcom|NOMB.=s|GENRE=m|CAS=r" >fil</w>
<w xml:id="t11" n="12" lemma="Felipe" type="POS=NOMpro|NOMB.=s|GENRE=m|CAS=r" >Felipe</w>
<w xml:id="t12" n="13" lemma="le" type="POS=DETdef|NOMB.=s|GENRE=m|CAS=r" >le</w>
<w xml:id="t13" n="14" lemma="roi2" type="POS=NOMcom|NOMB.=s|GENRE=m|CAS=r" >roi</w>
<w xml:id="t14" n="15" lemma="de" type="POS=PRE|MORPH=empty" >de</w>
<w xml:id="t15" n="16" lemma="Macedoine" type="POS=NOMpro|NOMB.=s|GENRE=f|CAS=r" >Macedoine</w>
```

Figura 2. XML/TEI di partenza del testo annotato

I metadati linguistici in formato XML/TEI vengono successivamente rimodellati secondo la struttura del testo in versi e integrati con gli elementi di carattere filologico e letterario, come l'apparato critico, le varianti, note di commento e, inoltre, la riproduzione facsimilare del codice. Ad arricchire questo già complesso sistema di informazioni linguistiche e testuali, si riscontra nel *Roman d'Alexandre* del codice Correr 1493 un abbondante drappello di elementi paratestuali, come rubriche e note al miniatore, che sono stati altresì oggetto di annotazione linguistica, e che richiedono, dunque, di essere integrati nell'edizione critica digitale. Essendo gli elementi paratestuali portatori di preziose informazioni linguistiche, la sfida è stata quella di cercare di integrare quest'ultimi all'interno del testo in versi, rispettando la loro posizione, in questo caso - almeno per i nostri scopi - non più liminale, all'interno del codice. Sono state vagliate diverse alternative per riuscire a rappresentare la presenza di marginalia all'interno di una carta del codice, come la possibilità di inserirli tra due lasse o di raggrupparli tutti all'interno della descrizione del codice nel `<teiHeader>`, per arrivare, alla fine, alla soluzione mostrata in Figura 3.

```
<pb type="carta" n="1r"/>
            <noteGrp type="paratesto">
               <note xml:id="min_rub1ba" target="#lassa3" n="nota1" type="miniatura_rubricata">
                  <figure type="miniatura" n="c1ba" place="margine lato dx">
                     <head type="attribuito"> La nascità di Alessandro </head>
                     <figDesc >
                        <mentioned rendition="rubrica">
                           <w xml:id="t146" n="146" lemma="le" pos="DETdef" msd="NOMB.=s|GENRE=m|CAS=r" >Li</w>
                           <w xml:id="t147" n="147" lemma="jor" pos="NOMcom" msd="NOMB.=s|GENRE=m|CAS=r" >jor</w>
                           <w xml:id="t148" n="148" lemma="que2" pos="PROrel" msd="NOMB.=s|GENRE=m|CAS=r" >qu&#39;</w>
                           <w xml:id="t149" n="149" lemma="Alexandre" pos="NOMpro" msd="NOMB.=s|GENRE=m|CAS=n" >Alexandre</w>
                           <w xml:id="t150" n="150" lemma="naistre" pos="VERcjg" msd="MODE=ind|TEMPS=psp|PERS.=3|NOMB.=s"
>nasqui</w>                <w xml:id="t151" n="151" lemma="estre1" pos="VERcjg" msd="MODE=ind|TEMPS=psp|PERS.=3|NOMB.=p" >furent</w>
                           <w xml:id="t152" n="152" lemma="le" pos="DETdef" msd="NOMB.=p|GENRE=m|CAS=n" >li</w>
                           <w xml:id="t153" n="153" lemma="signe1" pos="NOMcom" msd="NOMB.=p|GENRE=m|CAS=n" >signes</w>
                           <w xml:id="t154" n="154" lemma="en1" pos="PRE" msd="MORPH=empty" >en</w>
                           <w xml:id="t155" n="155" lemma="ciel" pos="NOMcom" msd="NOMB.=s|GENRE=m|CAS=r" >ciel</w>
                           <w xml:id="t156" n="156" lemma="." pos="PONfrt" msd="MORPH=empty" >.</w>
                        </mentioned>
                     </figDesc>
                  </figure>
               </note>
```

Figura 3. Integrazione degli elementi paratestuali nel testo in versi

Tutti gli elementi paratestuali sono stati inseriti all'inizio di ogni carta, all'interno del tag `<noteGrp>` personalizzato dal `<type>` "paratesto". All'interno è stata segnalata l'eventuale presenza di rubriche o di miniature rubricate, con la rispettiva analisi linguistica, oppure di istruzioni per il miniatore, anch'esse annotate linguisticamente. Ogni elemento è stato associato a un "id" a sua volta collegato alla lassa a cui questi elementi si riferiscono.
Per garantire la massima fruizione dell'edizione digitale in XML/TEI che riesca a rappresentare in modo ordinato tutti gli elementi, è stato necessario accompagnarla con una piattaforma Web personalizzata, nella quale è possibile consultare le immagini facsimilari del manoscritto (tipicamente giustapposte al testo), la trascrizione interpretativa, l'apparato, la presenza di marginalia e, infine, alcune note di accompagnamento alla lettura. A fianco a questa struttura di partenza, verrà introdotto un sistema di interrogazione lessicale sia di tipo qualitativo - tramite una visualizzazione a colori degli schemi sintattici delle parole nel testo ispirata a *Iliadoscope* (Bouvier, Jambé, Zenzaro, 2020) - sia quantitativo - tramite

ricerche di corrispondenza lessicale e morfologica che rendano conto dei diversi gradi di ibridismo linguistico dei testi del francese d'Italia.

BIBLIOGRAFIA

Benedetti, R. (1998). Le Roman d'Alexandre, riproduzione del ms. Venezia, Biblioteca Museo Correr, Correr 1493. Udine: Roberto Vattori Editore.

Bouvier, D., Jambé, A., & Zenzaro, S. (2020). Iliadoscope, Le devenir numèrique d'un texte fondateur: l'Iliade et le Genavensis Graecus 44 (1.0.0). https://www2.unil.ch/iliade/#/proj.

Clérice, T., Janès, J., Pilla, J., Camps, J-B., Pinche, A., Gille-Levenson, M., & Jolivet, V. Pyrrha, A language indipendent post correction app for POS and lemmatizazion (4.0.0) [Computer sotware], Zenodo. https://doi.org/10.5281/zenodo.2325427.

Gambino, F., & Beretta, A. (2023). Antologia del francese d'Italia: XIII-XV secolo. Bologna: Pàtron.

Gambino, F., Beretta, A., Barillari, S., Ceresato, F., Costa, G., Fassanelli, R., Favaro, M., Fois, J., Guadagnini, E., Guariglia, F., Parodi, M., & Rettore, C. Il francese d'Italia e il progetto "FrIngE". Panoramica generale e casi di studio. Francigena V. 10 (2024), 285-340. https://www.francigena-unipd.com/index.php/francigena/article/view/100.

Guillot, C., Prévost S., & Lavrentiev, A. (2013). Principes d'annotation Cattex09 (Version 2.0), Lyon, École normale supérieure de Lyon, 8 avril 2013, http://bfm.ens-lyon.fr/IMG/pdf/Cattex2009_principes_2.0.pdf.

La Du, M. S. (1937). The Medieval French Roman d'Alexadre, vol. I, Text of the Arsenal and Venice Versions, prepared with an introduction and a commentary by Milan S. La Du. Princeton: Princeton University Press; rist. (1965) New York: Kraus Reprint Corporation ("Elliott Monographs", 36).

Pierazzo, E., & Mancinelli, T. (2020). Che cos'è un'edizione scientifica digitale. Roma: Carocci («Le bussole» 599).

RIALFrI - Repertorio Informatizzato Antica Letteratura Franco – Italiana. Diretto da Francesca Gambino, Università degli Studi di Padova, Dipartimento di Studi Linguistici e Letterari, versione 2.0, 2021. www.rialfri.eu.

TEI (Text Encoding Initiative), TEI P5: Guidelines for Electronic Text Encoding and Interchange, by the TEI Consortium; Originally edited by C.M. Sperberg-McQueen and Lou Burnard for the ACH-ALLC-ACL; Text Encoding Initiative Now entirely revised and expanded under the supervision of the Technical Council of the TEI Consortium, 2023. https://tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf.

Descrivere la catastrofe: documentare la diegesi per la catalogazione di opere distopiche e post-apocalittiche

Luca Paolo Bruno[1], Valeria Stabile[2], Juan Scassa[3], Carmelo Caruso[4] e Ludovica Pannitto[5]
[1] Alma Mater Studiorum Università di Bologna, Italia – lucapaolo.bruno@unibo.it
[2] Alma Mater Studiorum Università di Bologna, Italia – valeria.stabile2@unibo.it
[3] Università degli Studi di Torino, Italia – juanfrancisco.scassa@unito.it
[4] Alma Mater Studiorum Università di Bologna, Italia – carmelo.caruso@unibo.it
[5] Alma Mater Studiorum Università di Bologna, Italia – ludovica.pannitto@unibo.it

ABSTRACT (ITALIANO)
Questo contributo descrive le fasi di progettazione volte allo lo sviluppo e all'implementazione di ENDLIT, un sistema per la catalogazione di opere distopiche e post-apocalittiche per la ricerca umanistica multi-area. Il sistema risponde all'esigenza di identificare elementi comuni, transazioni e interconnessioni culturali (o la loro assenza) tra diverse aree geo-socioculturali in relazione alla produzione di opere di genere distopico e post-apocalittico. Questo si traduce nello sviluppo di un modello dati che non solo è capace di accogliere opere provenienti da contesti di produzione culturali diversi per lingua e locazione geografica e caratteristiche, ma anche la molteplicità della produzione distopica e post-apocalittica in sé, caratterizzata da una sostanziale multi- e trans-medialità. Il risultato è lo sviluppo e l'implementazione di un sistema di catalogazione che pone in enfasi non l'opera come manufatto, bensì come diegesi, come narrazione in un mondo fittizio dagli attributi distinti, riconoscibili e ascrivibili alla produzione distopica e post-apocalittica internazionale.
Parole chiave: distopia; post-apocalisse; database; modellazione.

ABSTRACT (ENGLISH)
Describing Catastrophes: documenting storyworlds for cataloging dystopian and post-apocalyptic fiction.
The following paper provides an overview of the development and implementation of ENDLIT, a system for cataloging and documenting dystopian and post-apocalyptic narrative works. The system is born out of the necessity of identifying common elements, cultural interconnections and transactions (or absence thereof) in dystopian and postapocalyptic fiction between and across different geo-sociocultural areas. This translates to developing a data model that allows incorporation of works originating in production contexts differing in languages, location and characteristics while also accounting for dystopian and post-apocalyptic fiction diffusion across media, beyond the confines of prose narrative and printed matter. The resulting model privileges works of fiction not as physical artifacts, but as storyworlds with distinct, recognizable attributes linked to international dystopian and post- apocalyptic production.
Keywords: dystopian fiction; post-apocalyptic fiction; database; data modelling

1. INTRODUZIONE E OBIETTIVI

ENDLIT descrive un database pensato per interfacciarsi con gli specialisti delle discipline umanistiche, con particolare riferimento agli studi d'area, interessati alla produzione di carattere distopico e post-apocalittico. La distopia è focalizzata nella di mondi dove le condizioni vitali sono una proiezione estremizzata di sviluppi culturali, sociali e politici esperiti come negativi dall'autore vis-à-vis la sua contemporaneità (Moylan 2018). Il post-apocalisse, sebbene focalizzato sul rappresentare gli effetti di una determinata tipologia catastrofe sottoforma di "rovina globalizzata" [globalized ruin], nello scegliere il tipo di evento catastrofico produce anch'esso un avvertimento e un commentario sulla realtà attuale dell'autore (Hicks 2016). Oltre al successo di pubblico, la distopia e il post-apocalisse sono ormai fuori dai confini della narrativa di genere, ottenendo risonanza globale e infiltrandosi in altri settori della produzione culturale. La catastrofe annunciata, imminente o già avvenuta, narrata dalla distopia e dal post- apocalisse è sempre meno avvertimento e sempre più riflesso finzionale di condizioni attuali, collocata nell'ambito del "realismo aumentato" (Deotto, 2018). Questi due settori della produzione culturale globale, nella loro compenetrazione in quanto descrittori di catastrofi avvenute o imminenti, sono accumunate dalla volontà di presentare un avvertimento al lettore, in un'apparente convergenza di idee, concetti, archetipi narrativi in molteplici contesti socioculturali, materializzatasi sotto forma di testi in prosa, audiovisivi, fino ai videogiochi. L'idea di avvertimento tramite rappresentazione diegetica di un evento scatenante che altera negativamente il mondo è la caratteristica che accomuna le opere distopiche e post-apocalittiche. Questa è l'attributo fondamentale delle opere catalogate in ENDLIT.

In questa vasta produzione, le opere che hanno origine nel contesto dei paesi di lingua inglese e di cultura angloamericana, la cosiddetta anglosfera, ne costituiscono la parte più visibile e accessibile, rendendo la globalità della distopia e del post-apocalisse un riflesso dell'egemonia delle culture dell'occidente angloamericano. Il ruolo egemonico dell'anglosfera porta all'oscuramenti di altri contesti di produzione culturale, il cui apporto e dialogo è invece altrettanto significativo. La situazione non è solo un riflesso delle dinamiche di produzione culturali contemporanee, nell'ombra dell'egemonia delle lingue e delle culture dell'anglosfera. È anche e soprattutto un problema di accesso per gli studiosi a fronte di una produzione culturale globale ormai impossibile da seguire per i singoli attori umani (cf. Manovich 2020).
Nell'ambito della ricerca d'area, l'accesso alle informazioni sulle opere oltre i riferimenti bibliografici tradizionali – autore, editore, anno di pubblicazione ecc. – è una problematica sempre più rilevante, che si aggiunge alle sfide d'ordine culturale e linguistico. Se queste ultime sono necessarie per la disamina critica di singoli testi, diviene sempre più impegnativo raffrontare singoli testi con contesti di produzione sempre più disparati, sempre più connessi e sempre più vasti. Anche se rimane comunque possibile stilare, a seconda dell'oggetto e dell'ambito di studi, insiemi di testi iconici, di larga e larghissima diffusione sedimentatisi in canoni più o meno diffusi e riconosciuti, esiste una maggioranza di testi dal pubblico e dalla diffusione ridotti, che però sono comunque in contatto e in dialogo con testi maggiormente conosciuti. Questo è particolarmente riscontrabile nella produzione distopica e post- apocalittica: a fronte di un insieme di testi di ampissima diffusione, spesso con origine nell'anglosfera, è possibile rilevare moltitudini di opere in diverse aree geo-socioculturali, in dialogo e in ricezione con le opere 'globali'. Lo sviluppo di database di ricerca è uno dei modi in cui è possibile realizzare degli strumenti atti a supportare la ricerca in ambiti umanisti alla scala della produzione culturale contemporanea.
Questo si traduce nello sviluppo di un modello dati che non solo è capace di accogliere opere provenienti da contesti di produzione culturali diversi per lingua e locazione geografica e caratteristiche, ma anche la molteplicità della produzione distopica e post-apocalittica in sé, caratterizzata da una sostanziale multi- e trans-medialità. Di conseguenza, l'oggetto privilegiato di documentazione non possono essere le informazioni bibliometriche tradizionali delle opere tradizionali. L'attività di catalogazione deve invece privilegiare la catalogazione di informazioni pertinenti a questioni di studio attuali come l'inclusione della diversità, la sostenibilità e la crisi climatica, al fine di intersecarsi con i dibattiti femministi, postcoloniali e postumani in merito al progresso e all'umanesimo, fornendo strumenti ulteriori per le discipline umanistiche in dialogo con la contemporaneità. In altre parole, deve privilegiare informazioni di tipo diegetico, accogliendo l'opera non come manufatto, ma come narrazione in un mondo fittizio dagli attributi distinti, riconoscibili e ascrivibili alla produzione distopica e post-apocalittica internazionale.
Il progetto si inserisce in un filone di ricerca e sviluppo volto ad attenzionare maggiormente le necessità della ricerca in ambito umanista fornendo informazioni di tipo diegetico. Iniziative esistenti, collegate alla raccolta e collazione di dati esistenti includono il Japanese Visual Media Graph Project (cf. Kacsuk 2023; Pfeffer e Kyriakos 2021; Pfeffer & Roth 2020), il progetto GOLEM (Pannach et al. 2023) e il World Literature Knowledge Graph (Stranisci et al. 2023).

2. STATO DELLE COSE

Attualmente lo sviluppo del database ENDLIT, redatto in inglese, ha prodotto una collezione di dati interrogabile di 281 opere e 190 autori, catalogate con l'apporto di esperi afferenti a tre studi d'area (studi giapponesi, 96 opere e 59 autori, angloamericani, 113 opere e 103 autori, e iberoamericani, 72 opere e 28 autori) che servono come 'campi-pilota'. La forbice temporale selezionata è collocata nell'ultimo decennio (2014-2024), con eccezioni motivate dalla loro importanza relativa al discorso locale o globale. La scelta della forbice temporale è stata operata per testare i criteri del database vis-à-vis l'emergere di una sempre più intensa tendenza all'impiego della distopia e del post-apocalisse all'interno di altri generi come le opere supereroistiche e le opere *young adult*. Applicare il modello dati in questa forbice temporale permette di testare velocemente la comparsa o meno di tratti stabili o non all'interno delle opere distopiche e post-apocalittiche. Il verificarsi di un evento scatenante che funziona da separatore dalla quotidianità dell'autore rispetto alla catastrofe imminente o già verificatasi concorre a risolvere le problematiche collegate alla porosità della produzione distopica e post-apocalittica globale rispetto ad altri generi, che rende problematica la collocazione sulla base di criteri collegati a generi/sottogeneri letterari specifici.
La continua evoluzione della produzione rende di conseguenza maggiormente produttiva la considerazione di distopia e post-apocalisse non come un genere letterario, bensì quello che il critico giapponese Azuma Hiroki definisce come "meta-genere": un insieme di condizioni narratologiche che condizionano lo sviluppo dell'opera prima dell'intento autoriale o del genere stesso (Azuma 2007). Da qui la collocazione dell'opera

narrativa come unità documentale primaria, la cui inclusione dipende dalla presenza di un evento o una serie di eventi (denominato Evento Scatenante o *Trigger Event*), che produce un'alterazione delle condizioni di vita - ambientali, sociali, politiche – che finiscono per risultare nettamente diverse e negative rispetto al momento storico in cui l'opera viene creata. Questa negatività è rappresentata all'interno dell'opera tramite descrizioni di eventi, stili di vita, percezioni dei personaggi in maniera tale che risulti negativa anche alla percezione del fruitore della storia. Questa condizione è pensata per essere indipendente dal media che ospita la narrazione o da criteri collegati all'interpretazione della poetica dell'autore.

3. MODELLO DATI

La catalogazione dell'opera richiede la scelta di una tra cinque possibili tipologie (*type*): romanzo (*novel*); serie TV (*serial*), videogame per opere interattive elettroniche, arte sequenziale (*sequential art*) per i fumetti e lungometraggi (*feature-length*) per le opere audiovisive non seriali di lunghezza superiore ai 60 minuti. A seconda della tipologia dell'istanza, vengono poi identificati (obbligatoriamente) il primo titolo conosciuto dell'opera e l'anno di prima pubblicazione o messa in circolazione, e una lista di titoli alternativi (*Alternative titles*), l'editore (*Publisher*) o la casa di distribuzione (*Distributor*). Per ogni opera è poi possibile inserire all'interno del database una breve sinossi, manualmente compilata dal curatore della scheda o citata da altre fonti. In questo modo al crescere del database sarà disponibile un repository di descrizioni di opere distopiche su cui poter svolgere analisi di tipo linguistico. Ad ogni istanza è poi associata la lista dei suoi autori o creatori, anch'essi presenti come istanze nel database. Infine, l'attributo chiave di ogni opera presente nella collezione dati, l'evento scatenante, viene descritto tramite l'esplicitazione di quattro dimensioni:

- **Agentività (Agency)** descrive l'evento scatenante sulla base dell'agente a cui se ne attribuisce la responsabilità: 'antropica' (a*nthropogenic*) o 'non-antropica' (*nonanthropogenic*) a seconda che la causa sia attribuibile all'azione umana, 'ambigua' (*ambiguous*) e 'sconosciuta' (*unknown*) negli altri casi.
- **Tipologia (Type)** descrive la natura del Trigger event. I possibili descrittori sono quattro. Il parametro 'sovversivo' (*subversion*) nei casi in cui l'alterazione si manifesti a livello personale o di routine quotidiana: casi in cui il cambiamento ha a che fare con il normale corso delle relazioni umane o con azioni quotidiane quali il bere, mangiare, dormire etc. Il parametro 'ambientale' (e*nvironmental*) viene invece impiegato se le condizioni negative derivano da modificate condizioni ambientali o climatiche, il parametro 'bellico' (*war*) nei casi in cui l'evento introduce circostanze di guerra da cui derivano le mutate condizioni. Infine, l'evento viene etichettato come di tipo 'metafisico' (*metaphysical*) se le alterazioni si verificano a un livello così fondamentale da comportare effetti come la scomparsa della morte come cessazione delle funzioni vitali, leggi fisiche come la gravitazione o la continuità cronologica.
- **Nucleo (Kernel)** descrive le modalità in cui l'evento si manifesta. Queste possono essere 'extrasociale' (*extra-societal*), se l'evento è indotto da attori che si posizionano all'infuori della società descritta nell'opera (es. Altri popoli ma anche forme di vita extra-terrestri), 'intrasociale' (*intra-societal*) se al contrario il motore dell'evento è da rintracciarsi all'interno della società stessa oggetto di alterazione (si pensi ad esempio a colpi di stato o elezioni di governi dittatoriali), 'naturale' (*natural*) se l'alterazione si manifesta ad esempio come disastro ambientale o alterazione delle condizioni climatiche, 'non specificato' *(unspecified*) e 'sconosciuto' *(unknown*).
- **Scala (Scale)** descrive le dimensioni dell'influenza che l'evento ha sulla diegesi dell'opera, e queste possono essere 'locale' (*local*) se gli effetti si localizzano in un territorio ben delimitato al cui esterno le alterazioni non sono avvertite, 'continentale' (*continental*) nei casi in cui, anche implicitamente, le alterazioni ricadono su un numero di territori delimitato da confini per lo più fisici (es. Grandi catene montuose oppure oceani), mentre il resto del pianeta non sembra esserne soggetto, 'planetaria' (*planetary*) o 'multi-planetaria' (*multi-planetary*) nei casi in cui le condizioni si presentino alterate per l'intero pianeta o per più corpi celesti, 'universale' (*universal*) nel caso in cui le alterazioni siano percepite nell'intero universo in cui l'opera è ambientate, 'multi-versale' (*multi-versal*) per effetti che si propagano a molteplici universi, ignota nel caso la scala dell'evento non sia specificata

Sono state classificate come eventi con agentività di tipo antropico anche casi in cui la responsabilità dell'evento scatenante è attribuibile a personaggi non-umani ma con caratteristiche antropomorfe tali da

renderli assimilabili ad esseri umani al fine di incorporare intenti ironici e/o satirici all'interno di un'opera. L'evento scatenante di un'opera è poi affiancato e supportato da sistema di etichette (in questo caso, a compilazione libera), raggruppate sotto la dicitura 'eventi seguenti' (aftermaths). Questi eventi, da intendersi comunque come posteriori all'evento scatenante, non sono conseguenze dirette dell'evento scatenante ma contribuiscono alla situazione diegetica descritta come attuale. L'inserimento degli eventi seguenti è libero, non necessario e di numero variabile. Questa catalogazione separata degli eventi seguenti permette una descrizione granulare dell'opera, accogliendo il maggior numero di configurazioni diegetiche possibili in quanti più contesti geo-socioculturali possibili. Come risultato preliminare della catalogazione gli eventi scatenanti delle opere nelle loro quattro dimensioni descrittive presentano la seguente distribuzione:

Dimensione	Sottotipo	Opere d'area giapponese	Opere d'area angloamericana	Opere d'area iberoamericana	Totale
Agentività	Antropica	87	42	80	209
	Non antropica	19	9	10	38
	Ambigua	3	21	2	26
	Sconosciuta	4	0	4	8
Tipologia	Sovversivo	48	39	46	133
	Ambientale	41	8	11	60
	Bellico	17	19	23	59
	Metafisico	7	6	16	29
Nucleo	Extrasociale	11	12	4	27
	Intrasociale	69	42	78	189
	Naturale	16	6	5	27
	Sconosciuto	12	10	7	29
	Non-specificato	5	2	2	9
Scala	Locale	30	24	29	83
	Continentale	7	8	1	16
	Planetaria	73	36	48	157
	Multi-planetaria	1	2	5	8
	Universale	0	0	2	2
	Multi-versale	0	2	10	12
	Ignota	0	0	1	1

Gli autori di un'opera sono catalogati a seconda del loro status come singola persona fisica o come entità commerciale o sociale, e prevedono informazioni anagrafiche come nome e cognome, la data di nascita o di fondazione, l'area geosocioculturale di appartenenza, l'autopresentazione di sé (pronomi personali) e, in vista di ricerche future, provvigioni per accogliere eventuali identità alias.

4. IMPLEMENTAZIONE

L'implementazione del database ha come criterio fondamentale la leggibilità e operabilità sia per operatori umani sia da parte di processi computazionali. A tale scopo è stato redatto un modello per la compilazione assieme a un primo prototipo di codebook. Così facendo si facilita la compilazione e la contribuzione da parte una molteplicità di attori, senza richiedere particolari competenze in ambito informatico. I file sono poi attualmente mantenuti su un repository git (https://github.com/LaboratorioSperimentale/prepararsi-al-presente): al momento dell'inserimento di una nuova scheda o dell'aggiornamento di una scheda esistente, tramite tecniche di continuous integration, le informazioni vengono automaticamente validate

grazie alla libreria cerberus (https://docs.python-cerberus.org/), restituendo un errore all'utilizzatore nel caso in cui la scheda non sia stata compilata correttamente. In questo modo viene preservata la consistenza delle informazioni presenti all'interno della risorsa, lasciando allo stesso tempo la possibilità a chiunque di aggiungere, integrare o modificare le informazioni presenti.
I vari descrittori verranno poi trasformati in filtri di ricerca che, tramite un sito web statico, permetteranno all'utente di esplorare le relazioni presenti tra i dati all'interno della risorsa. La scelta di un'implementazione statica piuttosto che un'architettura più complessa basata sull'interazione client-server è stata dettata essenzialmente da due necessità:

- La possibilità di hostare permanentemente i dati su piattaforme che ne permettano in modo continuato la condivisione, l'arricchimento e l'interrogazione anche al di là dei limiti imposti dal progetto di ricerca: un'architettura basata su modello client-server avrebbe infatti richiesto un impegno, tanto economico quanto di gestione, continuativo nel tempo per poter garantire la permanenza online della risorsa, e questo è reso impossibile dagli attuali modelli di finanziamento della ricerca.
- L'agilità di manutenzione dell'infrastruttura, che richiede, a parità di personalizzazione, minore quantità di codice e minore complessità implementativa grazie alle innumerevoli librerie software a disposizione nell'ecosistema *javascript*.

5. SVILUPPI FUTURI

Il progetto prevede i seguenti sviluppi nel breve periodo:

- Espansione della collezione dati, aprendo la compilazione anche a esperti esterni.
- La pubblicazione del codebook e guida alla compilazione (template per pull request) per il popolamento collaborativo della risorsa secondo i principi dello sviluppo open source.
- Connessione e arricchimento dei dati attraverso collegamento automatico ad entità e database esterni, istituzionali e amatoriali.
- Estrazione automatica di attributi linguistici e testuali dal testo dell'opera tramite pipeline NLP, qualora il testo sia disponibile.
- Sviluppo di piccole domande di ricerca (*Tiny Use Cases*, vd. Freybe et al. 2019) limitate nella scala e nel tempo per testare le capacità della collezione dati nella risposta.

RINGRAZIAMENTI

ENDLIT è parte del progetto di Rilevante Interesse Nazionale (anno 2022) "Prepararsi al presente: nuovi immaginari distopici globali e impegno pubblico", ospitato dall'Alma Mater Studiorum Università di Bologna e dall'Università di Torino e condotto dalla professoressa Paola Scrolavezza in veste di Principal Investigator. Si ringraziano inoltre Ludovica Pannitto, PhD e il dott. Carmelo Caruso del Laboratorio Sperimentale del Dipartimento di Lingue, Letterature e Culture Moderne – LILEC dell'Alma Mater Studiorum Università di Bologna.

RIFERIMENTI BIBLIOGRAFICI

Azuma, H. (2007). Gēmuteki riarizumu no tanjō: dōbutsuka suru posutomodān 2 [Japan's database animals 2: game-like realism]. Tokyo: Kōdansha Gendai Shinsho

Deotto, F. (2018). Il tempo del realismo aumentato. Un'idea di letteratura rivolta al "futuro prossimo". Il Tascabile 16.

Freybe, K., Rämisch, F., & Hoffmann, T. (2019). With small steps to the big picture: A method and tool negotiation workflow. In TwinTalks@ DHN (pp. 13-24).

Hicks, Heather J. (2016). The post-apocalyptic novel in the twenty-first century: Modernity beyond salvage. New York: Palgrave Macmillan.

Kacsuk, Z. (2023, dicembre). Working with vagueness: A pragmatic incremental approach to ontology development in the Japanese Visual Media Graph project. International Forum on Data, Information, and Knowledge for Resilient and Trustworthy Digital Societies (IFDIK 2023), Taipei, Taiwan

Kiryakos, S., & Pfeffer, M. (2021, novembre). Exploring the research utility of fan-created data in the Japanese visual media domain. In International Conference on Asian Digital Libraries (pp. 210-218). Cham: Springer International Publishing.

Manovich, Lev. (2020). Cultural analytics. Cambridge (MA): Mit Press.

Moylan, Thomas (2018). Scraps of the untainted sky: Science fiction, utopia, dystopia. London: Routledge.

Pannach, F., Cheng, L., & Pianzola, F. (2024, dicembre). The GOLEM-Knowledge Graph and Search Interface: Perspectives into Narrative and Fiction. In 2024 Computational Humanities Research Conference, CHR 2024 (pp. 462-471). CEUR Workshop Proceedings (CEUR-WS. org).

Pannach, F., Yang, X., Solissa, N. V., Yu, Z., Van Cranenburgh, A., Van Der Ree, M., & Pianzola, F.

(2024, maggio). The GOLEM Triple Store: A Graph-based Representation of Narrative and Fiction. In B. Sartini, J.Raad, P. Lisena, A. Meroño Peñuela, M. Beetz, I. Blin, P. Cimiano, J. de Berardinis, S. Gottschalk, F. Ilievski, N. Jain, J. Kim, M. Kümpel, E. Motta, I. Tiddi, & J.-P. Töberg (Eds.), ESWC 2024 Workshops and Tutorials Joint Proceedings (Vol. 3749). (CEUR Workshop Proceedings; Vol. 3749). CEUR Workshop Proceedings (CEUR-WS.org).

Pfeffer, M., & Roth, M. (2020, marzo). Japanese Visual Media Graph: Providing researchers with data from enthusiast communities. In Proceedings of the International Conference on Dublin Core and Metadata Applications. Dublin Core Metadata Initiative.

Stranisci, M. A., Bernasconi, E., Patti, V., Ferilli, S., Ceriani, M., & Damiano, R. (2023, ottobre). The World Literature Knowledge Graph. In International Semantic Web Conference (pp. 435-452). Cham: Springer Nature Switzerland.

Un'applicazione pratica per l'edizione digitale di testi agiografici e calendariali

Luca Avellis[1]
[1] Università degli Studi di Bari Aldo Moro, Italia – luca.avellis@uniba.it

ABSTRACT (ITALIANO)

Testi d'uso come martirologi e calendari pongono problemi peculiari alla critica testuale e all'edizione. La loro funzione pratica, legata a esigenze liturgiche e comunitarie, e la connotazione locale li ha esposti a modifiche, interpolazioni e trasformazioni che ne hanno frammentato e ampiamente contaminato la tradizione manoscritta. Questi aspetti rendono complessa la ricostruzione di un archetipo, ponendo la necessità di una ridefinizione di un metodo filologico adeguato. Il *Martyrologium Hieronymianum* (MH) è un caso particolare di tali difficoltà. Strutturato secondo il calendario solare, non segue i criteri formali di un'opera letteraria. Interpolazioni, compendi e contaminazioni rendono incerti i rapporti genealogici tra i testimoni. Le sole edizioni di questo testo, quella diplomatica di De Rossi e Duchesne del 1894 e quella critica di Quentin e Delehaye del 1931, irte di contrasti e traversie, hanno evidenziato i limiti delle tecniche di edizione tradizionali. Questo studio propone un'edizione digitale basata su EVT3, affrontando aspetti fondamentali, quali la gestione delle varianti, la rappresentazione delle contaminazioni e la conservazione delle peculiarità dei testimoni. L'obiettivo è sperimentare delle soluzioni, utilizzando anche l'IA, che possano offrire un modello applicabile ad altri testi simili, caratterizzati da una continua rielaborazione e stratificazione testuale.
Parole chiave: *Martyrologium Hieronymianum*; edizione scientifica digitale; EVT3; codifica TEI

ABSTRACT (ENGLISH)

A practical application for the digital edition of hagiographic and calendrical texts. Texts intended for Practical texts such as martyrologies and calendars pose unique challenges to textual criticism and editing. Their practical function, tied to liturgical and communal needs, along with their local character, has subjected them to modifications, interpolations, and transformations that have fragmented and extensively contaminated the manuscript tradition. These factors make the reconstruction of an archetype particularly challenging, necessitating the redefinition of an appropriate philological method. The *Martyrologium Hieronymianum* (MH) exemplifies these difficulties. Structured according to the solar calendar, it does not adhere to the formal criteria of a literary work. Interpolations, abbreviations, and contaminations create uncertainty in establishing genealogical relationships between the witnesses. The only editions of this text—the diplomatic edition by De Rossi and Duchesne (1894) and the critical edition by Quentin and Delehaye (1931)—marked by controversies and difficulties, have highlighted the limitations of traditional editorial techniques. This study proposes a digital edition based on EVT3, addressing fundamental issues such as the management of textual variants, the representation of contaminations, and the preservation of the witnesses' unique features. The goal is to experiment with solutions, including the use of AI, that could provide a model applicable to other similar texts characterized by ongoing reworking and textual stratification.
Keywords: *Martyrologium Hieronymianum*; Digital scholarly editing; EVT3; TEI encoding

1. INTRODUZIONE

Il *Martyrologium Hieronymianum* (MH) è un repertorio agiografico estremamente complesso, la cui natura mal si adegua alle tradizionali categorie della filologia. Redatto probabilmente tra IV e V sec., nell'Italia del Nord, o forse un secolo dopo in Gallia, per soddisfare esigenze pratiche della comunità cristiana, il MH si struttura come un calendario basato sull'anno solare, con commemorazioni di martiri e santi per ogni giorno dell'anno. Organizzato in latercoli, nella sua forma più completa, riporta in ognuno di questi località e nomi di personalità venerate in quel dato giorno, offrendo quelle che Delehaye definì le coordinate agiografiche, gli elementi minimi necessari a identificare univocamente l'identità di un martire. La forma testuale del MH, composto da liste di nomi, toponimi, attributi e brevi e rare notizie storiche, manca di coerenza letteraria e risente di una tradizione manoscritta frammentata e contaminata, segnata da un massiccio intervento di interpolazioni e abbreviazioni. La tradizione manoscritta, distinta in due famiglie, è composta da *codices pleniores*, versioni estese e dettagliate del testo, e da testimoni *contracti*, nei quali il testo è drasticamente accorciato, talvolta ridotto a poco più di un calendario. La disparità tra i testimoni è

notevole: alcuni manoscritti riportano lunghi elenchi di martiri, completi di località e attributi, altri, meno che un'epitome, si limitano a un solo nome o poco più per latercolo, senza neppure l'indicazione del luogo. Il testo ha ricevuto tra '800 e '900 due edizioni: una diplomatica di De Rossi e Duchesne (1894) e una critica di Quentin e Delehaye (1931). Mentre la prima offriva una trascrizione diplomatica sinottica dei quattro manoscritti ritenuti principali accompagnati da un apparato di varianti rappresentativo della restante tradizione (non priva di compromessi), l'edizione critica successiva, nonostante il tentativo di rappresentare la complessità del testo, si basò su un metodo filologico, quello quentiniano (Avellis, 2019; Avellis, 2020), che, alla ricerca di un testo oggettivo, faticò a gestire la natura fluida del MH. Entrambe le edizioni, pur rappresentando un progresso significativo nello studio del MH, hanno di fatto evidenziato i limiti delle tecniche tradizionali applicate a un'opera tanto peculiare (Avellis, 2022). La varietà testuale è tale,[1] da rendere quasi superfluo il tentativo di redigere uno *stemma codicum*.[2] Eppure, il MH è un testo di straordinaria importanza, fonte autoritativa sia per la storia del cristianesimo delle origini, poiché ha conservato preziose testimonianze su molti martiri altrimenti ignoti, sia per la storia del culto e dell'agiografia, offrendo una sguardo privilegiato sulle pratiche commemorative e sulla loro strutturazione liturgica. La trasmissione testuale del MH documenta l'evoluzione di questi culti nel corso dei secoli, attraverso epoche e contesti geografici differenti, rivelando come tradizioni locali e influenze regionali abbiano modellato il panorama agiografico cristiano europeo. Non solo, dunque, ogni elemento di questo testo contiene una possibile informazione che attende ancora di essere valorizzata, ma ogni testimone tramanda una storia che in gran parte non è stata ancora raccontata e che può essere compresa solo a seguito di una *constitutio* del testo che tenga conto delle complessità dello stesso e dei più recenti apporti metodologici e criteri testuali della filologia.

2. UNA STRUTTURA PARTICOLARE

La struttura del MH è solo apparentemente banale: il testo, preceduto da un titolo non sempre presente e spesso variabile, si articola per mesi, introdotti da formule fisse tratte dalla nomenclatura calendariale (e.g. *Ianuarius habet dies xxx*). Ogni mese è suddiviso in latercoli. Ogni latercolo, introdotto dall'indicazione del giorno cui fa riferimento (*Kl ian.*, *IIII non.*, *Idus*), elenca le memorie martiriali di quel determinato giorno ed è inteso come una unità testuale autonoma: non contiene infatti riferimenti espliciti ad altri latercoli, martiri o località citate nel resto del testo (anche quando tali citazioni sono effettivamente presenti). Anche il latercolo rispetta una struttura precisa: prima l'indicazione della località (dalla maggiore, come una *regio*, alla più piccola come una *civitas*, un *vicus* e per l'area romana anche una *via consularis*), poi seguono il nome o i nomi dei martiri e dei santi menzionati in quel dato luogo, accompagnati a volte da un titolo (*episcopus*, *presbiter*, *virgo* etc.) e molto raramente da altre indicazioni più elaborate, dette non senza una certa enfasi 'storiche' dalla tradizione degli studi (e.g. *una die coronatorum*). A questa struttura si affianca un altro apparato, anch'esso propriamente calendariale, costituito dalle lettere dominicali – e, in alcuni casi, da segni o lettere feriali di uso computistico – che scandiscono rispettivamente il ciclo settimanale e altri cicli liturgici o computistici, e che sono presenti nel margine sinistro di molti testimoni manoscritti in corrispondenza di ciascun latercolo.[3] Questa struttura ideale si piega e deforma, però, continuamente sotto le spinte della tradizione: alcuni latercoli – ad es. quelli realtivi ai primi giorni di maggio e di giugno – arrivano ad annoverare sin dalla *facies* più antica centinaia di memorie, accumulando nomi, titoli e toponimi. Nel corso della tradizione precedente quella a noi giunta, il testo era già stato sottoposto a redazioni diverse, la più recente delle quali, la gallicana è attestata in tutta la tradizione in modo ancora tanto evidente da poter essere parzialmente rimossa se necessario. All'interno del *MH* sono identificabili fonti più antiche: il calendario di Cartagine, il martirologio siriaco e un martirologio africano non meglio definibile, oltre alla *Depositio martyrum* e alla *Depositio episcoporum*. Nel corso della tradizione, a queste fonti si sono aggiunte memorie provenienti da altre fonti, quali calendari locali, *Vitae*, *Passiones* o inserzioni ex novo, introdotte per volontà del redattore di un manoscritto, al fine di registrare nomi ed eventi, ritenuti degni di essere annoverati. Se contro la

[1] I manoscritti tramandano forme corrotte e distorte tanto dei nomi che dei toponimi, spesso così alterate da rendere ardua, se non del tutto impossibile, una loro identificazione.
[2] I testimoni del MH sono 38, tra VIII e XVI sec., sia nella redazione *plenior* che in quella *contracta*, alcuni dei quali frammentari. A questi si aggiungono circa 20 testimoni misti, nei quali cioè il testo del MH è combinato con quello di altri martirologi, come quello di Beda.
[3] Questo apparato è assente in alcuni codici *pleniores* più antichi, come il Wissemburgensis (Wolfenbüttel, Herzog August Bibliothek, Cod. Guelf. 81), mentre è presente in diversi manoscritti appartenenti ad altri martirologi antichi, distinti dal MH. Questo apparato non fu incluso né nell'edizione diplomatica né in quella critica del MH, poiché non ritenuta pertinente al testo originale.

contaminazione, secondo il dettato maasiano, 'non vi è rimedio' (Maas, 2021: 72), le contaminazioni incrociate rendono quasi irrimediabile la ricostruzione del testo, soprattutto nei casi in cui le note marginali si configurano come varianti provenienti da altri testimoni, anziché semplici aggiunte o inserzioni deliberate (Maas, 2021: 15-17), e nel caso del MH, distinguere tra la lettura originale, le aggiunte e inserzioni deliberate e la contaminazione risulta ancora più complesso.[4] Il filologo deve spesso ricorrere a competenze provenienti da altri campi del sapere per affinare il testo (Maas, 2021: 23-25). Era il proposito di uno degli editori precedenti, Hyppolite Delehaye. Quando quest'ultimo cercò di colmare, non sempre in accordo con il collega Quentin, le lacune attraverso la tradizione, questo approccio si dimostrò limitato, andando in diversi casi a sovrainterpretare il dettato originale (Avellis, 2022: 34-35). Essendo evidente che la normale pratica filologica e la comune forma editoriale sono inadatte a gestire un'opera così strutturata tanto nel descriverne le forme, quanto nel raccontarne la capillare storia testuale, ho ritenuto che la sola via editoriale e critica da intraprende in questo caso fosse quella digitale. Quindi una edizione digitale non come una opzione o una ottimizzazione, ma una necessità pratica. L'approccio editoriale adottato si fonda su una posizione intermedia tra la filologia stemmatica di matrice lachmanniana e l'attenzione classificatoria di Quentin, secondo un equilibrio critico già intuito da Paul Collomp (Collomp, 1931: 18-25). In questo contesto, la prospettiva maasiana — sebbene non rigidamente lachmanniana — offre ancora strumenti utili per la descrizione delle relazioni tra testimoni, senza tuttavia condurre a una *restitutio*.[5] Lungi dal ridurre la complessità della tradizione a un modello unilineare, l'edizione assume i testimoni nella loro singolarità, documentandone le divergenze e rinunciando a un tentativo di unificazione testuale se non come ipotesi *in itinere*.

3. UNA EDIZIONE CRITICA DIGITALE DEL *MH*

Di fronte a una struttura così peculiare, avevo caldeggiato il proposito di progettare una applicazione proprietaria che contemplasse nativamente una segmentazione del testo speculare alla struttura precedentemente descritta e tenesse conto della stratificazione e differenziazione del testo. Ma le difficoltà tecniche di iniziare una applicazione da zero, i tempi di sviluppo, il vincolo, che avrei dovuto rispettare di far collimare tutti i testimoni nelle segmentazioni predisposte – cosa non sempre possibile – e la condivisione dell'idea presentata da Chiara Martignano (Martignano, 2021; Martignano, 2024) della necessità di sviluppare un modello teorico, basato su metodologie filologiche consolidate, da formalizzare in un'ontologia per garantire interoperabilità e riutilizzo degli strumenti software nelle edizioni digitali,[6] mi hanno portato a risolvermi all'uso di un'applicativo già operativo. La scelta è caduta quasi naturalmente su EVT.[7] Nella sua struttura, progettata per essere il più possibile adattabile a una varietà di esigenze tipiche di un testo critico, EVT mi consentiva la gestione delle diverse versioni di un medesimo testo — diplomatica, interpretativa e critica — all'interno dello stesso file XML.

La prima scelta da operare è stata sulla gestione dei testimoni. EVT3, permette, infatti, due approcci distinti (Cacioli & *alii*, 2022: 216): un approccio centralizzato che gestisce le varianti testuali dei diversi testimoni all'interno di un unico file XML attraverso una lista dei testimoni;[8] o, in alternativa, un approccio

[4] Il termine *Urfassung* non è impiegato da Paul Maas nella *Textkritik* e non rientra nel suo lessico tecnico. Esso viene talvolta usato in ambito filologico per designare una redazione primitiva di un'opera, spesso assimilata all'autografo o all'intenzione autoriale. Diversamente, l'*Archetyp*, secondo la definizione di Maas, è l'esemplare dal quale ha avuto inizio la prima divisione della tradizione manoscritta e rappresenta un testo già suscettibile di errori e non coincidente con l'originale (Maas, 2021: 9). Maas raccomanda espressamente di riservare il termine *archetipo* a questo livello e di non applicarlo ad altri anelli intermedi della tradizione.

[5] L'opzione bedieriana - la scelta di un unico codice base - risulterebbe in questo contesto metodologicamente insostenibile.

[6] Su questa stessa linea sta operando il gruppo coordinato dal prof. Dionysos Benetos, National and Kapodistrian University of Athens (Benetos & Pappadaki, 2024). Ringrazio gli autori per avermi anticipato il testo del loro contributo in corso di pubblicazione.

[7] Ho concepito l'ECD del MH basandomi sulla versione EVT2, allora disponibile. Tuttavia, con il rilascio di EVT3 il 7 ottobre 2024 (https://visualizationtechnology.wordpress.com/2024/10/07/evt-3-beta-available-for-download-and-testing/), ho ritenuto opportuno migrare il lavoro svolto fino a quel momento sulla nuova versione. Le modifiche iniziali hanno riguardato principalmente la personalizzazione dei file di configurazione, assenti nella versione precedente e rappresentativi di alcune delle nuove funzionalità introdotte.

[8] Si tratta dell'elemento <listWit>, che contiene i sottoelementi <witness> a ciascuno dei quali viene attribuito un xml:id univoco. Attraverso i marcatori <app>, <lem> e <rdg> è possibile descrivere le varianti testuali. EVT3 ha inoltre introdotto una nuova funzione, <changesView>, che permette di tracciare le variazioni testuali e di rappresentare i processi di riscrittura, correzione e, di particolare interesse nel caso in oggetto, di aggiunta di contenuti, offrendo la possibilità di sincronizzare la visualizzazione delle modifiche con altri livelli testuali e permettendo un confronto diretto ad es. tra il testo "iniziale" e le versioni "successive". Sebbene di non facile gestione tecnica, ho provato a utilizzare entrambi i livelli di marcatura: sia un apparato per varianti significative provenienti da altri codici, sia per le variazioni all'interno dello stesso testimone. Tuttavia questo richiede delle modifiche sui file di configurazione. La visualizzazione in

modulare dove ogni testimone è descritto in un file XML separato. Essendo il MH attestato da testimoni spesso estremamente difformi tra loro, la soluzione più adeguata mi è sembrata quella di proporre un'edizione per ciascuno dei manoscritti del MH. I vantaggi sono innazitutto pratici: l'organizzazione in un unico file, infatti, sebbene offra di base una maggiore coerenza, si è mostrata di difficile gestione. L'inserimento delle diverse varianti, a cui poi va sommata la gestione degli altri marcatori per note, nomi, toponimi, elementi storici, ecc.,[9] si è rivelato eccessivamente lungo – un test su un solo mese ha infatti comportato diverse centinaia di righe di codice –[10] con un alto rischio di sovrapposizioni di sezioni e nidificazioni complesse. Inoltre, trattandosi di un compito notevolmente articolato, questa modalità consente di suddividerlo tra più operatori, permettendo a ciascuno di lavorare in modo autonomo su un singolo testimone. La scelta è stata, però, guidata anche da una precisa posizione teorico-scientifica. Ritengo che, pur considerando importante l'obiettivo di una ECD del MH, questo non debba rappresentare l'unico scopo. È altrettanto importante produrre edizioni diplomatiche di ciascuno dei testimoni, al fine di valorizzare il ruolo che ciascun manoscritto ha avuto come testimone unico di una specifica fase della diffusione di un corpus di culti in un determinato luogo e tempo, aspetto che nella gestione delle varianti testuali all'interno di un unico file XML semplicemente va persa. La struttura a testimoni autonomi permette, inoltre, di utilizzare con maggiore facilità sia la visualizzazione della stratificazione del testo – elemento intrinseco alla natura stessa dei testimoni del MH e altrettanto significativo – sia un apparato di annotazioni secondo i principi TEI e predisposta per una strutturazione interrogabile in JSON-LD. Considero l'ECD del MH come 'opera' necessaria, ma anche, più di quanto accada in altre edizioni, come 'opera' in divenire. A differenza di un'edizione critica tradizionale, che di solito richiede aggiornamenti su una scala temporale di decenni, l'edizione del MH, per la sua natura di fonte autoritativa, sembra destinata a subire modifiche e correzioni con maggiore frequenza. Questo processo - per approssimazioni successive - diventa necessario ogni volta che uno dei migliaia di nomi, spesso privi di una storia definita, viene identificato, trovando così il proprio posto nella ramificazione di una memoria liturgica e di una prassi devozionale.

A tal riguardo, una funzione di EVT3 particolarmente utile per l'analisi storico-agiografica è la possibilità di creare liste strutturate. Questa opzione permette di categorizzare gli elementi testuali classificandoli come entità nominali – persone, luoghi, eventi, ecc. – e di filtrarli in base a specifiche esigenze. Si tratta di una funzione particolarmente rilevante per un testo composto quasi interamente da nomi propri, toponimi e attributi. Inoltre, le entità possono essere arricchite con elementi descrittivi che ne definiscono in modo più dettagliato le caratteristiche specifiche.[11] La lista delle categorie è facilmente estendibile, consentendo l'aggiunta di nuove voci nei file di configurazione edition_config_nome.json (sezione entitiesSelectItems).[12]

---

modalità 'Reading Text' si è dimostrata più stabile e fruibile, mentre in modalità 'Documental' ho riscontrato errori di visualizzazione, dovuti a tutti gli effetti al mio tentativo di 'forzare' il sistema.

[9] Non ho ritenuto, in questo caso, di inserire anche la funzione <changesView>.

[10] Si intende ovviamente in una formattazione multi-line.

[11] Ad esempio, il marcatore <person> – ma lo stesso principio vale per qualsiasi altro elemento analogo – popola la lista <listPerson> e può essere arricchito con diversi attributi. Ho inserito, con diverse prove, <name> (nome), <place> (luogo del martirio o di venerazione), <sex> (genere), <role> (carica, p. es. vescovo, presbitero, ecc.), <day> (giorno commemorativo), <daytype> (tipologia della commemorazione) e <note> (da distinguere dal tag <note> impiegato per annotazioni extratestuali). Le entità agiografiche codificate in liste strutturate (ad es. <listPerson>, <listPlace>) consentono, per ciascun elemento, la costruzione di una scheda descrittiva completa. Ho inoltre provato ad associare a ciascun soggetto un identificatore esterno stabile mediante il tag <idno type="URI">, come quelli forniti dal portale *Cult of Saints in Late Antiquity* (Oxford, http://csla.history.ox.ac.uk/). Sebbene EVT3 non generi errori evidenti, non è in grado di interpretare tali collegamenti in modo visibile, e, probabilmente poiché non prevede la presenza di elementi figli, causando problemi nell'interfaccia al momento della selezione dall'elenco di nomi. Ho tentato anche di richiamare nel testo l'URI attraverso l'attributo @ref, inserito nei nodi <person> o <persName>, secondo la prassi prevista dallo standard TEI. Tuttavia, nella versione attualmente disponibile, EVT3 non supporta ancora nativamente la visualizzazione attiva di tali URI: l'inserimento dell'attributo @ref – ad esempio in <person … ref="http://..."> – non produce alcun effetto visibile nell'interfaccia, pur non compromettendo il corretto funzionamento dell'edizione. L'informazione viene comunque mantenuta all'interno della struttura semantica del documento TEI. È stato possibile renderla visibile in forma testuale attraverso un attributo, come <link>, che – analogamente agli altri – è visualizzato ma non attivo (in sostanza, l'utente dovrebbe copiarne il contenuto e incollarlo nel browser). L'attributo @target, invece, non risulta funzionante in questo contesto.

[12] Questo lavoro viene svolto all'interno del file XML principale, ma rimane separato dal testo dell'edizione critica. Le liste sono popolate in modo indipendente compilando le singole entità nominali a mano, mentre gli xml:id corrispondenti sono inseriti nel corpo del testo. Si tratta di un lavoro complesso e impegnativo: ogni elemento, <person>, <place>, ecc., ha un'ampiezza minima di 8 righe di codice, ma a seconda degli attributi implementati può facilmente superare la dozzina, che moltiplicate per le migliaia di nomi presenti nei codici *pleniores* sviluppa diverse migliaia di righe di codice. Ho provato, ammetto con scarso successo, ad automatizzare il lavoro servendomi di un file excel che già possedevo con un limitato elenco alfabetico di nomi di martiri del MH, ampliato con i dati di cui avevo bisogno (come xml:id, sesso, ruolo, ecc.), attraverso uno script python e le librerie pandas. Lo script restituisce un file xml ben formattato e, se

EVT3 presenta di contro alcune limitazioni nella resa della *dispositio textus* della versione diplomatica. Sebbene sia possibile personalizzare l'aspetto grafico dell'edizione diplomatica per rispettare rientri, spaziature e font,[13] il sistema non supporta nativamente la modifica dei font o delle dimensioni del testo all'interno dello stesso paragrafo tramite semplici marcatori XML. Tali variazioni richiedono interventi attraverso modifiche al CSS, rappresentando un limite per un'edizione diplomatica in cui la posizione del testo all'interno dello specchio di scrittura è significativa.[14]
Il risultato atteso è un'edizione digitale del MH che comprenda una versione critica dell'archetipo, arricchita da varianti, note e liste analitiche di martiri, santi e luoghi, affiancata dalle edizioni diplomatiche di tutti i testimoni, inclusi quelli misti e incompleti, ciascuno dei quali rappresenta una fase unica della diffusione e della trasformazione dei culti cristiani in specifici contesti storici e geografici. L'ECD, concepita come un archetipo – ideale -, sarà progressivamente affinabile e arricchita con varianti e note. Le edizioni diplomatiche, invece, saranno progettate per garantire la massima fedeltà al testo originale, grazie anche alla funzione <changesView>, che permetterà di visualizzare le riscritture, le aggiunte e le modifiche che hanno caratterizzato la trasmissione del MH. In questo modo, sarà possibile restituire una rappresentazione dettagliata delle dinamiche storiche e filologiche associate a ciascun testimone.[15]

4. QUESTIONI E PROSPETTIVE

L'edizione del *MH* ha storicamente rappresentato un compito complesso per la critica testuale. Come rilevato in questo contributo, l'introduzione di strumenti digitali - come EVT - ha reso possibile affrontare questioni critiche altrimenti non risolvibili con i soli metodi tradizionali. Tali strumenti consentono di gestire volumi significativi di dati, di confrontare varianti testuali in modo sistematico e di rappresentare visivamente la molteplice complessità dei testi. L'edizione è attualmente in fase di sviluppo e si basa su metodologie consolidate, che hanno già prodotto risultati incoraggianti, nonostante le difficoltà operative e il notevole impegno richiesto. Tra gli obiettivi futuri emergono il miglioramento della gestione dei testimoni e l'ampliamento dell'accessibilità delle fonti per la comunità accademica, con particolare riferimento agli studi storico-agiografici. La fruizione dell'edizione è pensata in modo modulare e stratificato: il lettore deve poter esplorare i singoli testimoni in formato diplomatico, con la possibilità di visualizzare riscritture, integrazioni e interventi redazionali tramite il sistema <changes>. Sarà inoltre possibile consultare la versione critica con visualizzazione delle varianti, e accedere a liste interattive di martiri, luoghi ed eventi, filtrabili secondo diversi criteri. In prospettiva, potrà essere implementata una navigazione sincronizzata tra testo critico, edizioni diplomatiche e facsimili digitali, con link diretti tra varianti, testimoni e porzioni corrispondenti del facsimile digitale. Sarà accessibile da browser e progettata per supportare attività di ricerca filologica, storica e prosopografica, rivolgendosi a studiosi, studenti e specialisti del dominio agiografico. In questo contesto, l'adozione di un'interfaccia WYSIWYG potrebbe rappresentare un passo

necessario, può essere ampliato, a condizione di elaborare nuovamente l'intera tabella a ogni modifica. Tuttavia l'operazione rimane complessa e non risolve la difficoltà di inserire manualmente l'attributo xml:id nel corpo del testo, aspetto non secondario di fronte al gran numero di nomi e all'elevata frequenza di omonimie. Un approccio alternativo, potenzialmente più pratico ma che non ho ancora sperimentato, potrebbe consistere nel processo inverso, in cui, partendo dal testo XML compilato, si estrapolino gli xml:id per generare direttamente la ListPerson. Quest'ultima dovrebbe comunque essere completata in un secondo tempo con i relativi attributi. L'IA potrà semplificare la creazione e l'arricchimento delle liste: somministrando lo stesso file excel utilizzato con lo script python a chatGPT 4o con un numero estremamente limitato di indicazioni ha saputo classificare le entità nominali presenti nei manoscritti (nomi, toponimi, titoli, ecc.) con un errore di poco superiore del 15% circa, dovuto soprattutto alla forma corrotta dei nomi sottoposti, ma restituendo un xml sostanzialmente corretto. Una problematica analoga si presenta anche per le altre liste, in particolare quella dei toponimi, <listPlace>. Un'analisi specifica andrebbe dedicata all'attributo 'relation', che consente di stabilire un collegamento tra due elementi del testo secondo un vincolo definito dall'utente. Questo attributo si basa su un sistema di puntamento che utilizza l'xml:id per designare i soggetti coinvolti nella relazione. La flessibilità dell'attributo risiede nella possibilità di definire il tipo di relazione attraverso il parametro type, che può descrivere legami di varia natura.

[13] EVT3 consente la modifica dei font primari e secondari tramite i parametri mainFontFamily e secondaryFontFamily nel file ui_config.json.

[14] Anche la riproposizione di una disposizione del testo in colonne, tipica dello stile calendariale, risulta complessa: è possibile simulare tale formato con interventi mirati nel file XML, utilizzando una struttura basata sugli elementi <table>, <row> e <cell> per creare un layout tabellare.

[15] Durante il lavoro di preparazione, ho compreso che l'allestimento di un'ECD del MH, accompagnata dalle edizioni diplomatiche dei diversi testimoni, non sarebbe stato sufficiente. Risulta infatti necessario includere anche le quattro edizioni diplomatiche di De Rossi-Duchesne e l'edizione critica di Quentin-Delehaye. Questo permetterebbe non solo di correggere alcuni errori e compromessi dovuti ai limiti del supporto cartaceo - come l'organizzazione tipografica dell'edizione critica in cui le due famiglie testuali coesistono nella stessa colonna -, ma anche di rispondere all'importanza storica di queste edizioni. Per oltre un secolo, infatti, queste edizioni hanno rappresentato i principali riferimenti accademici. La loro pubblicazione digitale appare quindi indispensabile per garantire un accesso diretto e sistematico alle versioni testuali che hanno modellato la tradizione critica e la ricezione moderna del MH.

significativo per semplificare la gestione e l'imputazione dei contenuti, riducendo il ricorso diretto alla codifica in TEI e favorendo una maggiore efficienza nell'elaborazione e nella pubblicazione dei testi.[16] Infine, è utile sottolineare che il presente contributo si configura come un caso di studio limite. Il MH e, in misura ancora più ampia, i calendari antichi non rappresentano però gli unici esempi di testi caratterizzati da una struttura e una varietà comparabili. Oltre agli altri martirologi, in particolare quello di Beda,[17] si possono menzionare i necrologi e gli obituari, per i quali Jean-Loup Lemaître ha catalogato alcune migliaia di esempi relativi alla sola area francese (Lemaître, 1980). A questi si aggiungono i *Tacuina sanitatis* e altri testi che, pur non presentando una struttura intrinsecamente calendariale, risultano spesso organizzati secondo un calendario liturgico. Si tratta di opere accomunate da una varietà di contenuti distribuiti all'interno di una struttura stabile, come i passionari, i leggendari e, spostandosi oltre l'ambito latino, le catene, i menologi e i sinassari. Questi testi, nella maggior parte dei casi, non hanno ancora beneficiato di un'edizione critica *in extenso*, ma rappresentano un potenziale ricco di informazioni e di possibili scoperte. Tale lacuna editoriale rende questi documenti un campo di studio promettente per future ricerche filologiche e storico-culturali oltre che un banco di prova significativo per le metodologie dell'umanistica digitale.

RINGRAZIAMENTI

In questo contributo ho incluso parte dei risultati delle mie ricerche e delle sperimentazioni svolte nell'ambito del progetto "La piattaforma digitale intelligente per la ricostruzione dell'identità europea attraverso lo studio della gestione del tempo sacro e profano nell'Europa antica", Codice progetto: H95F21001360006. Questo progetto è stato finanziato nell'ambito dell'iniziativa REACT-EU (Recovery Assistance for Cohesion and the Territories of Europe) con risorse del Fondo Europeo di Sviluppo Regionale (FESR) ed è stato coordinato dalla Prof.ssa Immacolata Aulisa presso l'Università degli Studi di Bari Aldo Moro. Desidero inoltre ringraziare il prof. Nicola Barbuti per la preziosa, sincera e sempre attenta guida.

BIBLIOGRAFIA

Avellis, L. (2019). Il ruolo di Dom Henri Quentin nella filologia del '900. Vetera Christianorum, 56, 31-45.

Avellis, L. (2020). Il metodo ecdotico di Quentin: gli Essais de critique textuelle (Ecdotique). Classica & Christiana, 15, 291-310.

Avellis, L. (2022). Il Martyrologium Hieronymianum. Storia critica della critica di un testo. Invigilata Lucernis, 44, 29-42.

Benetos, D., & Papadaki, A. (2024). Well-defined Critical Editing: The 'Manuscribus' Platform for Manuscript Transcription, Annotation, and auto-Collation. Invigilata Lucernis, 46, cs.

Cacioli, G., Cerretini, G., Di Pietro, C., Maenza, S., Rosselli Del Turco, R., & Zenzaro, S. (2022). There and back again: what to expect in the next EVT version. AIUCD 2022 - Proceedings. Culture digitali. Intersezioni: filosofia, arti, media. Ciracì, F., Miglietta, G., & Gatto, C. (Eds.). 212-217. https://dx.doi.org/10.6092/unibo/amsacta/6848.

Collomp, P. (1931). La critique des textes et l'édition des auteurs grecs et latins. Paris : Les Belles Lettres.

Lemaître, J.L. (1980). Repertoire des documents necrologiques francais. Paris : Academie des inscriptions et belles-lettres.

Maas, P. (2021). La critica del testo. Traduzione a cura di Giorgio Ziffer. Roma : Edizioni di Storia e Letteratura.

Martignano, C. (2021). A Conceptual Model to Encourage the Development and Reuse of Apps for Digital Editions. Umanistica Digitale, 10, 71–88. https://doi.org/10.6092/issn.2532-8816/12620.

Martignano, C. (2024). Critical Edition Ontology: A Conceptual Model for Digital Critical Editions. Umanistica Digitale, 17, 71–94. https://doi.org/10.6092/issn.2532-8816/19469.

[16] Immagino qualcosa di simile a Xeditor o Oxygen XML Editor. L'interfaccia WYSIWYG proposta dovrebbe essere ottimizzata per rispondere alle specifiche richieste dall'implementazione in EVT3, garantendo la compatibilità con la struttura XML-TEI utilizzata per le edizioni critiche digitali.

[17] I martirologi storici successivi, come quelli di Adone e, in particolare, di Usuardo, si caratterizzano per un grado più elevato di 'letterarietà', che li ha resi meno aperti a interventi esterni. Questi testi tendono a configurarsi come opere letterarie a tutti gli effetti, sottoposte così alle norme canoniche di trasmissione tipiche della tradizione letteraria.

Embracing flexibility: new EVT features for critical editing, accessibility and inclusivity

Roberto Rosselli Del Turco[1], Davide Cucurnia[2], Marina Buzzoni[3]
[1] Università di Torino, Italia - roberto.rossellidelturco@unito.it
[2] Università di Pisa, Italia - davide.i.cucurnia@gmail.com
[3] Università Ca' Foscari Venezia, Italia - mbuzzoni@unive.it

ABSTRACT (ENGLISH)
After a few years of fluctuating progress, caused by difficulties in maintaining continuity within the development team, EVT was given new impetus by a collaboration with a new software company and with an unexpectedly large number of digital edition projects. The requests from various research groups have led to the belief that the key concept on which to build the new features should be "flexibility/adaptability", both to ensure better support for critical editions and to offer more inclusive editions. The paper will discuss the design and development of these features in light of the digital projects in which the authors are involved.
Keywords: digital scholarly editing; digital philology; accessibility; inclusivity; critical edition.

ABSTRACT (ITALIANO)
Embracing flexibility: new EVT features for critical editing, accessibility and inclusivity. Dopo qualche anno di andamento altalenante, dovuto alla difficoltà di dare continuità al team di sviluppatori, EVT ha ricevuto nuovo impulso grazie alla collaborazione con una nuova ditta software e con un numero inaspettatamente elevato di progetti di edizioni digitali. Le richieste dei vari gruppi di ricercatori hanno portato a ritenere che il concetto chiave su cui costruire le nuove funzionalità debba essere quello della "flessibilità/adattabilità", sia per garantire un migliore supporto alle edizioni critiche, sia per offrire edizioni più inclusive. Il paper tratterà della progettazione e dello sviluppo di tali funzionalità, alla luce dei progetti digitali in cui gli autori sono coinvolti.
Parole chiave: edizione scientifica digitale; filologia digitale; accessibilità; inclusività; edizione critica.

1. INTRODUCTION: WHERE WE ARE WITH EVT DEVELOPMENT
The development of EVT has been a bit bumpy over the last few years: as explained in Cacioli *et al*. 2022, the change of development framework from AngularJS to Angular was essential to the birth of EVT 3, but as a first effect it meant a hiatus of more than two years just to reach a basic functionality comparable to that of EVT 2. The alpha version released in late 2022 thus represents a significant step forward for the future of the project, but its features are very limited: more of a demonstration prototype to be tested than something actually usable. The beta version, released in autumn 2024, is much more complete in terms of available features, but still has gaps that are being filled. The road to 1.0 is still a long one, and 2025 will be crucial.

The 2023-24 biennium also saw a revolution in the development team: the core group of developers associated with EVT's origins, our "old guard", gradually stepped aside, an event that was not painless, considering the network of human relationships and shared experiences that accumulated during the ten years of EVT development, but entirely understandable given work, family, etc. commitments that have become increasingly dense over the years. The new developers who joined the team between 2022 and 2024 were funded by grants, but this temporary relationship did not lead to their integration into the team when the grants expired. The collaboration with the software company SilentWave[1] starting in 2024 is probably the most important event for the future of EVT.

Not all is negative in 2022-24, though, as completely new and potentially very useful features have been introduced for a 1.0 version of EVT 3. These include authorial philology, thanks to the project on Saba's 1919 *Canzoniere* (Buzzoni 2024; Buzzoni *et al*. 2024), experimentation with TEI data processing and user annotation (Cacioli *et al*. 2022), and a major refinement of the digital facsimile, with the reintroduction of

[1] Home page: https://www.silentwave.eu/. See this page for EVT support: https://www.silentwave.eu/it/evt.

the bookreader view and a new view for comparing different images. A general improvement in support for critical editions has also been initiated in 2024 thanks to a collaboration with the PRIN RETI project.[2]

Before these new features could be successfully implemented, however, the new development team encountered a rather serious problem: EVT 3's XML parser was designed, and works well, for progressive handling of the TEI document, creating an in memory data structure as the file is read. This allows smooth handling of, for example, lists of named entities, and in general any information involving elements placed within the `<body>`. Unfortunately, this approach has proved less suitable for handling stand-off-type markup, which is typically placed in the `<back>` of a TEI document. Mainly for this reason, but also on the basis of other technical considerations, the SilentWave team proposed to rewrite the parser in such a way as to have a more flexible and powerful tool capable of handling all situations in which some text-related information is located anywhere in the TEI document (or even in another separate TEI document).

2. NEW FEATURES FOR THE *LEGES LANGOBARDORUM* PROJECT

This paper discusses new features introduced in the version developed to meet the needs of the *Leges Langobardorum* project, or otherwise related to the objectives of this project (Buzzoni & Rosselli Del Turco 2024).

2.1 Critical edition support

One of the new features introduced in EVT 3 is support for the double-end-point attached method (hereafter depa) for linking the critical apparatus to the text. As is well known, depa works both as inline markup and in stand-off mode. While both modes are supported, the implementation has focused particularly on stand-off markup because it offers an important advantage: it allows the critical text to be left "clean" so that another markup layer of any kind can be added. This feature will be particularly useful in the context of the *Leges* project, as it will allow named entities and especially Lombard terms to be marked up without fear of tag overlap.

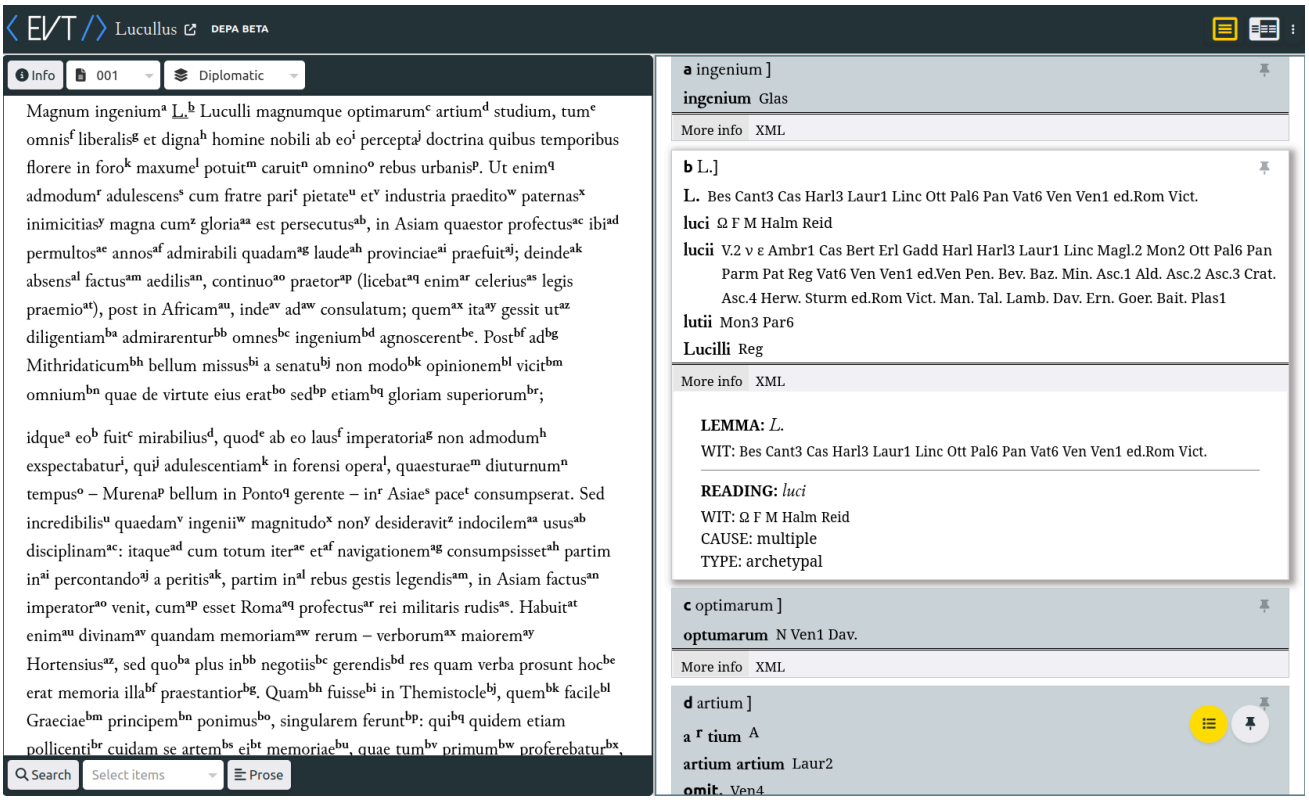


Figure 1. The TEI double-end-point attached method implementation for the RETI project

[2] PRIN 2022 "RETI - REndering Texts and Images. Digital Scholarly Editions with Edition Visualization Technology". See the paper "Il progetto RETI (REndering Texts and Images): metodologia e primi risultati" by C. Barbero, M. Di Franco, F. Lazzerini and A. Persia in this volume of the proceedings.

EVT will support the standard depa method proposed in the TEI *Guidelines*,[3] which is based on the positioning of empty `<anchor/>` elements, each given a unique identifier by the `xml:id` attribute, and used in pairs (hence the name "double endpoint") to delimit text spans that are to be associated with apparatus entries. However, the method proposed by the EVT developers and used by the RETI project is slightly different: it is still based on a stand-off approach, but the way it identifies text spans does not depend on the insertion of `<anchor/>` elements, but on the identifiers automatically assigned to each word in the text.[4]

```
<p n="1"><seg n="1" xml:id="Luc-001">
      <w xml:id="Luc-001-1">Magnum</w>
      <w xml:id="Luc-001-2">ingenium</w>
      <w xml:id="Luc-001-3">L.</w>
      <w xml:id="Luc-001-4">Luculli</w>
      <w xml:id="Luc-001-5">magnumque</w>
      <w xml:id="Luc-001-6">optimarum</w>
      <w xml:id="Luc-001-7">artium</w>
      <w xml:id="Luc-001-8">studium</w>,
```

This method lets the encoder skip the tedious addition of `<anchor/>`s to the critical text, and you also don't need the `@to` attribute in the `<app>` if the variant corresponds to only one word:

```
<listApp n="Luc-001">

   <app from="#Luc-001-2">
      <rdg type="lectio.sing" cause="palaeographicConfusion"
      wit="#Glas">ingen<add hand="#Glas.hand.c">i</add>um </rdg>
   </app>

   <app from="#Luc-001-3">
      <lem
         wit=" #Bes #Cant3 #Cas #Harl3 #Laur1 #Linc #Ott #Pal6 #Pan
#Vat6 #Ven #Ven1 #ed.Rom #Vict."/>
      <rdg type="archetypal" cause="multiple" wit="#Ω #F #M #Halm
#Reid">luci</rdg>
      <rdg type="editorial"
         wit="#V.2 #ν #ε #Ambr1 #Cas #Bert #Erl #Gadd #Harl #Harl3
#Laur1 #Linc #Magl.2 #Mon2 #Ott #Pal6 #Pan #Parm #Pat #Reg #Vat6 #Ven #Ven1 #ed.Ven
#Pen. #Bev. #Baz. #Min. #Asc.1 #Ald. #Asc.2 #Asc.3 #Crat. #Asc.4 #Herw. #Sturm
#ed.Rom #Vict. #Man. #Tal. #Lamb. #Dav. #Ern. #Goer. #Bait. #Plas1"
         >lucii</rdg>
      <rdg type="polygenetic" cause="multiple" wit="#Mon3
#Par6">lutii</rdg>
      <rdg type="lectio.sing" cause="multiple" wit="#Reg">Lucilli</rdg>
   </app>
```

[3] TEI Consortium, eds. "13.2.2 The Double End-Point Attachment Method." *Guidelines for Electronic Text Encoding and Interchange*. 4.9.0. 24 January 2025. https://www.tei-c.org/release/doc/tei-p5-doc/en/html/TC.html#TCAPDE.
[4] This task can easily be performed by a script in Python or XSLT.

```
            <app from="#Luc-001-6">
               <rdg type="polygenetic" cause="diction" wit="#N #Ven1
#Dav.">optumarum</rdg>
            </app>
```

Using this method, all you need is an `<anchor/>`, so that you can link an `<app>` to it, if the variant or textual phenomenon to be described is not present in the critical text.

2.2 Editorial conventions

This is a feature that was introduced in an experimental form in the very first versions of EVT 3, but only fully developed during the last development cycle (after the beta 1 release). The principle behind this feature is very simple: each philological tradition has its own editorial conventions (see, among many others, Bausi et al. 2015; Roelli et al. 2020), so to introduce fixed and unchangeable solutions in EVT is to propose a way of publishing DSE that only partially satisfies, or completely displeases, philologists other than the one who set these solutions. The remedy introduced in EVT is an additional configuration file, `editorial_conventions.json`, in which the typographical rendering for the desired type of phenomenon can be specified.

In EVT 2 it was already possible to use custom CSS rules to highlight and format the text contained by specific elements. This is particularly useful when these elements are ignored by EVT, for example the content of `<supplied>` is rendered "as is", with no special handling of its appearance. To specify that its content should be inserted between square brackets, two simple CSS rules can be used in the `custom-styles.css` file:

```
.supplied:before {content: "["}
.supplied:after {content: "]"}
```

However, this approach is not perfect, firstly because when a TEI element is actually handled by EVT, there may be a conflict with the internal stylesheets, which means that the custom CSS rules may fail altogether. It would also require each editor to create their own list of editorial elements and custom CSS rules to render them on screen. In EVT 3, the `editorial_conventions.json` file groups together all the solutions devised for editorial markup, and makes it possible to check that they are appropriate for the current edition, and to change them if necessary. Using a separate configuration file, it is also possible to create pre-defined sets corresponding to the main philological traditions (e.g. a set for Leiden style editions, one for medieval Latin texts, etc.) and to create a web front-end to allow any editor to modify and adapt them without knowing CSS syntax. This is how `<supplied>` might be handled for all edition levels:

```
  "editSuppliedText": {
      "markup": {
          "element": "supplied"
      },
      "layouts": {
          "diplomatic": {
              "style": {
                  "pre": "[",
                  "post": "]",
                  "background-color": "#e7b0e8"
              }
          },
          "interpretative": {
              "style": {
                  "pre": "[",
                  "post": "]"
                   }
```

```
            },
            "critical": {
                "style": {
                    "pre": "[",
                    "post": "]"
                }
            }
        }
    },
```

This feature is an important element of flexibility, as it does not impose a single set of conventions applicable to all philological traditions, but allows for choice and customisation according to the needs of the philologist. Also note that using this feature doesn't prevent the editor from adding custom CSS rules to further customize the appearance of the text, allowing for even more flexibility in the final output.

2.3 Inclusivity and accessibility

One of the advantages of digital editions is that they are a very viable alternative to consulting a document in person. This not only facilitates research in general, since it is only in very specific and quantitatively limited cases that it is necessary to travel in order to physically examine a document, but also an inclusive approach to the dissemination of cultural heritage, since people with serious motor disabilities can also easily use published texts. However, there are other types of disability that can make it difficult or even impossible for large groups of users to make use of a DSE: motor control problems, poor eyesight, hearing difficulties and other issues can have a serious impact on the usefulness (or otherwise) of a digital edition for many people.

As a first step, we have started to analyse the accessibility status of EVT 3 using a browser extension called SilkTide (see Fig. 2).[5]

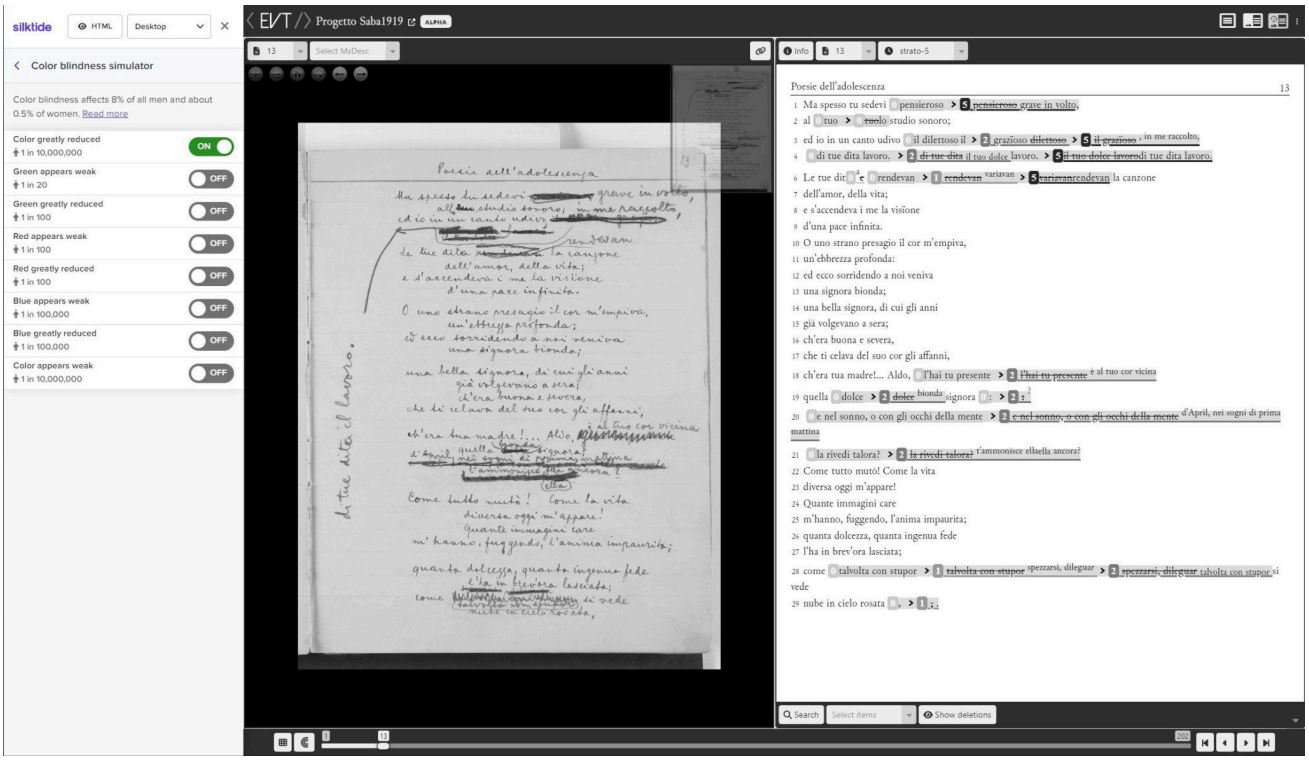


Figure 2. Checking the performance of EVT 3 in relation to visual impairment

In EVT 3 we plan to address this type of requirement in two ways:

[5] https://silktide.com/toolbar/.

- first, by improving the general layout and allowing navigation of the DSE with arrow keys and tabs, introducing shortcuts for selecting views, etc.; the use of ALT text for images and icons is also included in this part;
- the other key resource for improving the inclusivity and accessibility of the editions published using EVT will be the introduction of themes: this is also a feature introduced right away in EVT 3, with the intention of allowing, however limited, customisation of the appearance according to the type of edition (e.g. modern theme for critical editions, classic with warmer colours for diplomatic editions). The new task for themes is to customize the look and feel of EVT with the aim of mitigating visual impairments as much as possible.

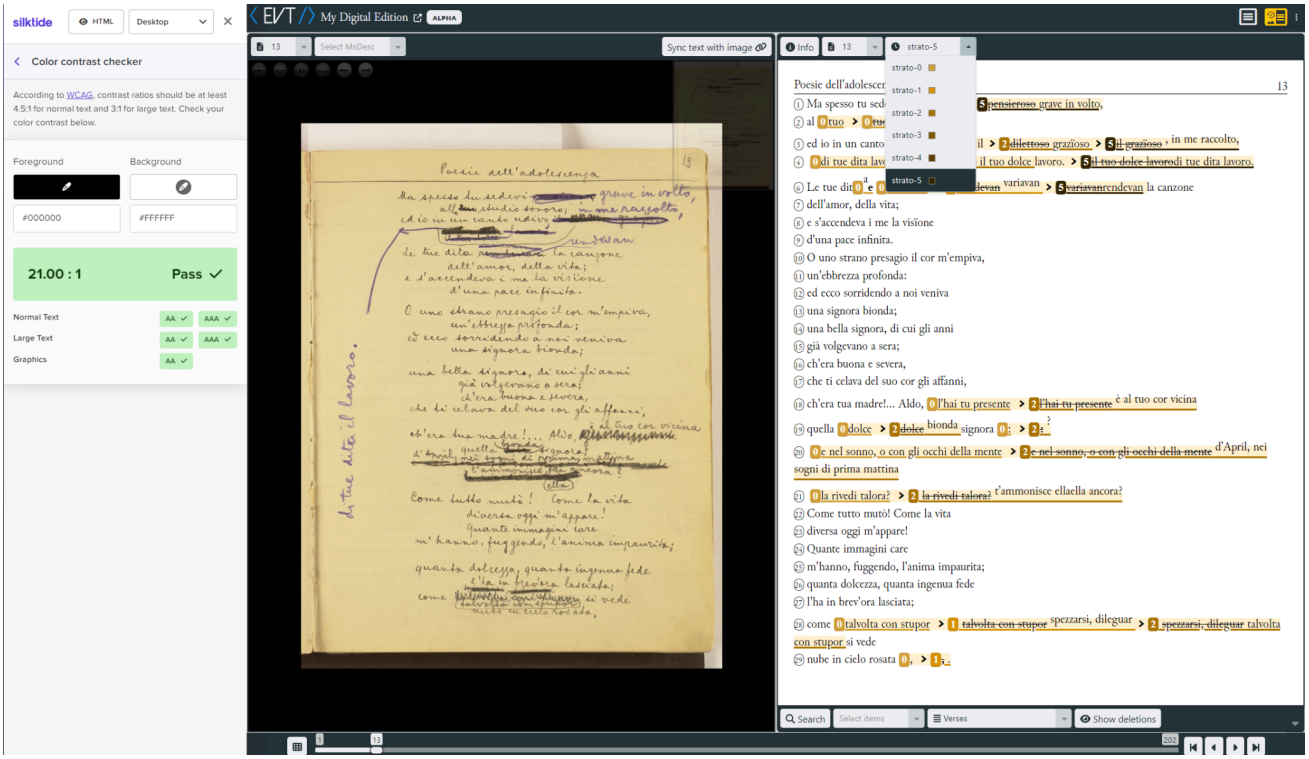


Figure 3. Testing the high contrast theme with SilkTide

Our main point of reference is the latest version of the Web Content Accessibility Guidelines 2.2 (WCAG 2024) which require alternative text for images, the availability of keyboard shortcuts, transcripts for audio (which we do not have in our project) and, above all, the possibility of modifying the contrast ratio according to specific needs by acting on the Graphical User Interface.

CONCLUSION

EVT has always been a visualization software focused on user needs, primarily the philologist who intends to create a critical edition. It is precisely by addressing user requirements—on the one hand, greater flexibility to create products better aligned with the practices of their communities, and on the other, increased attention to readers with specific needs—that the new features related to critical editions, accessibility, and inclusivity have been designed and developed. The "bottom-up" model, deemed effective by the authors given the results achieved, will continue to be employed in the future, with the addition of focus groups made up of readers from specific categories, capable of providing the team of designers and developers with informed feedback.

ACKNOWLEDGEMENTS

The *Leges Langobardorum* project is a PRIN PNRR 2022 project funded by the European Union – Next Generation EU, cod. P2022X2TXN – CUP H53D23009340001. EVT is developed with the support of the Digital Scholarship for the Humanities (DISH) research center (https://www.dish.unito.it/).

REFERENCES

Andrews, T. L., & van Zundert, J. J. (2018). What Are You Trying to Say? The Interface as an Integral Element of Argument. In R. Bleier, M. Bürgermeister, H. W. Klug, F. Neuber, & G. Schneider (Eds.), *Digital Scholarly Editions as Interfaces* (Vol. 12, pp. 3–34). BoD. https://kups.ub.uni-koeln.de/9106/.

Bausi, A. et al. (Eds.) (2015). *Comparative Oriental Manuscript Studies: An Introduction.* COMSt.

Bleier, R., Bürgermeister, M., Klug, H. W., Neuber, F., & Schneider, G. (Eds.). (2018). *Digital Scholarly Editions as Interfaces* (Vol. 12). BoD. https://kups.ub.uni-koeln.de/9085/.

Buzzoni, M., & Rosselli Del Turco, R. (2024). Towards an integrated digital edition of the *Leges Langobardorum*. In A. Di Silvestro & D. Spampinato (Eds.), *Atti del XIII Convegno Annuale AIUCD 2024* (pp. 226–231). https://doi.org/10.6092/unibo/amsacta/7927.

Buzzoni, M., Cucurnia, D., Fenu, C., Rosselli Del Turco, R., & Tancredi, G. (2024). Progetto di edizione genetica digitale del Canzoniere manoscritto di U. Saba (1919-20). In A. Di Silvestro & D. Spampinato (Eds.), *Atti del XIII Convegno Annuale AIUCD 2024* (pp. 215–220). https://doi.org/10.6092/unibo/amsacta/7927.

Buzzoni, M. (2024). Il «Progetto Saba»: Dare voce a un manoscritto inedito del Canzoniere. In M. Gatto, A. Squeo, & S. Silvestri (Eds.), *Informatica umanistica, Digital Humanities: Verso quale modernità?* (pp. 153–166). Cacucci Editore. https://iris.unive.it/handle/10278/5016061.

Cacioli, G., Cerretini, G., Chiara Di Pietro, Maenza, S., Roberto Rosselli Del Turco, & Zenzaro, S. (2022). There and back again: What to expect in the next EVT version. *AIUCD 2022 - Culture Digitali. Intersezioni: Filosofia, Arti, Media. Proceedings Della 11a Conferenza Nazionale, Lecce, 2022*, 212–217. https://conference.unisalento.it/ocs/index.php/aiucd2022/index/pages/view/proceedings

Cappellotto, A., & Cioffi, R. (2024). Towards an Inclusive and Accessible Digital Scholarly Editing: A Critical Assessment. *Journal of the Digital Humanities Association of Southern Africa*, *5*(1), Article 1. https://doi.org/10.55492/dhasa.v5i1.5012.

Di Pietro, C., & Rosselli Del Turco, R. (2018). Between Innovation and Conservation: The Narrow Path of User Interface Design for Digital Scholarly Editions. In *Digital Scholarly Editions as Interfaces* (Vol. 12, pp. 133–163). BoD. https://kups.ub.uni-koeln.de/9115/.

Drucker, J., El-Assady, M., Hinrichs, U., Windhager, F., & Akbaba, D. (2024). *Visualization and the Humanities: Towards a Shared Research Agenda (Dagstuhl Seminar 23381)*. https://doi.org/10.4230/DAGREP.13.9.137.

EVT – Edition Visualization Technology. (2014-). Home page: https://evt.labcd.unipi.it. GitHub source code repository: https://github.com/evt-project/evt-viewer-angular/.

Roelli, Ph. *et al*. (Eds.) (2020). *Handbook of Stemmatology. History, Methodologies, Digital Approaches*. De Gruyter.

Rosselli Del Turco, R. (2019). Designing an advanced software tool for Digital Scholarly Editions: The inception and development of EVT (Edition Visualization Technology). *Textual Cultures*, 12(2), 91–111. https://doi.org/10.14434/textual.v12i2.27690.

TEI Consortium (Eds.). (2025). *TEI: Guidelines for Electronic Text Encoding and Interchange.* 4.9.0. 24 January 2025. http://www.tei-c.org/Guidelines/P5/.

Web Content Accessibility Guidelines (WCAG) 2.2. (2024). W3C Recommendation 12 December 2024. https://www.w3.org/TR/WCAG22/.

Riscoperte poliane: edizione digitale di un manoscritto inedito de *Il Milione*

Giulia Fabbris[1], Samuela Simion[2], Fabio Soncin[3]
[1] Università Ca' Foscari Venezia, Italia – giulia.fabbris@unive.it
[2] Università Ca' Foscari Venezia, Italia – samuela.simion@unive.it
[3] Università Ca' Foscari Venezia, Italia – fabio.soncin@unive.it

ABSTRACT (ITALIANO)
Lo scopo di questo poster è la presentazione di un'edizione scientifica digitale di un manoscritto inedito del *Devisement dou Monde*, il codice A.II.9 conservato presso la Biblioteca Diocesana *Lodovico Jacobilli* di Foligno. L'edizione che si intende realizzare è di tipo diplomatico-interpretativo e verrà implementata secondo le linee guida XML-TEI. La visualizzazione avverrà tramite il noto software di visualizzazione Edition Visualization Technology (EVT), ormai giunto alla sua terza versione. Durante l'ideazione dello schema di codifica, sarà riservata particolare attenzione all'adattamento dei metadati descrittivi alle linee guida di Manus, al fine di allineare la codifica e permettere l'integrazione dell'edizione presso il portale. Particolare attenzione sarà prestata anche all'integrazione dei LOD.
Parole chiave: edizione scientifica digitale; Marco Polo; Devisement dou Monde; XML-TEI; Manus online

ABSTRACT (ENGLISH)
Polian Rediscoveries: Digital Edition of an Unpublished Manuscript of The Travels of Marco Polo.
The aim of this poster is to present a digital scholarly edition of an unedited manuscript of the *Devisement dou Monde*, ms. A.II.9 preserved at *Lodovico Jacobilli* library in Foligno. The edition we want to create is a diplomatic-interpretative one, and it will be implemented according to the XML-TEI guidelines. The visualization will be carried out through the well-known software Edition Visualization Technology (EVT), which has now reached its third version. Particular attention will be reserved to the adaptation of the descriptive metadata to the Manus guidelines, in order to align the encoding and allow the integration of the edition on the portal. Particular attention will be given also to the integration of LOD.
Keywords: digital scholarly edition; Marco Polo; Devisement dou Monde; XML-TEI; Manus online

1. INTRODUZIONE
Con questo poster si presenta un progetto di allestimento di un'edizione scientifica digitale del manoscritto Jacobilli A.II.9, della Biblioteca Diocesana *Lodovico Jacobilli* di Foligno,[1] latore di una importante versione del *Devisement dou Monde* (DM), opera odeporica tradizionalmente nota come *Il Milione*. Il testimone in oggetto, d'ora in avanti indicato come VA6, noto ma mai effettivamente rinvenuto, è stato segnalato nel 2024 da Fabio Soncin, ma già Petrini ne diede notizia nello stesso anno (Petrini, 2024). Appartiene al ramo VA[2] della tradizione poliana (si veda la sezione successiva per una panoramica generale), che trasmette la versione dell'opera più diffusa in Europa. L'edizione digitale in preparazione è un'edizione diplomatico-interpretativa sviluppata in conformità con le linee guida XML-TEI e di Manus Online (MOL),[3] al fine di garantire il supporto dell'edizione sul sito dei manoscritti delle biblioteche italiane. Si prevede inoltre di implementare l'edizione con i Linked Open Data (LOD),[4] in modo da collegare le diverse occorrenze di luoghi e persone a diversi server tramite collegamenti ipertestuali. Il prodotto finale sarà invece pubblicato su Edizioni Ca' Foscari (ECF)[5] in modalità open access e visualizzato con il software di visualizzazione Edition Visualization Technology (EVT).[6]

2. LA TRADIZIONE DEL DEVISEMENT DOU MONDE E L'EDIZIONE DIGITALE
Il DM, in Italia meglio noto come *Milione*, è uno dei testi più famosi dell'Europa medievale, scritto da Rustichello da Pisa e da Marco Polo durante la loro prigionia a Genova (1298-1299).

[1] https://www.jacobilli.it/ (cons. 24/01/2025).

[2] L'edizione di riferimento è (Barbieri & Andreose, 1999).

[3] https://manus.iccu.sbn.it/ (cons. 24/01/2025).

[4] https://www.ontotext.com/knowledgehub/fundamentals/linked-data-linked-open-data/ (cons. 24/01/2025).

[5] https://edizionicafoscari.unive.it/it/edizioni/ (cons. 24/01/2025).

[6] https://visualizationtechnology.wordpress.com/ (cons. 24/01/2025).

La versione originale scritta a Genova sparì rapidamente dalla circolazione, lasciando posto a un gran numero di traduzioni nelle principali lingue europee, tanto che la tradizione manoscritta del *Milione* è quasi interamente una tradizione indiretta. Il testo primitivo fu scritto in franco-italiano, e l'unico testimone completo pervenutoci in questa lingua è il codice BnF fr. 1116, conosciuto come F: i restanti 143 testimoni sono tutti traduzioni.[7] I copisti, che agirono di fatto come veri e propri redattori, intervennero pesantemente sul contenuto, con tagli e amplificazioni, sulla struttura e sulla lingua, dando così vita a una tradizione estremamente complessa e contaminata.
Malgrado negli ultimi vent'anni la maggior parte delle redazioni del DM sia stata oggetto di edizioni critiche, il tradizionale supporto cartaceo risulta inadeguato qualora si intenda pubblicare un'edizione critica "integrale". Infatti, mentre queste ultime possono essere efficacemente rappresentate nelle edizioni classiche, un'edizione integrale del *Milione* necessita di un supporto più dinamico che sia in grado di tenere conto sia della *mouvance* che caratterizza l'opera, nonché della pluralità delle lingue in cui questa è attestata.

3. IL MARCO POLO DIGITALE

Per affrontare le sfide della tradizione manoscritta poliana, è stato creato un progetto di edizione scientifica digitale dell'opera, denominato DEDM (*Digital Edition of the Devisement dou monde*),[8] coordinato da Eugenio Burgio,[9] Marina Buzzoni[10] e Samuela Simion dell'Università Ca' Foscari Venezia e in pubblicazione con ECF. L'obiettivo è realizzare un'edizione critica basata su dodici redazioni selezionate, seguendo la più recente proposta stemmatica (vd. Figura 1).[11] Il testo critico, disponibile in inglese, si basa su F, il testimone che conserva la struttura e la lingua più primitive, nonché il testo più completo. La scelta dell'inglese sopperisce all'impossibilità di ricostruire l'originale genovese e permette di raggiungere un pubblico ampio e internazionale. I tredici testi (dodici redazioni e la traduzione) sono presentati in una visualizzazione sinottica che consente di navigare facilmente tra le versioni grazie all'allineamento a livello dei capitoli.[12]

[7] Esiste un altro testimone frammentario e molto breve, *f*, diviso in due collezioni private.

[8] https://edizionicafoscari.unive.it/it/edizioni4/libri/978-88-6969-872-9/ (cons. 24/01/2025).

[9] burgio@unive.it

[10] mbuzzoni@unive.it

[11] Le redazioni incluse nel progetto sono: F = redazione franco-italiana del XIV sec.; Fr = redazione francese; K = redazione «catalana»; L = compendio latino trecentesco; R = la redazione allestita da Giovanni Battista Ramusio per il secondo volume della silloge odeporica *Navigationi et Viaggi* (1559), con il titolo *Dei Viaggi di messer Marco Polo gentiluomo veneziano*, esito di un collage di almeno tre esemplari (afferenti ai rami P, VB e Z); TA = redazione toscana primo-trecentesca; VA = redazione veneto-emiliana; P = traduzione in latino condotta su un esemplare VA da Francesco Pipino, entro il primo quarto del XIV sec.; TB = redazione toscana tardo-trecentesca; VB = rimaneggiamento veneziano; Z = versione latina.

[12] Inoltre, sono state predisposte schede dettagliate per una numerosa selezione di antroponimi, toponimi ed entità generali. I file sono stati codificati in XML seguendo le linee guida TEI. La visualizzazione avviene tramite EVT anche in questo caso. Per un'introduzione più dettagliata alla tradizione manoscritta e ai motivi che hanno portato all'ideazione dell'edizione digitale del DEDM si veda Eusebi, Burgio e Simion (2024).

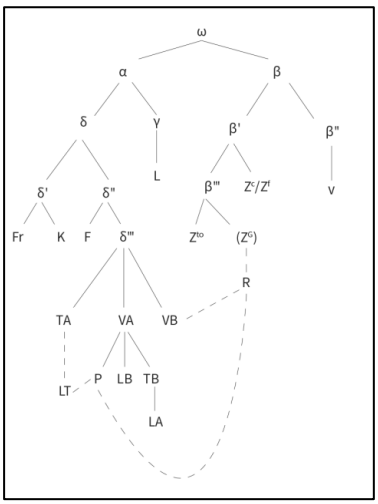


Figura 1. *Stemma editionum* proposto da Simion (2017: 27).

4. LA TRADIZIONE VA E IL TESTIMONE FOLIGNATE

Come già accennato, il testimone del quale si propone l'edizione digitale appartiene al ramo VA, per la cui posizione nello *stemma editionum* si riporta la recente proposta di Simion (2017: 27) (vd. sopra, Figura 1).

Il testimone A.II.9 (= VA6), conservato presso la Biblioteca Diocesana *Lodovico Jacobilli* di Foligno, è il 145° conosciuto a trasmettere il testo del DM (Figura 2). Presente nei cataloghi,[13] ma ignoto agli studiosi poliani, il manoscritto trasmette una traduzione trecentesca del DM appartenente alla famiglia veneto-emiliana VA (ricavata da un testimone franco-italiano perduto affine a F).

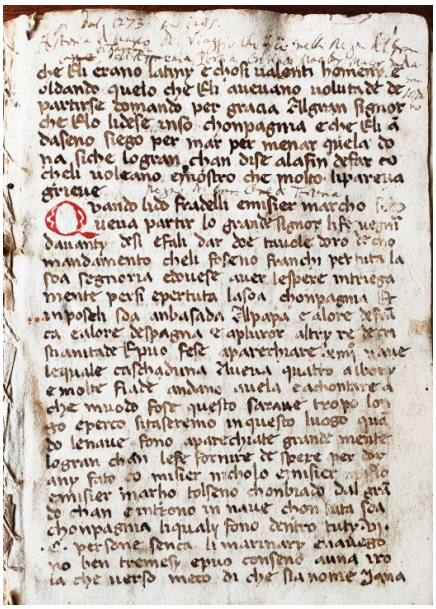


Figura 2. Foligno, Biblioteca Diocesana *Lodovico Jacobilli*, A.II.9, f. 1r

Tale redazione, a lungo conosciuta come la redazione "veneta per eccellenza" (Benedetto 1928: C), è fondamentale nella storia della diffusione e della fortuna del DM, malgrado il suo apporto secondario in un'ottica ricostruttiva. Il gruppo è oggi costituito da sei testimoni: VA1[14] e VA2[15] (trecenteschi), VA3[16] e

[13] https://manus.iccu.sbn.it/risultati-ricerca-manoscritti/-/manus-search/detail/673521?biblioteca_s=Foligno%2C+Biblioteca+L.+Jacobilli& (cons. 24/01/2025).

[14] Roma, Biblioteca Casanatense, ms. 3999.

[15] Firenze, Biblioteca Riccardiana, ms. 1924.

[16] Padova, Biblioteca Civica, ms. CM 211.

VA4 - quest'ultimo irreperibile - (quattrocenteschi), VA5[17] (cinquecentesco). La datazione del codice Jacobilli, VA6, è fissata al quindicesimo secolo nella scheda MOL; alcuni indizi suggeriscono tuttavia una retrodatazione al quattordicesimo secolo; per averne conferma sarà necessario completare la perizia paleografica.[18]
Accanto a questi testimoni diretti, VA vanta una nutrita discendenza, tanto da essere di fatto la redazione più diffusa in Europa tra Medioevo e età moderna: da esemplari perduti verranno precocemente ricavate svariate traduzioni: P, LB, TB, VL e LT.

5. L'EDIZIONE DIPLOMATICO-INTERPRETATIVA: SCHEMA DI CODIFICA

L'edizione del codice A.II.9 è un progetto condiviso con la Biblioteca Diocesana *Lodovico Jacobilli* e con la Soprintendenza Archeologia, Belle Arti e Paesaggio dell'Umbria.[19] L'edizione sarà pubblicata con ECF e, punto di particolare importanza e su richiesta della Soprintendenza stessa, sarà modellata nel rispetto delle linee guida di Manus al fine di garantirne il supporto nell'archivio digitale.
L'edizione sarà diplomatico-interpretativa al fine di valorizzare la storicità di un manoscritto che fino ad ora è rimasto nell'ombra. Poiché il testo è ben leggibile e non presenta grosse difficoltà, gli interventi editoriali saranno leggeri. Tuttavia, la doppia codifica, ossia quella del livello diplomatico e quella del livello interpretativo, è necessaria al fine di rendere il testo più accessibile, dando modo anche a utenti meno esperti o con difficoltà di fruire dell'edizione (vd. Figura 3).

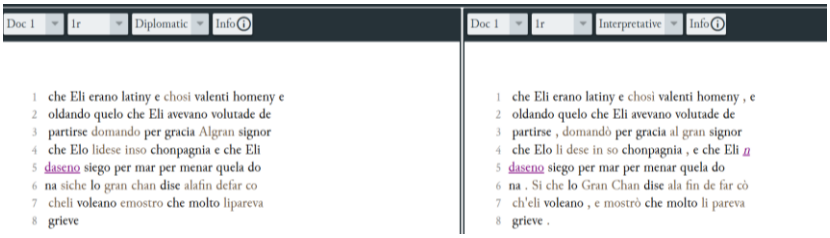


Figura 3. Esempio di codifica diplomatica (a sinistra) e interpretativa (a destra)

Lo schema di codifica, che segue le linee guida XML/TEI, è basato ampiamente su quello ideato per il DEDM, essendo quest'ultimo un punto di riferimento per un lavoro di questa tipologia. Tuttavia, le due codifiche divergono in alcuni punti.
Innanzitutto, il <teiHeader> sarà, come già accennato sopra, allineato con il tracciato di Manus.[20] Secondo questo documento, le tre componenti principali del file sono l'intestazione del documento (<teiHeader>), la gestione delle immagini facsimile (<facsimile>) e la registrazione dei dati esterni (<standOff>). Ad ogni modo, la maggior parte degli elementi e degli attributi usati per la rappresentazione XML/TEI della scheda di MOL si ritrovano nella descrizione del manoscritto (<msDesc>). Poiché anche EVT richiede una serie di elementi e attributi in <facsimile> per l'allineamento testo-immagine, questo punto richiederà particolare attenzione al fine di avere una codifica che sia compatibile con i requisiti di entrambe le linee guida.
Verranno inoltre predisposte delle liste di termini, che saranno suddivisi in toponimi, antroponimi e altre entità generiche. Ove possibile, a queste voci sarà abbinato un riferimento ai LOD (supportati da EVT a partire dalla versione 2.0, Monella & Rosselli Del Turco 2020) al fine di contribuire alla costruzione di reti di relazioni semantiche tra testi, con la possibilità di effettuare ricerche più specifiche e così creare informazioni. Segue un breve esempio di codifica secondo questo principio.

```
<listPerson>
        <person xml:id="marcoPolo">
          <persName
ref="https://www.wikidata.org/wiki/Q6101"><forename>Marco</forename><surname>Polo</surname>
</persName>
```

[17] Berna, Biblioteca Civica, ms. 557.

[18] La perizia paleografica sarà effettuata nel corso del 2025, in quanto attualmente il manoscritto è in prestito presso l'Accademia dei Lincei di Roma. Tale perizia sarà importante al fine di datare con precisione il manoscritto attraverso l'analisi delle filigrane.

[19] https://sabapumbria.cultura.gov.it/ (cons. 24/01/2025).

[20] https://manus.iccu.sbn.it/tracciato-xml-tei-di-manus-online (cons. 24/01/2025).

```
            <occupation>viaggiatore</occupation>
          </person>
        </listPerson>
```

Per quanto riguarda il testo in sé, l'attenzione è stata posta sui seguenti punti, con lo scopo di favorire la leggibilità del testo in edizione moderna:

- modernizzazione della punteggiatura;
- normalizzazione delle lettere maiuscole;
- aggiunta di accenti e segni diacritici secondo l'uso moderno;
- aggiunta di spazi laddove le parole sono state scritte unite;
- scioglimento delle abbreviazioni;

Tutte le modifiche sopra citate sono codificate all'interno di un elemento <choice> con l'utilizzo di <orig> per la lezione presente nel manoscritto – che verrà quindi visualizzata nella resa diplomatica – e di <reg> per la voce normalizzata che creerà il testo interpretativo. Fa eccezione l'ultima voce della lista, per la quale sono stati usati invece gli elementi <abbr> e <expan>, trattandosi appunto di abbreviazioni e le loro espansioni (vd. Figura 4). Questo set di elementi potrà essere arricchito con l'attributo @choice al fine di specificare il tipo di normalizzazione, informazione che non sarà poi computata da EVT, ma che sarà conservata all'interno del file XML al fine di garantire comunque la presenza del metadato.

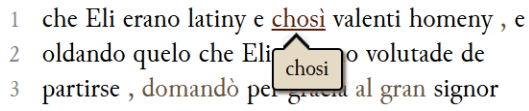


Figura 4. Resa grafica del set di elementi all'interno di <choice>

La tipologia di edizione richiede inoltre l'accostamento delle immagini del manoscritto al testo codificato, e ciò è reso possibile grazie alle funzioni di EVT che permettono l'allineamento testo-immagine. Le due componenti comunicano a livello di pagina e a livello di singola linea. Ci saranno da una parte gli elementi <surface/> e <zone/> contenuti in <facsimile> collegati rispettivamente con <pb/> e <lb/> contenuti in <body>. Ognuno di questi elementi è identificato da un @xml:id che verrà richiamato tramite un @corresp o un @facs nell'elemento complementare. L'elemento <zone/> contiene inoltre i punti delle coordinate cartesiane all'interno dell'immagine che permettono di creare il rettangolo che isola la singola riga. Di seguito si propone un esempio di codifica per la sincronizzazione testo-immagine e in Figura 5 la resa.[21]

```
<facsimile xml:id="A.II.9_fac">
    <surface xml:id="A.II.9_surf_1r" corresp="#A.II.9_fol_1r">
      <graphic url="1r.jpg" width="2481px" height="3509px"/>
      <zone xml:id="A.II.9_line_1r_1" corresp="#A.II.9_lb_1r_1" ulx="349" uly="486" lrx="2216"
lry="569" rend="visible" rendition="Line"/>
    </surface>
  </facsimile>
<text>
    <body>
     <pb n="1r" xml:id="A.II.9_fol_1r" corresp="#A.II.9_surf_1r" facs="data/images/single/1r.jpg"/>
     <div xml:id="numero_capitolo">
       <p>
         <seg xml:id="numero_pericope1" n="1">
           <lb n="1" xml:id="A.II.9_lb_1r_1" facs="#A.II.9_line_1r_1"/>[…]</seg>
       <p/>
     </div>
   </body>
 </text>
```

[21] Nell'esempio qui proposto sono assenti gli attributi richiesti dal tracciato di Manus, che verranno aggiunti successivamente.

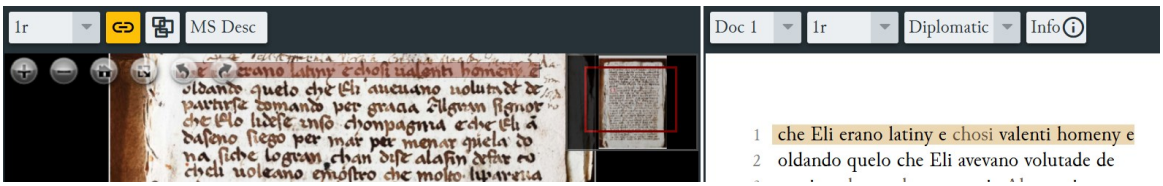


Figura 5. Esempio di collegamento testo-immagine

L'importanza di questa edizione scientifica è duplice: dal punto di vista digitale si crea un'edizione che da un lato si conforma agli standard di codifica internazionali e dall'altro viene adeguata al fine di permetterne la conservazione in un database, MOL, estremamente importante per la tutela dei beni manoscritti delle biblioteche italiane. Tutto questo, insieme con la pubblicazione in open access dell'edizione tramite ECF, garantisce la disseminazione e la libera fruizione del manoscritto, finora inedito, permettendone l'uso sia agli studiosi che a un pubblico più ampio; dal punto di vista prettamente filologico si restituisce l'edizione di un testimone appartenente a un importante ramo della tradizione poliana a beneficio del DEDM.

RINGRAZIAMENTI

Si ringrazia la Biblioteca Diocesana *Ludovico Jacobilli* di Foligno, in particolar modo il bibliotecario Ivan Petrini per l'accesso al manoscritto, la Soprintendenza della Regione Umbria per l'interesse manifestato nella realizzazione del progetto, e i collaboratori del progetto DEDM per l'attivo interesse e contributo nella realizzazione di questo lavoro.

BIBLIOGRAFIA

Barbieri, A. (1999). Introduzione. In A. Barbieri & A. Andreose (A cura di), *Il «Milione» veneto, ms. CM 211 della Biblioteca di Padova* (pp. 23–65). Marsilio Editori s.p.a. http://opac.regesta-imperii.de/id/1103439

Benedetto, L. F. (1928). *Marco Polo: Il Milione; Prima edizione integrale*. L.S. Olschki. https://archive.org/details/marcopolomilionebenedetto/page/n31/mode/2up

Bertolucci Pizzorusso, Valeria (A cura di). (1975). *Marco Polo: Milione. Versione toscana del Trecento*. Milano: Adelphi.

Buzzoni, M. (2024). Il «Devisement dou monde» in forma digitale. In E. Burgio & S. Simion (A cura di), *Marco Polo. Storia e mito di un viaggio e di un libro* (pp. 165–179). Carocci. https://www.carocci.it/prodotto/marco-polo

Eusebi, M., Burgio, E., & Simion, S. (2024). *On the Way to an Integral Edition of the Book of Marco Polo: A First Attempt to Create a Digital Edition*. In H. U. Vogel & U. Theobald (A cura di), Marco Polo Research: Past, Present, Future (pp. 91–131). Tübingen Library Publishing. https://iris.unive.it/handle/10278/5062381

Monella, P., & Rosselli Del Turco, R. (2020). Extending the DSE: LOD Support and TEI/IIIF Integration in EVT. In C. Marras, M. Passarotti, G. Franzini, & Litta Eleonora (A cura di), *La svolta inevitabile: Sfide e prospettive per l'informatica umanistica*. Milano: Università Cattolica del Sacro Cuore. https://doi.org/10.6092/unibo%2Famsacta%2F6316

Petrini, I. (2024, marzo 17). *Carovanieri e copisti. Sulla Via della Seta fino a Foligno*. Gazzetta di Foligno, 9.

Simion, S. (2017). *Tradizioni attive e ipertesti: Ramusio «editore» del Milione*. 6(2), 9–30. https://dx.doi.org/10.30687/QV/1724-188X/2017/02/001

«Proximior perfectioni»: criticità e future prospettive del progetto Dante Limina

Elisabetta Tonello
Università di Salerno, ITALIA – etonello@unisa.it

ABSTRACT (ITALIANO)

Il contributo vuole illustrare un recente progetto legato al testo della *Commedia* di Dante e affrontare alcuni problemi di metodo sorti in seno al suo sviluppo informatico, collegato ai diversi approcci ermeneutici e concettuali dell'oggetto di ricerca su cui si basano tutti i sub-progetti. Si cercherà di ricostruire il cammino di progressivo miglioramento nell'utilizzo delle DH per il processo di acquisizione delle conoscenze delle diverse prospettive e oggetti di ricerca. Si discuteranno quindi le opportunità concesse dagli archivi digitali per la dantistica e le criticità connesse.

Parole chiave: Archivi digitali; Digitalizzazioni; Patrimonio culturale; Dante; *Commedia*

ABSTRACT (ENGLISH)

"Proximior perfectioni": Challenges and Future Prospects of the Dante Limina Project

The contribution aims to illustrate a recent project related to Dante's *Divine Comedy* and address some methodological issues that arose during its development in the digital context, connected to the different hermeneutic and conceptual approaches to the research object on which all sub-projects are based. The aim is to reconstruct the path of progressive improvement in the use of DH for the process of acquiring knowledge from different perspectives and research objects. Finally, the opportunities provided by digital archives for Dante studies and the related challenges will be discussed.

Keywords: Digital Archive; Digitization; Cultural Heritage; Dante; *Commedia*

1. INTRODUZIONE

In questo contributo si intende illustrare e problematizzare l'incontro tra dantistica e DH, attraverso il caso studio del progetto *Dante limina* diretto da E. Tonello (Università di Salerno) e C. Perna (Università della Campania "Luigi Vanvitelli"). L'esperienza del filologo e dello storico della letteratura che incontra la *Commedia* di Dante può essere disorientante per la massa di testimoni manoscritti del poema (sono circa 800, sparsi nelle biblioteche di tutto il mondo) e per la complessa e multiforme storia della ricezione del testo che ha coinvolto tutte le forme artistiche, che ha dato vita e nutrito esperienze letterarie, pittoriche, plastiche e musicali autonome e che ha creato un immaginario collettivo su scala mondiale. In altre parole, gli studiosi della *Commedia* sentono più di altri la necessità dell'ausilio informatico per riunire, ordinare, razionalizzare imponenti messe di dati. Non si deve però pensare che queste operazioni siano neutre: l'organizzazione della conoscenza all'interno dell'ambiente informatico comporta scelte di metodo e in alcuni casi alterazione della percezione degli oggetti della ricerca finendo per tracciare un nuovo paradigma di rappresentazione e sapere.

2. IL PORTALE DANTE LIMINA

L'esperienza di allestire una edizione critica basata sulla collazione (per *loci critici*) di tutto il testimoniale superstite e gli studi sull'esegesi e l'illustrazione antica del poema di chi ha pensato questo progetto, che è poi un grappolo di progetti, è sfociata in modo naturale nell'esigenza di adottare soluzioni informatiche per rendere disponibili i risultati delle ricerche, per coinvolgere la comunità accademica in problemi di non facile soluzione, per rendere fruibili alla comunità (e dunque conservati e preservati) gli oggetti di base delle analisi (manoscritti, stampe antiche, ma anche disegni e manga…) e per stimolare la curiosità e diffondere il patrimonio culturale presso gli studenti di tutti i gradi e la popolazione tutta.
Il nome scelto per il progetto che viene presentato in queste pagine è DANTE LIMINA. *Limina* fa riferimento al contempo sia alla marginalità, a tutto ciò che si trova in posizione decentrata, sia alla soglia di accesso al testo dantesco. Si tratta di un contenitore cui fanno capo (sub-)progetti, ognuno con una sua identità e un suo team dedicato: https://www.dantelimina.it. Come vedremo, ogni progetto ha rappresentato uno step

di una lunga e costante sperimentazione e progressiva crescita di consapevolezza dei principi di base per le DH, a partire dallo sforzo di astrazione nella rappresentazione dei dati in nostro possesso e della replicabilità per altri progetti simili, fino alla inclusività degli utenti e della comunità accademica.

3. DANTEMATRIX

In ordine di tempo il primo sub-progetto ad essere stato sviluppato è stato DanteMatrix (https://dantelimina.it/matrix/public/). Questo programma rappresenta l'esito dei risultati delle collazioni del *Gruppo di Ferrara*, guidato da P. Trovato e che ha come obiettivo l'edizione critica della *Commedia* di Dante basata, per la prima volta nella storia della critica, sulla collazione per *loci critici* dell'intero testimoniale non frammentario (ca. 600 mss.). Durante la fase di collazione erano stati fatti dei tentativi di sviluppo di un software[1] che aiutasse a interrogare gli esiti dei versi in esame, annotati in fogli Excel e resi in codice binario, ma è stato solo al termine della classificazione e con la prima cantica già pubblicata (Tonello, Trovato, 2022) che Tonello e Trovato, grazie all'intervento informatico di L. Tessarolo, sono riusciti a trasformare le matrici di collazione in un database relazionale MySQL, strutturato per consentire interrogazioni complesse basate su parametri personalizzabili dall'utente. I dati, originariamente organizzati in un foglio di calcolo .xlsx, sono stati convertiti in formato .tsv (Tab-Separated Values) per favorire l'elaborazione automatizzata. Successivamente, attraverso una pipeline di parsing e normalizzazione, sono stati inseriti nel database, rispettando uno schema progettato per rappresentare in modo efficace le relazioni tra varianti testuali. L'interfaccia web è stata sviluppata utilizzando il framework PHP e la libreria jQuery, allo scopo di restituire in modo interattivo le ipotesi di relazioni tra manoscritti, costruite sulla base di modelli statistici applicati alle occorrenze di varianti.
Questo impianto consente non solo di rendere navigabili e interrogabili dati filologici complessi, ma anche di proporre visualizzazioni dinamiche in grado di supportare nuove ipotesi critiche sulla tradizione manoscritta.[2] Si tratta, con ogni evidenza, di un progetto molto ricco di informazioni, ma che interessa pochissimi specialisti e che necessita di una guida all'uso molto dettagliata e per nulla semplice. Vanno incontro all'utente altre funzioni rese disponibili sul sito, come la possibilità di consultare uno stato dell'arte sulla storia delle edizioni critiche del poema, di scaricare i canti della *Commedia* con i *loci critici* evidenziati da un lato, quello più divulgativo, e di sfruttare i raggruppamenti di manoscritti come subarchetipi a loro volta ricercabili o le liste ordinate di innovazioni comuni a gruppi di codici dall'altro lato, quello più tecnico.
Questa prima prova, che mostra evidenti criticità, ha rappresentato per me l'ingresso nel complesso mondo delle DH. Al di là di alcune ingenuità nella strutturazione del sito, il problema più grande su cui sono stata costretta a interrogarmi riguardava la strutturazione dei dati e la possibilità di renderli accessibili, di metterli in relazione con altri enti, e riutilizzabili. Per prima cosa, la forma nella quale si trovavano questi dati dall'origine costituiva già di per sé una razionalizzazione matematica che il linguaggio macchina era automaticamente in grado di comprendere. Tuttavia la perdita di informazione che questa standardizzazione dei dati comportava in alcuni casi poteva inficiare i risultati della ricerca (penso soprattutto al peso che certi esiti di lezione hanno sia per ragioni di potenziale di poligenesi sia per ragioni di effettiva poligenesi dell'errore). Apro una piccola parentesi per sostanziare di un dato storico questo passaggio. All'inizio del secolo scorso, Giuseppe Vandelli fu incaricato dalla Società Dantesca Italiana di produrre l'edizione critica della *Commedia* per l'edizione nazionale di tutte le opere di Dante Alighieri. Lo studioso collazionò tutti i codici della *Commedia* superstiti (che ricordo ancora essere 600) e annotò le varianti riscontrate (su un canone di circa 600 versi) di ognuno di essi in grandi fogli, oggi secretati e non accessibili. Quando fu il momento di allestire l'edizione si rese conto di non essere in grado di produrre uno stemma e tantomeno di sfruttare i risultati di quella campagna di collazione (la prima e unica così estesa, almeno fino a oggi). La forma dei dati e il supporto sui quali si reggevano costituivano gli ostacoli principali a quel principio di razionalizzazione che è alla base (e in parte il risultato) della ricerca. È stato sufficiente annotare gli esiti delle collazioni in linguaggio matematico (linguaggio binario 0;1) su un foglio di calcolo per poter osservare chiaramente la situazione genealogica e procedere alle formulazioni di gruppi e sottogruppi. Il software ha poi permesso di rendere automatico il processo di individuazione delle affinità.
In secondo luogo, l'estrema specificità e unicità di questo lavoro (anche rispetto ad esperienze passate – è la prima volta che si raccolgono tutte le collazioni di così tanti *loci critici* sull'intero testimoniale superstite – ha finito per condizionare (in negativo!) l'architettura informatica del DB. Non sono stati fatti sforzi per concepire DanteMatrix secondo le teorie e i modelli del Semantic Web, con la conseguenza che i dati, da

[1] Una relazione si può leggere in Renello, 2013.

[2] Per una descrizione dettagliata si veda Malatesta & Tonello, 2024.

quelli legati a entità fisiche (come i manoscritti) a quelli più teorici (la lezione sostanziale di un determinato codice in un determinato verso mondata dei suoi attributi linguistico formali), non risultano metadatati, associati a URI e interrelati attraverso l'RDF.

4. LIMINA

Un altro sub-progetto, LiMINA (Lost in Manuscripts. Ideas, Notes, Acknowledgments) https://limina.dantelimina.it, finanziato Prin 2022, ha dato il via a questo hub di incontri e intersezioni, DANTE LIMINA, e riguarda le tracce marginali (ed "emarginate") dei lettori nei mss. della *Commedia* di Dante. Il corpus è stato ristretto ai codici della cosiddetta "Antica vulgata" (ovvero ai mss. datati *ante* 1355), per rientrare nei limiti del finanziamento. L'obiettivo del progetto era dare voce alla borghesia colta cittadina medievale (di cui poco o nulla si sa) che aveva lasciato traccia di sé, dei propri gusti, dei propri orizzonti culturali, in un supporto come gli spazi bianchi delle pagine di un ms. in relazione (o meno) al testo principale, da sempre oggetto di attenzione prediletto da parte della critica. Per mettere a sistema la mole di dati che deriva dalle indagini su queste tracce e trarne informazioni utili per la storia della cultura, della sociologia, della ricezione del testo, dell'antropologia e della psicologia era necessario organizzare le tipologie, così disomogenee, di interventi nei codici (segni di attenzione, note, disegni, *maniculae*, annotazioni biografiche personali, brevi trascrizioni di testi…), classificarle, dotarle di informazioni aggiuntive e renderle così ricercabili. L. Tessarolo ha sviluppato un database relazionale (MySQL) su misura per queste esigenze. S. Malatesta ha diretto e orchestrato, insieme all'illustratore Angelo Rauso, una accattivante e significativa grafica, che alleggerisse l'aspetto altamente specialistico dell'operazione. Il workflow, che prevede la schedatura del manoscritto, la catalogazione dei relativi *limina* e la possibilità di visualizzarli nella riproduzione digitale in IIIF, ha avuto come punto di forza la collaborazione con ManusOnline. Il riconoscimento di progetto speciale da parte dell'ICCU (https://manus.iccu.sbn.it/limina) ha infatti reso possibile l'esportazione della scheda paleografica depositata in Manus nel DB Limina. Inoltre, proprio in questi mesi, MOL sta aggiungendo ai suoi campi proprio la categoria "Limina" con le sue macro-tipologie, segnando così l'ingresso nella comunità scientifica del nuovo oggetto di ricerca, con una sua autonomia e funzione specifica. L'intero sistema è stato realizzato come un'unica applicazione web integrata, sviluppata in PHP 8.2, che gestisce sia il back-end sia il front-end in modo modulare. L'applicazione è strutturata secondo una logica MVC (Model–View–Controller) e si appoggia a un database MySQL. Di particolare rilievo è il sistema Ariadne, sviluppato per semplificare la gestione back-end di relazioni complesse tra tabelle, integrato in interfacce dinamiche via JavaScript. L'intero impianto favorisce la modularità, la riusabilità del codice e l'adattabilità tra ambienti, e riflette un approccio orientato alla sostenibilità del software di medio-lungo termine. Anche in questo caso lo sviluppo del DB, anch'esso relazionale, era per lo più tagliato sulle specificità che questa complessa materia di studio richiedeva. In sostanza si stava finendo per replicare l'esperienza di un database che riflettesse la molteplicità del reale e la complessità della materia, ignorando le potenzialità che un oggetto di ricerca come questo, i limina, possiede proprio da un punto di vista orizzontale e trasversale. La scelta di andare a fondo, verticalmente, nell'esame della questione (catalogando minuziosamente tutte le manifestazioni fenomeniche della presenza delle tracce del lettore in un corpus limitato e definito) ha scontato nuovamente la possibilità di apertura a una organizzazione del sapere più ampio e interconnesso. Per questa ragione si è deciso di optare, da quel momento in avanti, su architetture più libere. Con questo progetto si è dunque iniziato a riflettere sull'efficacia della rappresentazione dei dati attraverso strutture standardizzate. «La realtà che vogliamo descrivere è spesso molto complessa e lo schema (qualunque schema, per quanto ricco e variegato) correrà sempre il rischio di essere posto di fronte al caso eccezionale che non può esservi descritto», dice F. Tomasi.[3] Infatti, il problema ha iniziato lentamente a spostarsi verso una più compiuta teoria della conoscenza nell'ambito delle DH. Ho l'impressione, infatti, che non si insista abbastanza sul fatto che la nostra percezione del reale, degli oggetti che lo compongono (ma anche, con termine husserliano, delle essenze) sia riflessa solo in parte dalle potenzialità dell'infrastruttura informatica e che, viceversa, quando operiamo con le DH, siamo costretti necessariamente a sfruttare le nostre intuizioni eidetiche, per restare con Husserl, cioè le essenze dei fenomeni della nostra coscienza, potenziando e liberando il sapere dalla materia, esplorandone punti di osservazione inusitati.

[3] Tomasi, Vitali, 2023: 138-139.

5. D.A.N.T.E.

Per D.A.N.T.E. (Digital Archive and New Technologies for E-content) https://dante.dantelimina.it/home è stata sviluppata una struttura di DB a grafo, da F. Masucci (con l'aiuto di S. Malatesta). D.A.N.T.E. è un progetto finanziato dal fondo V:ALERE 2022 dell'Università della Campania "Luigi Vanvitelli" che prevede lo studio dell'iconografia dantesca negli incunaboli e nelle stampe cinque e seicentesche del poema. Lo scopo ultimo è la generazione di esperienze immersive in realtà virtuale e ottimizzata per il visore Meta Quest 3. Insomma, iniziava una nuova esperienza, da un lato ancora legata alle precedenti, perché ancorata al testo dantesco, alla catalogazione, e alla possibilità di rendere disponibile in digitale il patrimonio delle prime stampe della *Commedia* (e in una sottosezione, quella che descrive l'esemplare dell'edizione impiegato, lo sfruttamento delle etichette e delle gerarchie pensate per i *limina* dei manoscritti), ma dall'altro lato totalmente inedite perché legate ai problemi e alle potenzialità delle DH in relazione alla storia dell'arte, sia dal punto di vista teorico che di organizzazione dei dati.[4] Relab è stata incaricata di realizzare col team di D.A.N.T.E. un filmato immersivo a 360 gradi, sviluppato con Blender per la modellazione e Unreal Engine per la creazione dell'ambiente virtuale che riproduce il passaggio di Dante e Virgilio all'interno di un piccolo dettaglio della pianta di Botticelli, per far provare allo spettatore le sensazioni delle descrizioni (e traduzioni in immagini) della *Commedia*. È in fase di completamento un altro filmato che farà partecipare l'utente a una narrazione in grado di spiegare, muovendosi nei diversi ambienti, anzitutto come operare in biblioteca e consultare le xilografie, osservandole e confrontandole a piacimento e poi come lavorare con il torchio delle prime stamperie e con le scansioni digitali al giorno d'oggi. Chiaramente la finalità "didattica" al centro di questa esperienza ha determinato molti degli orientamenti della ricerca. Ora che la campagna di digitalizzazioni e descrizioni è conclusa, si è aperta una nuova fase di studio delle immagini che hanno maggiormente influenzato l'immaginario connesso alla *Commedia*, del perpetuarsi di questi paesaggi e figure nell'arte e nell'iconografia legata al testo dantesco, in vista della selezione di una galleria di "scene" che prenderanno vita nei filmati realizzati da A. Calamo e che speriamo di vedere un domani impiegati in contesti museali e scolastici.
Il database ospiterà invece le descrizione dei cicli illustrativi conservati nelle stampe e delle singole immagini che compongono il ciclo. Le campagne di descrizione, delle edizioni e dei disegni, sono confluite in un catalogo tradizionale di esemplari a stampa e di illustrazioni. Questa esperienza ci ha permesso di riflettere a fondo sull'oggetto della nostra ricerca (e sull'oggetto libro stesso) e sulle diverse prospettive che è possibile adottare attraverso l'approccio digitale. Anzitutto era ormai chiaro che la progettazione non poteva partire da un modello centrato sul focus della nostra ricerca ma doveva ispirarsi a un modello di sapere universalistico in grado di comprendere le molteplici possibilità del reale. Roncaglia afferma che «l'enciclopedia rappresenta l'esempio paradigmatico della concezione architettonica del sapere, e l'enciclopedismo digitale ha cercato di aggiornare questa tradizione sfruttando le nuove armi offerte dalla multicodicalità, dall'ipertestualità, dall'interattività, nonché – sul piano della descrizione e categorizzazione dei contenuti – dalle ontologie e dai metadati, al centro di progetti come quelli del web semantico e dei linked data».[5] La necessità di repertoriare e inventariare i diversi enti e i loro attributi ci ha offerto la possibilità di interrogarci sui livelli gerarchici non solo di organizzazione della conoscenza del dato ma anche su un piano teorico di rappresentazione del sapere. E dunque di potenzialità degli sviluppi della ricerca. Si è finito per capovolgere la struttura del catalogo tradizionale, che ordina oggetti (le stampe) e ne riassume i contenuti. Liberi dal vincolo gnoseologico del sensibile abbiamo potuto immaginare uno schema che mettesse al centro oggetti che la realtà relega in secondo piano o lega a contingenze fenomenologiche. Abbiamo così creato gallerie di illustrazioni in qualche misura senza spazio (della pagina) e senza tempo (stampate in una edizione e poi ristampate identiche in un'altra a distanza di anni). Abbiamo lavorato sul concetto di filiazione dell'immagine, sia dal punto di vista pratico che dal punto di vista ideale, agevolati dalla possibilità di affrontare e confrontare illustrazioni differenti (miniature, xilografie e disegni fino al manga) per verificare il grado di interferenze e interconnessioni. Il risultato non è un particolare ordine, ne un semplice rovesciamento gerarchico, ma a tutti gli effetti un incremento ontologico. L'immagine, slegata dai suoi supporti materiali e avvicinata a enti simili, ha finito per divenire una nuova entità, che infatti nuovamente – e adamiticamente – è nominata. Si conosce e riconosce sempre su base sensibile, ma è diversamente percepita per via del contesto all'interno del quale la si osserva e delle informazioni che la accompagnano. Lo studioso, in altre parole, si trova di fronte a nuovi oggetti di ricerca, fortemente potenziati dal punto di vista dei dati a corredo e delle possibilità di

[4] Si vedano almeno Pinotti, 2021 e Crupi, 2022.

[5] Roncaglia, 2023: 12.

abbinamento che la macchina è in grado di offrire. Insomma, ci appare sempre più evidente che il mondo della ricerca che sfrutta le DH rappresenti un mondo parallelo rispetto agli studi ante DH perché intrinsecamente diverso; popolato cioè da enti che non coincidono con quelli della ricerca tradizionale e che permettono nuove riflessioni e nuove forme di sapere.

6. DANTE JUYŌ E D:VERSE

Con Dante Juyō abbiamo esplorato un aspetto "estremo" della ricezione della *Commedia*: la fortuna di Dante in Giappone. L'Oriente rappresenta un confine che Dante è riuscito a oltrepassare grazie all'universalità dei contenuti e al fascino senza tempo del suo viaggio oltremondano. Un caso di studio di estrema rilevanza è rappresentato dal Giappone, dove la cultura jpop (manga, anime, videogames) e l'arte contemporanea (es. pittori "danteschi" come Fukuzawa Ichiro e Kazumasa 'Dante' Chiba) dialogano costantemente con l'ispirazione della *Commedia*. L'ingresso del poema in Giappone è avvenuto già nel Seicento con le prime stampe italiane del testo (incunaboli e cinquecentine) forse ad opera dei missionari gesuiti e poi con l'imposizione dell'assimilazione di un canone di autori cardine della cultura europea durante l'epoca Meiji, che ha prodotto traduzioni, rielaborazioni e adattamenti. Lo scopo del progetto è organizzare la rete di input (item relativi alla *Commedia* e a Dante che entrano in Giappone) e di output (item ispirati alla *Commedia* e a Dante prodotti in Giappone); per questo si struttura come una digital library and archive costruite con il content manager Omeka S, i cui metadati RDF e standard DublinCore permettono una gestione semantica delle risorse e una piena interoperabilità con archivi e biblioteche internazionali.

L'adozione di Omeka S si inscrive in una più ampia scelta progettuale fondata su software open source, standard condivisi e accesso aperto: ciò garantisce trasparenza metodologica, sostenibilità a lungo termine e la riusabilità del sistema.

Omeka S consente inoltre l'integrazione nativa con lo standard IIIF, permettendo la visualizzazione di immagini ad alta risoluzione. La struttura della piattaforma, completamente open access, consente una fruizione estesa e partecipativa: ogni pagina è dotata di uno spazio per segnalazioni dal pubblico, in un'ottica di citizen scholarship e partecipazione collettiva alla mappatura delle occorrenze dantesche nella cultura giapponese. Ogni item (stampa antica, manga, mostra pittorica, rappresentazione teatrale, traduzione ecc.) godrà così delle relazioni con altri, interni ed esterni al nostro archivio, grazie ai vantaggi dell'interoperabilità del sistema adottato. La scelta di puntare su un sistema open source e open access condiviso con tutto il mondo delle biblioteche e degli archivi ci permetterà di migliorare l'interoperabilità e i collegamenti tra le varie risorse del web. Questo ci consente di dialogare con l'esterno, in un'ottica di espansione verso altre aree geografico-culturali e di possibilità di replicazione per altre opere fondanti.

Questo sub-progetto ha assistito alla cooperazione di gruppi di lavoro misti italo-giapponesi, i cui componenti possedevano talvolta approcci metodologici molto distanti e diversificati, per i quali è stato possivbile trovare un punto d'incontro proprio grazie all'utilizzo di uno strumento "prêt-à-porter". Inoltre, per questo progetto la collaborazione con il public-at-large diviene fondamentale: ogni pagina del sito contiene uno spazio dove è possibile fare segnalazione di una citazione dantesca (nei videogames, nelle pubblicità, nei manga ecc.).

Concludo spendendo qualche parola sul neonato D:verse (Dante in Vernacular Experience: Readings, Studies, Exegesis). Gli studi raccolti sulle traduzioni otto e novecentesche della *Commedia* nei dialetti italiani ha dato vita a questo ulteriore progetto che raccoglierà il corpus completo di circa 150 traduzioni (integrali e di singoli canti) della *Commedia* in varietà regionali dell'italiano, digitalizzate e annotate.

- Edizioni digitali critiche con note filologiche, glossari e metadati.
- Trascrizioni automatizzate e verificate (HTR - Handwritten Text Recognition).
- Dati linguistici strutturati (lemmatizzazione, POS tagging, metadati geolinguistici).

Si tratta del progetto in ordine di tempo più giovane e per questo ancora meno sviluppato. In particolare i tag XML per il tagging linguistico sono ancora in via di definizione. Tuttavia questi progetti rappresentano l'approdo, nella nostra personale esperienza, a DB basati su standard condivisi e dunque aperti, condivisibili e accessibili nei dati e nei risultati della ricerca. Anche in questo caso lo sforzo di strutturazione della conoscenza (nel caso di Dante Juyō molto disomogenea perché comprende item che vanno da libri di carta a eventi cinematografici, da relazioni dell'ambasciata a sculture), sebbene all'interno di maglie prestabilite dagli standard Dublin core, ci ha dato l'impressione della costruzione di un universo intellettivo di relazioni e di enti non riproducibile e non corrispondente alla realtà sensibile, ma soprattutto – e proprio in virtù di questo – più potente dal punto di vista della possibile riflessione scientifica.

7. CONCLUSIONI

La continua interrogazione sul fine e sullo sviluppo di questi progetti ci obbliga a riflettere sempre più a fondo sulla conoscenza, sulle capacità delle DH di aumentare l'informazione specialistica dando allo studioso basi di partenza sempre più raffinate per i suoi studi.
Credo che sia evidente a chiunque abbia, come me, contribuito alla progettazione teorica di un DB, che i dati immessi e i risultati ricavabili dalla loro interazione partecipino di una natura intellettiva che necessita di un corrispondente paradigma gnoseologico. Analogamente a quanto prospetta Platone nell'arcinoto mito della caverna, è forse possibile anche in questo caso immaginare di distinguere un livello di conoscenza sensibile e uno astratto, che pertiene al mondo delle idee, che però dipende dal taglio specifico che ogni studioso concepisce per il suo lavoro di ricerca ed è per questo, franto nelle molteplici prospettive individuali. Per fare un esempio tolto dai progetti presentati: sotto l'etichetta *limina* (appunti testuali secondo il Dublin Core) faremo rientrare sia le *maniculae* (disegnino di una mano che indica un passo) che le note di possesso (dichiarazioni di acquisizione e proprietà del codice) che troviamo nei manoscritti medievali. Questi progetti ci aiuteranno quindi a riflettere su come gli enti prodotti possano entrare in contatto virtuoso gli uni con gli altri, non solo da un punto di vista del sapere scientifico, ma anche dal rispetto culturale. La condivisione deve infatti produrre associazione, che è il primo frutto di una valida e robusta intelligenza. Senza di essa nessun progresso è possibile. Se il linguaggio è infatti condiviso dal punto di vista del lessico (e in parte della sintassi) si dovrà lavorare sulla semantica, che consentirà di compiere inferenze valide. La nominazione dei dati (che sono per loro natura astratti) non coincide infatti con la nominazione delle cose reali e dunque l'ontologia informatica finisce per rappresentare una metafisica individuale (a seconda dei progetti), svincolata dalla realtà sensibile, che ha in compenso una grande capacità informativa ma scarse potenzialità associative. Credo che lavorare su questi aspetti, a partire da esperienze pratiche come DANTELIMINA, permetterà di dare al Semantic Web e ai LOD maggiori possibilità di intervento.
Su un piano diverso, questi progetti ci hanno spinti a ragionare su come le DH possano diffondere la conoscenza ed essere fortemente inclusive, attraverso la tutela del patrimonio (conservazione di patrimoni manoscritti e a stampa *in primis*) e crescita della consapevolezza delle proprie radici attraverso la capacità di coinvolgere, incuriosire, insegnare a studenti e pubblici di diverso tipo.
Inoltre, il soggetto alla base dei nostri studi accademici, la *Commedia* di Dante, ci costringe a riflettere sull'inclusione da un ulteriore punto di vista. Spesso le iniziative informatiche su Dante e il suo poema, data la portata e la vastità degli studi coinvolti, si sovrappongono (anche se a volte solo in parte) senza riuscire a integrarsi. La ricerca di partenariati e collaborazioni può essere complessa e guardata con diffidenza. Il risultato è che la dispersione delle energie per lo sviluppo di progetti simili e l'assenza di scambio di idee e di risorse può dar vita a un caos di archivi dispersi e non dialoganti. Senza nulla togliere alla – fondamentale – ricerca e metodologia individuale, l'auspicio è di veder nascere una infrastruttura in grado di disciplinare e orientare le molteplici iniziative intorno a temi come Dante (e ad altre grandi figure del "canone *mondiale*"). Assegnare a consorzi e federazioni *ad hoc* progetti "neutri" (come la digitalizzazione dell'intero corpus mss. o a stampa dell'opera) servirebbe a creare la base per progetti con diverse declinazioni che si sviluppino e si integrino a partire da un centro condiviso.
In attesa che l'inclusione scientifica si realizzi a pieno, i prossimi anni ci vedranno impegnati a migliorare le nostre strutture (da database relazionali vorremmo passare a dataset LOD) e a implementare gli strumenti dell'AI per rendere più performativi i nostri archivi digitali, creando mappe dinamiche della ricezione letteraria e culturale. I prossimi obiettivi riguardano infatti le potenzialità dell'utilizzo sinergico dei tanti dati raccolti, e come si potrebbero sviluppare, attraverso l'AI, paradigmi di analisi predittiva e modelli statistici avanzati in grado di correlarli con eventi storici, movimenti culturali o cambiamenti linguistici.
Non dobbiamo dimenticare che il database rappresenta, con le parole di Lev Manovich, «una nuova forma simbolica, […] un nuovo modo di strutturare la nostra esperienza per noi stessi e per il mondo»[6] e che di questa esperienza dobbiamo fare tesoro per noi stessi e dono per gli altri, in un inesauribile circolo virtuoso.

[6] Manovich, 2002: 274.

BIBLIOGRAFIA

Crupi, G. (2022), Presentazione del convegno "Fare per non sprecare". Nei laboratori del riuso digitale (Roma, giugno 2022).

Malatesta, S., & Tonello, E. (2024). DanteMatrix: un software per le indagini stemmatiche sui manoscritti della «Commedia» di Dante. DigItalia, 19 (2), 191–201.

Manovich, L. (2002). Il linguaggio dei nuovi media. Milano: Olivares.

Pinotti, A. (2021). La storia dell'arte nello specchio dei visual Studies e delle Digital Humanities. Intersezioni, 3, 349-363.

Renello, G. P. (2013). Un programma per la classificazione "computer-assisted" delle copie della «Commedia» e di altre tradizioni sovrabbondanti. Nuove prospettive sulla tradizione della «Commedia». Seconda serie (2008-2012), Tonello, E. & Trovato P. (Eds). Padova: libreriauniversitaria.it. 207-222.

Roncaglia, G., (2023) L'architetto e l'oracolo. Forme digitali del sapere da Wikipedia a ChatGPT. Bari-Roma: Laterza.

Tonello, E., & Trovato, P. (2022) (Eds.) Dante. Commedia. Inferno. ed. critica. I vol. Padova: libreriauniversitaria.it.

Tomasi, F., & Vitali, F., (2023) Semantic Web, Linked Data e beni culturali. Digital Humanities. Metodi, strumenti, saperi. Ciotti, F. (Ed.). Roma: Carocci. 137-159.

TEI Encoding as a Unified Structure for Multilingual Digital Editions: The *LeggoManzoni* Case Study

Mariia Levchenko[1], Beatrice Nava[2], Ersilia Russo[3]
[1] Università "Alma Mater" di Bologna, Italy- maria.levchenko@studio.unibo.it
[2] Universität Wien, Austria - beatrice.nava@univie.ac.at
[3] Università degli Studi di Firenze, Italy- ersilia.russo@unifi.it

ABSTRACT (ENGLISH)

This paper presents *LeggoManzoni*, a digital edition of Alessandro Manzoni's *I promessi sposi* that tests the effectiveness of TEI encoding as a unified framework for managing complex textual relationships. The project aligns forty Italian commented editions from 1893 to 2021 and fourteen translations in ten languages (1845-2022) with the original text of the 1840-42 edition. We describe an investigative stand-off markup approach that keeps the source text separate from commentaries and translations while maintaining accurate references through automated alignment pipelines. The commentary pipeline processes digitized historical editions and implements a flexible text-matching algorithm with a 70% match threshold, achieving 87% accuracy on 2.441 processed comments. For translations, we use the Bertalign algorithm, configuring broader alignment parameters to account for cross-linguistic variation. Both pipelines generate TEI-compliant files using the same encoding pattern based on <note> elements with word-level references, allowing easy integration through XSLT transformations. The outcome of the project demonstrates TEI's ability to serve as a sustainable foundation for multilingual digital editions, bridging scholarly research and educational applications while ensuring long-term preservation through convertibility to emerging formats.
Keywords: *I promessi sposi*; XML/TEI; Multilingual Digital Editions; stand-off markup; text alignment

ABSTRACT (ITALIANO)

La codifica TEI come struttura unificata per le edizioni digitali multilingue: Il caso di studio LeggoManzoni*.* Questo articolo presenta *LeggoManzoni*, un'edizione digitale de *I promessi sposi* di Alessandro Manzoni che testa l'efficacia della codifica TEI come framework unificato per la gestione di relazioni testuali complesse. Il progetto allinea quaranta edizioni commentate in italiano dal 1893 al 2021 e quattordici traduzioni in dieci lingue (1845-2022) con il testo originale dell'edizione del 1840-42. Adotta un markup stand-off investigativo che tiene separato il testo di partenza dai commenti e dalle traduzioni, mantenendo riferimenti accurati attraverso pipeline di allineamento automatizzate. La pipeline dei commenti elabora le edizioni storiche digitalizzate e implementa un algoritmo flessibile di corrispondenza del testo con una soglia di corrispondenza del 70%, raggiungendo un'accuratezza dell'87% su 2.441 commenti. Per le traduzioni, utilizziamo lo strumento Bertalign, configurando parametri di allineamento più ampi per tenere conto delle variazioni linguistiche. Entrambe le pipeline generano file conformi a TEI utilizzando lo stesso modello di codifica basato su elementi <note> con riferimenti a livello di parola, consentendo una facile integrazione attraverso XSLT. Il risultato del progetto dimostra la capacità di TEI di fungere da base sostenibile per le edizioni digitali multilingue, stabilendo un legame tra la ricerca scientifica e le applicazioni didattiche e garantendo al contempo la conservazione a lungo termine attraverso la convertibilità in formati in espansione.
Parole chiave: *I promessi sposi*; XML/TEI; edizioni digitali multilingue; stand-off markup; allineamento testuale

1. INTRODUCTION. *LEGGOMANZONI*: OLD ROOTS, NEW LEAVES

Our community is nowadays deeply aware of the potential of digital editions and knowledge sites (Tomasi, 2016, p. 133). These resources serve as repositories of vast amounts of information, presented at various levels and tailored to diverse users - humans or machines. We are of course also fully committed to addressing the challenges that these resources pose. This proposal seeks to highlight the constructive potential of digital editions by presenting *LeggoManzoni* as a successful case study of an edition project. The project is examined from a twofold perspective. First, we provide an overview of its architecture and content, focusing on the output of the project, the digital environment *LeggoManzoni*.[1] Second, we explore

[1] https://projects.dharc.unibo.it/leggomanzoni/ (cons. 25/01/2025).

how TEI encoding can serve as a unified framework for creating Digital Editions with commentaries, as well as Multilingual Digital Editions.
LeggoManzoni was conceived by Paola Italia and Francesca Tomasi (FICLIT Department of the University of Bologna) and launched in 2019 as part of the PRIN project *Manzoni Online: unpublished manuscripts and documents, tradition and translation*.[2] Built upon a collaboration between the University and 20 Italian high schools, the project has resulted in a comprehensive digital environment that provides access to the complete text of the 1840–42 edition of *I promessi sposi* (*The Betrothed*). The edition is enriched with critical commentaries, translations and a digitized version with metadata of the original edition. *LeggoManzoni* therefore provides access to the novel from multiple perspectives, allowing users to engage with the masterpiece both as it was originally conceived by the author and as it has been interpreted by scholars over time. The purpose of this resource is to present Manzoni's well-known work and his cultural legacy in a new format, integrating various elements - such as commentaries, translations, and illustrations - to enhance the experience of reading and studying the text. The very construction of the project, which has evolved over time, has in fact also enabled technical experimentations that could serve as a starting point for the development of similar projects, adhering to best practices for reusing data, ideas, and code. Specifically, it has created opportunities to develop and apply digital methodologies and technologies for encoding, integrating, and aligning texts, through the design of an automated and reproducible pipeline. What the project actually aims to achieve is to create new leaves from old, established roots, both in terms of knowledge of Manzoni's work - for scholars as well as for schools - and in terms of digital technologies and methodologies.

2. THE TEI ENCODING FRAMEWORK

At the core of the edition lies a TEI framework. Both the text of the novel and the commentaries are encoded according to this standard, which serves as the foundation of the whole infrastructure, informing the entire pipeline. The main concern when modelling the connection between the novel's text and the various critical commentaries was, of course, managing overlaps. In fact, the notes from different commentaries often refer, as expected, to the same portions of text or to overlapping sections. To address this, we chose a stand-off markup approach, which keeps the novel's text separate from the commentaries while allowing references between the text file and each individual commentary encoded as a separate file, thus avoiding prohibited nesting and intersection of tags. The encoding model is therefore based on creating separate XML/TEI files for each chapter of the novel along with a rough tokenization of each chapter's text using a simple XSLT stylesheet.[3] The stylesheet tags each word in the novel (keeping punctuation with the preceding word) as a <w> element, which includes a unique identifier (@xml:id attribute), enabling cross-referencing from the commentary files. An encoded sample word appears as: <w xml:id="c1_10001">Quel</w>, where the first part of the @xml:id refers to the chapter, and the second part numbers the word in sequence.
For the commentaries, the encoding model uses <note> elements to store essential identification information (@xml:id and @type attributes) and links to the corresponding text (@target and @targetEnd attributes). Each note contains two child elements: a <ref> that holds the plain-text reference to the portion of text being commented, followed by a colon, and the plain-text commentary itself. A commentary entry looks as follows:

```
<note xml:id="Russo_cap1-n51"
     type="comment"
     target="quarantana/cap1.xml#c1_13095"
     targetEnd="quarantana/cap1.xml#c1_13101">
   <ref rend="bold">chiudendo il libro con le due mani</ref>:
   Il breviario era rimasto dunque aperto, durante il colloquio…
</note>
```

More in detail, each <note> element includes a unique identifier (@xml:id) combining the commentator's name, chapter, and entry number; the type attribute specifying it as commentary (type="comment"); a precise reference link to the source text through @target and @targetEnd attributes, using word-level IDs

[2] https://www.alessandromanzoni.org/ (cons. 25/01/2025).
[3] We didn't use a specific tokenization software since a perfect accuracy of this step was not required (see below the commentary pipeline) and a custom XSLT easily served the scope.

from the original TEI files and, finally, the quoted text wrapped in a <ref> element with bold rendering, plus the commentary entry following the quoted passage.
The commentaries' texts have been enriched with other information such as place names, bibliography references, placeholders for images, internal references from one note to another and so on.[4] However, the simple structure described above is the key concept we want to focus on, since it allowed us to define and build the automatic alignment mechanism described in the next section.
The alignment system has been created for the massive corpus of forty commented editions of *I promessi sposi*, representing almost a century of Italian literary scholarship, from Policarpo Petrocchi's 1893-1902 to Raimondi and Bottoni's 2021. The collection represents the work of major Italian literary scholars and includes both academic and educational editions from prominent Italian publishers.
Currently the project has completed the alignment and encoding of the Introduction and Chapter I for all forty editions, with two editions - Luigi Russo (1935) and Mariarosa Bricchi (1996) - fully completed. Work is underway to extend the alignment to the remaining chapters in all editions. The temporal distribution of these editions is significant: from the historical publications beginning with Petrocchi in 1893 through to contemporary critical works in the 2020s, demonstrating the breadth of critical perspectives and interpretative approaches to the text.

3. COMMENTARY PIPELINE

The preparation of the commentaries follows a modular approach in which each commented edition of *I promessi sposi* is processed in parallel, maintaining a clear separation between the source text and its commentaries. The process starts with the digitised historical editions, which are OCR'd into plain text files, organised chapter by chapter. Each chapter's commentary is stored in a separate .txt file, with each line following this format: "commented text": corresponding commentary.
The alignment process uses the TEI-encoded chapters of the novel as reference points but maintains the separation of content. The system reads the TEI files to create a normalised word index, which serves as the basis for identifying the exact locations of annotated passages. The alignment algorithm uses a flexible text-matching technique that accounts for OCR variations and encoding differences common in digitised historical texts. It can in fact tolerate small gaps (up to two words) and requires a match rate of 70%, making it robust enough to handle textual variation while maintaining accuracy.
Once the text fragments have been aligned, the system generates new TEI files for the chapter commentaries. Each file contains a series of note elements structured according to the encoding pattern defined in the encoding model, namely a note with a starting and ending point (see above section 2). This encoding structure maintains precise alignment with the source text while preserving the scholarly apparatus of the commentary. The reference system allows the exact location of commented passages while maintaining the independence of the commentary files from the base text.
The resulting TEI files contain full metadata about the source edition, including bibliographic and editorial information. This separation of content - retaining the original TEI-coded novel text while creating separate, linked commentary files - has several advantages. It allows for easier management of multiple commented editions, independent updating of either text or commentary, and flexible presentation options in the digital edition interface using XSLT files.
Alignment Accuracy Analysis. The performance of the alignment algorithm was evaluated across all forty commented editions, processing a total of 2,441 comments. Analysis of the alignment results showed an overall error rate of 13.03%. The errors fell into several different categories: 44 cases (1.8%) where the start position could not be identified, 44 cases where the end position was missing, 152 cases (6.2%) where both start and end positions were not identified, and 78 cases (3.2%) where the algorithm linked to incorrect occurrences of text spans appearing more than once in the chapter.
These results highlighted specific challenges in processing comments from the raw text. This error rate motivated the development of a web-based validation interface designed to facilitate the efficient identification and correction of misaligned comments. The interface provides specific tools for handling each type of error, allowing annotators to quickly locate problematic alignments and make the necessary corrections while preserving the TEI structure.

[4] The model was initially tested through the manual encoding of a selection of commentaries, thanks to the collaboration with the high schools previously mentioned. Students were tasked with encoding the structure of the commentary for a single chapter after receiving an introduction to Digital Scholarly Editing, as well as encoding languages and practices, during three PCTO (Pathways for Soft skills and Orientation) organized by the University of Bologna. Details on the schools involved in the project can be found here: https://projects.dharc.unibo.it/leggomanzoni/progetto.

Web-Based Alignment Management System. The alignment system was implemented as a Next.js web application that provides an intuitive interface for both automatic and manual alignment of comments (see Fig. 1). This interface integrates with the backend alignment service, while giving annotators full control over the alignment process and validation.



Figure 1. User Interface for Comment Alignment (https://manzoni-comments-aligner.vercel.app)

The workflow begins with the annotator selecting a chapter and uploading a text file containing comments for that chapter. Each comment in the file follows the standard format of quoted text followed by commentary text, separated by a colon. The system processes this input through the alignment algorithm and presents the results in an interactive interface. The interface displays the TEI text of the novel chapter alongside the processed comments, with clear visual indicators of alignment status. Comments requiring manual intervention are visually highlighted in red. For each comment, the system displays both the referenced text and the corresponding comment, along with the precise alignment positions derived from the TEI word identifiers.

When manual alignment is required, annotators can use an interactive text selection mode. They click on a comment that needs correcting, then select the correct text passage directly in the TEI content view. The system automatically updates the alignment references, maintaining the proper TEI linking structure. This combination of automatic alignment and manual validation ensures both efficiency and accuracy.

Once all alignments have been verified, the system generates a new TEI file for the chapter commentary. A metadata dialogue allows annotators to select the specific edition details from the *LeggoManzoni* project catalogue and provide their attribution information. The resulting TEI file contains proper namespaces, reference systems and metadata according to the project's specifications.

This web-based tool significantly streamlined the processing of forty historical commentaries by providing an efficient interface for both automatic alignment and manual verification, while ensuring consistent TEI encoding across all processed editions.

4. TRANSLATION ALIGNMENT PIPELINE

The translation alignment pipeline for the *LeggoManzoni* project has been implemented on a multilingual corpus of translations from 1845 to 2022, including English (1845, 1876, 1972, 2022), French (1874), German (1880), Spanish (1858), Dutch (1849), Polish (1882), Russian (1854, 1936, 1999), Ukrainian (1982), Chinese (1998), Finnish (1910). The process uses Bertalign (Liu & Zhu, 2022), a sentence alignment module that uses multilingual sentence transformer models, in this case the Language-agnostic BERT Sentence Embeddings (LaBSE) (Feng et al., 2020) model, to produce accurate alignments between the Italian source text and its translations.

The pipeline starts with text pre-processing, where both the Italian source text (Quarantana) and each translation are normalised and segmented. The core alignment process takes place in two stages. The first stage identifies basic sentence pairs, establishing initial correspondences between the source and translated texts. The second stage implements a more sophisticated alignment algorithm that can handle complex translation patterns, including one-to-many, many-to-one and many-to-many sentence relationships. This two-stage approach allows the system to handle structural differences between languages while maintaining semantic correspondence. The alignment system uses a maximum alignment window of 8 sentences, 5 nearest target neighbors and a sliding window of 5 sentences for context, ensuring both precision in local alignments and consideration of the wider textual context. While Sprugnoli & Sartor (2023) used more restrictive parameters (maximum alignment types: 6, k nearest neighbors: 3, search window: 5) for aligning the Ventisettana with its English translation, our multilingual pipeline requires broader settings to accommodate greater syntactic variation across languages.

Following the same TEI encoding pattern established for commentaries, the system generates XML files in which each aligned segment is encoded as a <note> element with a unique identifier that combines language, chapter, and segment number (see Levchenko, 2024 for text segmentation details). Using the same target attribute structure to reference the source text enables precise word-level linkage between the original text and the translations. Like DiScEPT's approach using the <annotation> element for stand-off markup (Hohenegger et al., 2024), this unified TEI approach allows translations to be seamlessly integrated into the existing web interface, while maintaining comprehensive metadata about each edition and taking advantage of independent file management and flexible presentation options through XSLT transformations.

5. CONCLUSIONS

In conclusion, the project led to the creation of an enriched edition of Manzoni's text and gave the opportunity for experimentation with automated pipelines for text alignment, starting from a simple XML/TEI structure. Despite common critiques of TEI's complexity, our implementation proved highly effective in two key areas: (1) enabling precise and flexible alignment between source text, commentaries, and translations through automated pipelines, and (2) bridging academic research and classroom education through an accessible digital interface, thanks to the mentioned PCTO (and, hopefully, the edition will remain a useful teaching tool: see Levchenko, Menna & Nava, 2024). The automated alignment pipelines and web-based validation tools we developed provide a reproducible methodology that can serve as a model for similar digital edition projects. While digital humanities may shift toward newer formats like JSON, TEI's well-structured nature ensures easy conversion of editions built over past decades to emerging standards, preserving the scholarly value of digital cultural heritage.

REFERENCES

Feng, F., Yang, Y., Cer, D., Arivazhagan, N., & Wang, W. (2022). Language-agnostic BERT sentence embedding. Muresan, S., Nakov, P., & Villavicencio, A. (Eds.), Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin: Association for Computational Linguistics, 878–891. https://doi.org/10.18653/v1/2022.acl-long.62.

Hohenegger, H., Mancinelli, T., Ciotti, F., Boschetti, F., Del Grosso, A. M., & De Longis, E. (2024). Paul Klee, Tunisreise e Bildnerische Formlehre: un caso studio di DiScEPT (Digital Scholarly Editions Platform and Aligned Translations). Di Silvestro, A., & Spampinato, D. (Eds.), Me.Te. Digitali. Mediterraneo in rete tra testi e contesti. Proceedings of the 13th Annual Meeting of AIUCD 2024, Catania 28-30 May 2024. Catania: AIUCD2024, 186-190. https://doi.org/10.6092/unibo%2Famsacta%2F7927.

Levchenko, M. (2024). Automatic Translation Alignment Pipeline for Multilingual Digital Editions of Literary Works. Proceedings of the Computational Humanities Research Conference 2024. Aarhus, Denmark, December 4-6, 2024, 1086-1104. https://doi.org/10.48550/arXiv.2410.13255.

Levchenko, M., Menna, G., & Nava, B. (2024). *LeggoManzoni*. Quaranta commenti alla Quarantana. Russo, E. (Ed.), Manzoni e Leopardi in digitale. Idee e proposte per la scuola. Bologna: Clueb, 73-86. ISBN: 978-88-491-5800-7.

Liu, L., & Zhu, M. (2023, june). Bertalign: Improved word embedding-based sentence alignment for Chinese–English parallel corpora of literary texts. Digital Scholarship in the Humanities, 38, 2, 621-634. https://doi.org/10.1093/llc/fqac089.

Sprugnoli, R., & Sartor, M. (2024). That branch of the lake of Como...: Developing a new resource for the analysis of *I Promessi Sposi* and its historical translations. Proceedings of the 9th Italian Conference on Computational Linguistics (CLiC-it 2023). Torino: Accademia University Press, 421–427. ISBN: 9791255000846. https://ceur-ws.org/Vol-3596/paper48.pdf.

Tomasi, F. (2016). Edizioni o archivi digitali? Knowledge sites e apporti disciplinari. Bonsi, C., & Italia, P. (Eds.), Edizioni Critiche Digitali Digital Critical Editions: Edizioni a confronto / Comparing Editions. Roma: Sapienza Università Editrice, 130-136. https://doi.org/10.13133/9788893770033.

Verso l'implementazione di un sistema di riconoscimento di allusioni al lessico dantesco nelle testimonianze del Lager: il caso d'uso in Voci dall'Inferno

Carla Congiu[1], Angelo Mario Del Grosso[2], Marina Riccucci[3]
[1] Università di Pisa, Italia- c.congiu1@studenti.unipi.it
[2] CNR-ILC, Italia – angelomario.delgrosso@cnr.it
[3] Università di Pisa, Italia- marina.riccucci@unipi.it

ABSTRACT (ITALIANO)

Voci dall'Inferno è un progetto di ricerca dell'Università di Pisa, sviluppato con il supporto dell'Istituto di Linguistica Computazionale "A. Zampolli". L'iniziativa ha due principali obiettivi scientifici: a) digitalizzare il primo corpus di testimonianze non letterarie di deportati sopravvissuti ai campi di concentramento e b) identificare al suo interno la presenza di citazioni e/o allusioni al lessico di Dante (Del Grosso et al.,2024). Al fine di raggiungere questo secondo obiettivo è stato sviluppato un prototipo di applicazione web denominata *Voci dall'Inferno Verse Similarity Search*. Il sistema è progettato per individuare citazioni e allusioni al lessico dantesco mediante approcci computazionali alla ricerca di frasi presenti nelle testimonianze e il confronto di essi con i versi presenti nella *Divina Commedia* di Dante Alighieri. L'applicazione, realizzata in Python, utilizza tecnologie avanzate come Weaviate, una piattaforma open-source per la ricerca vettoriale, e Streamlit, un framework per lo sviluppo di applicazioni web. Basandosi su metriche di *Sentence Similarity*, l'applicazione sfrutta modelli di machine learning per trasformare i testi in rappresentazioni di *embeddings* e in seguito misurarne la similarità. Attualmente l'applicazione non è ancora disponibile per l'uso da parte del pubblico, ciononostante l'infrastruttura di ricerca CLARIN-IT (H2IOSC) è stata contattata per ospitare l'applicazione garantendone accesso e sostenibilità. Una demo sarà predisposta per la conferenza qualora il contributo venisse accettato.
Parole chiave: Sentence Similarity; Sentence Transformers; vector database; embeddings; Voci dall'Inferno

ABSTRACT (ENGLISH)

Toward the implementation of a system for recognizing allusions to Dante's lexicon in Lager testimonies: the Voci dall'Inferno use case. Voci dall'Inferno is a research project by the University of Pisa, developed with the support of the Istituto di Linguistica Computazionale "A. Zampolli". The initiative has two main scientific objectives: a) to digitize the first corpus of non-literary testimonies from concentration camp, and b) to identify the presence of citations and/or allusions to Dante's lexicon within them (Del Grosso et al., 2024). To achieve this second objective, a prototype web application called *Voci dall'Inferno Verse Similarity Search* was developed. The system is designed to detect citations and allusions to Dante's vocabulary through computational approaches by searching for expression within the testimonies and comparing them with verses from Dante's Commedia. The application, built in Python, leverages advanced technologies such as *Weaviate*, an open-source vector search platform, and *Streamlit*, a framework for web application development. Adopting sentence similarity metrics, the application uses machine learning models to transform texts into embedding representations and subsequently measure their similarity. Currently, the application is not yet publicly available. However, the CLARIN-IT research infrastructure (within H2IOSC PNRR project) has been contacted to host the application, ensuring accessibility and sustainability. A demo will be prepared for the conference if the contribution will be accepted.
Keywords: Voci dall'Inferno; Sentence Similarity; Sentence Transformers; vector database; embeddings

1. INTRODUZIONE

Voci dall'Inferno è un progetto di ricerca dell'Università di Pisa, curato da Marina Riccucci (Università di Pisa) dal 2016 con il supporto di Angelo Mario Del Grosso (CNR-ILC, Pisa).
Nel corso degli ultimi anni il gruppo di lavoro ha avviato la digitalizzazione e la codifica, mediante lo standard XML-TEI (Burnard 2014), del primo corpus di testimonianze non letterarie provenienti dai campi di concentramento e ha sviluppato una applicazione web mediante la piattaforma eXist-db per la consultazione e la fruizione del repertorio (Del Grosso et al, 2024b). Il corpus ad oggi consta di 23 testimonianze appartenenti a 18 testimoni (vedi Figura 1) suddivisi in 17 deportati nei Lager, di cui 12 ebrei e 5 internati militari italiani, e 1 perseguitato, vale a dire il partigiano ebreo Emanuele Artom (Del Grosso et al., 2024). Di queste testimonianze oltre al recupero e alla digitalizzazione delle fonti scritte –

quasi tutte conservate presso il centro di Documentazione Ebraica Contemporanea di Milano oppure in fondi privati di familiari ed eredi dei testimoni (ad esempio i diari e le lettere di Nicola Ricci, di Luigi Giuntini, di Emanuele Artom, di Alberto Pacini e Romana Feld) vi sono molteplici testimonianze rilasciate in forma orale spesso conservate in microcassette e supporti non digitali (vedi Figura 2). Attualmente più di 18 ore di parlato sono presenti nell'archivio del progetto - per la maggior parte provenienti dal fondo inedito Segre-Pavoncello (13 testimonianze del fondo Segre-Pavoncello sono state digitalizzate su un totale di circa 25 sopravvissuti).

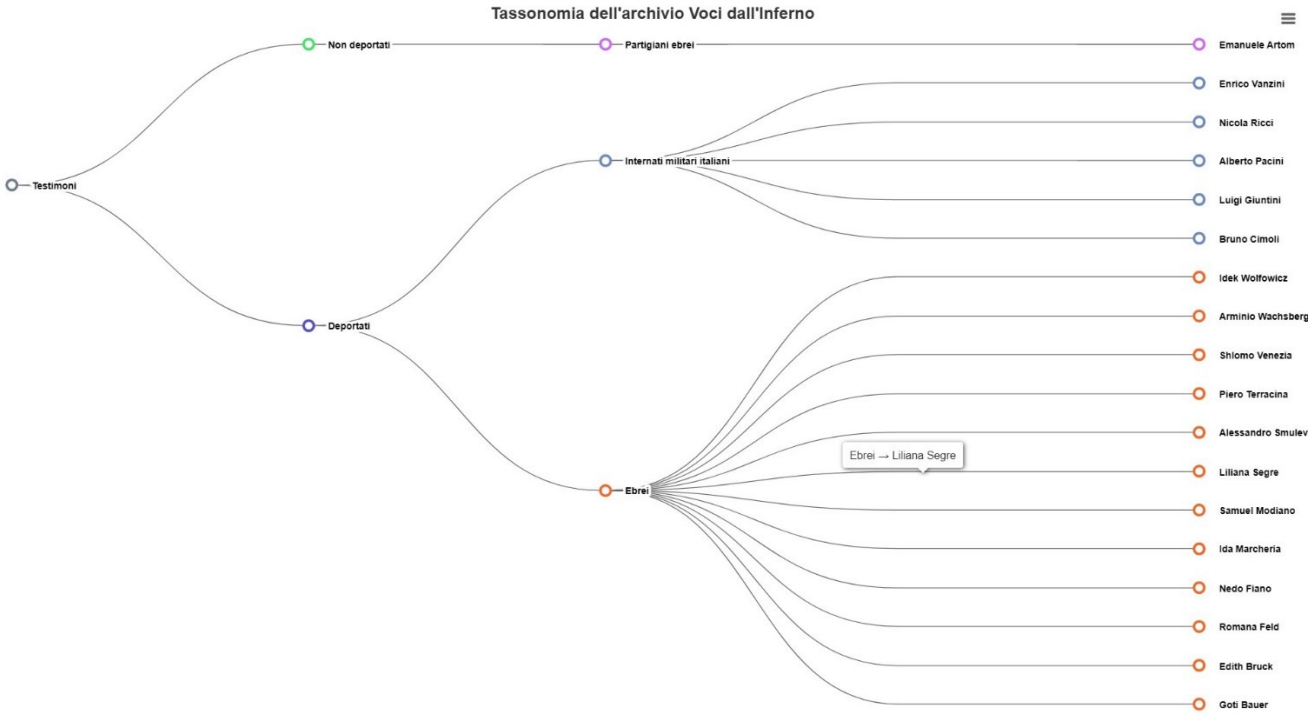


Figura 1. Tassonomia dell'archivio Voci dall'Inferno

Da una prima analisi generale del corpus è emerso che molti testimoni si servono del vocabolario dantesco per riportare le proprie esperienze "ineffabili" (Calderini, 2020). Per studiare le tessere dantesche si sono svolte differenti analisi e una classificazione puntuale in citazioni implicite, citazioni esplicite, allusioni e vocaboli danteschi. I dati che sono emersi dallo studio dell'archivio digitale dimostrano che la maggior parte delle citazioni sono riconducibili alla prima cantica. Nonostante ciò, anche passi del Purgatorio e del Paradiso sono stati individuati, seppur in misura minore (Del Grosso et al., 2024). Ad esempio, nel diario di Giuntini è citato il verso 12 del Canto I della seconda cantica per descrivere l'azzurro cielo estivo privo di nuvole:

> Stamani, guardando il cielo, potrei finalmente descriverlo come padre Dante: «Dolce colore d'oriental zaffiro» tanto è azzurro e senza una nuvola.

Lo studio delle testimonianze all'interno del progetto Voci dall'Inferno mira dunque ad analizzare il lessico e le allusioni dantesche presenti nel corpus, focalizzandosi più di altro sull'adozione del linguaggio dell'*Inferno* della *Divina Commedia* da parte dei sopravvissuti per descrivere le esperienze nei Lager. A supporto di questo obiettivo è stata sviluppata l'applicazione web *Voci dall'Inferno Verse Similarity Search*, finalizzata a identificare citazioni e allusioni al lessico dantesco mediante la ricerca di frasi contenute nelle testimonianze e il confronto con i versi della *Divina Commedia*.
L'applicazione che si introduce in questo contributo prende le mosse dal progetto *Latin Vulgate Verse Similarity Search*, sviluppato da William Mattingly per automatizzare l'identificazione e l'estrazione di potenziali citazioni bibliche in testi latini medievali (Mattingly, 2024).



Figura 2. Donut Chart con la distribuzione delle tipologie di testimonianze all'interno del corpus Voci dall'Inferno

2. METODOLOGIA

La realizzazione dell'applicazione *Voci dall'Inferno Verse Similarity Search* è stata sviluppata usando il linguaggio di programmazione Python e ha richiesto l'utilizzo di diversi strumenti tecnologicamente avanzati (vedi Figura 3). Nello specifico, sono state impiegate due tecnologie: la prima è *Weaviate* (Weaviate, 2025), una piattaforma di vector search open-source che, sfruttando modelli di Machine Learning allo stato dell'arte, trasforma dati non strutturati in un database vettoriale ricercabile sul quale è possibile eseguire interrogazioni e analisi semantiche. Il secondo strumento utilizzato è *Streamlit* (Streamlit, 2025), un framework open-source dell'ecosistema Python che facilita la creazione e lo sviluppo di applicazioni web personalizzate.

L'approccio adottato per l'individuazione dei tasselli danteschi nel corpus di testimonianze fa leva sull'uso di metriche per il calcolo della "Sentence Similarity" (Reimens 2019), ovvero il processo di valutazione della similarità tra due testi costituiti da poche parole che abbiano una coerenza sintattica/semantica. Ciò viene realizzato utilizzando un modello matematico di sentence similarity, che permette di trasformare i dati da testo piano in punti di uno spazio vettoriale misurabili, detti *embeddings* (Cao, 2024). Questo permette di misurare la similarità tra frammenti di testo, calcolando la loro vicinanza "semantica".

Per lo specifico sviluppo del caso d'uso presentato in questo contributo è stato adottato il modello di sentence similarity denominato SentenceTransformers *all-mpnet-base*, pubblicamente accessibile per mezzo della piattaforma *HuggingFace* (Huggingface, 2025).



Figura 3. - Tecnologie e Workflow dell'applicazione

Per il corretto funzionamento dell'applicazione è stato necessario prevedere diverse fasi di sviluppo. Dopo aver raccolto le terzine delle cantiche della Divina Commedia in una struttura tabellare, mediante il modello di SentenceTransformers sono stati calcolati i corrispondenti embeddings. L'utilizzo di Weaviate permette la configurazione di una collezione per archiviare le informazioni strutturate relative alla Divina Commedia e la creazione di un database per categorizzare e archiviare i dati relativi a cantiche, canti, versi e terzine. In questo modo è possibile archiviare gli embeddings ed eseguire ricerche semantiche sui dati. Dopo aver chiamato la funzione che consente il calcolo del punteggio di similarità tra la query ricercata e gli embeddings delle terzine il sistema restituisce i passi più simili ordinati secondo un punteggio di similarità (*similarity score*). Per esempio, data una sequenza di parole quali "vuolsi così colà" le prime due terzine restituite dal sistema sono; 1) "Non impedir lo suo fatale andare: Vuolsi così colà dove si puote ciò che si vuole, e più non dimandare" [Inf. V 22-24]; 2) "E 'l duca lui: Caron, non ti crucciare: vuolsi così colà dove si puote ciò che si vuole, e più non dimandare" [Inf. III 94-96]. Rispettivamente con similarity score 0.63 e 0.62.

3. L'APPLICAZIONE

L'applicazione "Voci dall'Inferno Verse Similarity Search" ha previsto lo sviluppo di uno strato software di front-end dedicato all'interazione web con l'utente studioso. A tal fine è stato adottato il framework Streamlit che dispone l'applicazione sul browser. Essa si presenta con una barra di ricerca nella parte alta della pagina web in cui inserire la query da analizzare e fonte testuale dell'interrogazione alla Commedia dantesca vettorializzata in Weaviate; in aggiunta, l'interfaccia mostra due componenti per la scelta delle sezioni che permettono l'eventuale selezione all'interno della Commedia rispettivamente di cantica e canto; in ultimo un componente slider dedicato alla personalizzazione del numero di risultati da presentare a video dopo l'interrogazione; il tutto è avviato dal pulsante di inoltro della ricerca (vedi Figura 4).

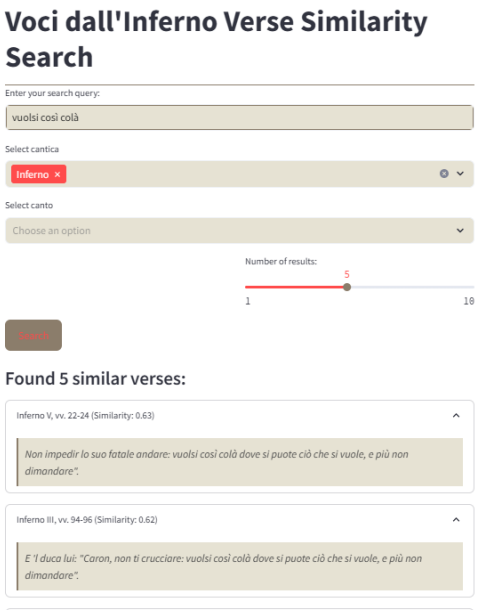


Figura 4. Screenshot dell'applicazione Voci dall'Inferno Verse Similarity Search

4. CONCLUSIONI

Il presente contributo ha illustrato l'idea progettuale e lo stato di sviluppo dell'applicazione "Voci dall'Inferno Verse Similarity Search" dedicata alla ricerca di citazioni alla Commedia di Dante (esplicite e implicite), di espressioni e vocabolario dantesco nonché di allusioni e riferimenti alle cantiche del poeta fiorentino. Il lavoro si inserisce nel più ampio contesto del progetto di ricerca Voci dall'Inferno che ha già raccolto e analizzato il primo corpus italiano di testimonianze non letterarie scritte e orali di sopravvissuti ai Lager nazisti.

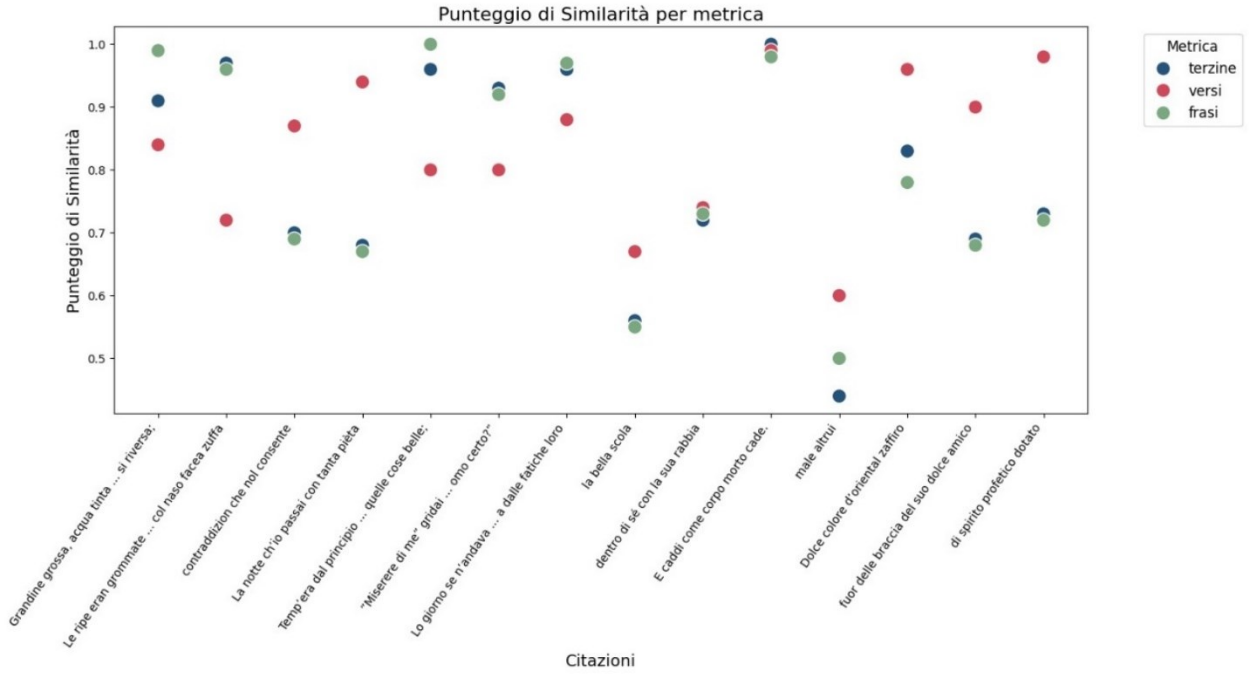


Figura 5. - Diagramma scatter plot che sintetizza i risultati della fase di valutazione dell'applicazione

L'applicazione è stata sottoposta a una fase di valutazione, mostrando un'elevata efficacia nell' individuazione dei versi danteschi in presenza di citazioni esplicite, ossia citazioni più o meno fedeli al testo originale (vedi Figura 5). Per quanto riguarda, invece, le citazioni implicite, ovvero i riferimenti meno immediati e diretti alla Commedia, l'applicazione attualmente sviluppata richiede ulteriori interventi di perfezionamento. In questa direzione, è già in corso una prima fase di *fine-tuning* del modello preposto alla generazione degli *embeddings,* con l'obiettivo di migliorare il reperimento automatico dei versi danteschi sulla base di esempi opportunamente codificati presenti nel corpus Voci dall'Inferno.
Ad esempio, con il modello attuale di *embeddings*, l'applicazione non è stata in grado di riconoscere l'allusione ai versi 109-114 del canto XIII attraverso la seguente espressione: "stare sempre agitato come selvaggina che può essere colta di sorpresa". Il sistema computazionale propone invece, con un indice di similarità pari a 0.69, il passo "Non han sì aspri sterpi né sì folti quelle fiere selvagge che 'n odio hanno tra Cecina e Corneto i luoghi còlti" [Inf. XIII 7-9].
Inoltre, la similarità tra testi viene calcolata considerando tre granularità testuali: 1) le terzine, 2) i singoli versi e 3) le frasi. Questa scelta consente di valutare in modo puntuale la granularità delle citazioni rilevate. La figura 5 mostra anche questa differenza di granularità. In particolare i valori di similarità sono rappresentati tramite piccoli cerchi colorati di colore blu per le terzine, di colore rosso per i versi e di colore verde per le frasi. Dal confronto emerge come i valori di similarità varino in funzione del livello di granularità considerato. Le citazioni più estese, anche meno precise, tendono ad avere valori di similarità più alti quando la ricerca avviene a livello di frase. Al contrario, le citazioni più brevi e aderenti al testo originale ottengono valori di similarità maggiori quando la ricerca avviene a livello di verso.
In conclusione, è attualmente in corso un'interlocuzione con l'infrastruttura CLARIN-IT, nell'ambito del progetto infrastrutturale H2IOSC, con l'obiettivo di individuare la soluzione di hosting più adeguata. Tale attività è finalizzata a garantire all'applicazione accessibilità e fruibilità per gli studiosi di Dante.

RINGRAZIAMENTI

Il Progetto Voci dall'Inferno si è avvalso del supporto dell'infrastruttura di ricerca CLARIN-IT, che ospiterà i dati e il software del progetto per la long-term preservation. Il CLARIN Knowledge Centre DiPText-KC ha messo a disposizione del progetto il proprio gruppo di esperti per consulenze scientifiche e operative (https://diptext-kc.clarin-it.it/helpdesk). Inoltre, si ringrazia la dott.ssa Elvira Mercatanti per i dati forniti e per la revisione del contributo.

BIBLIOGRAFIA

Burnard, L. (2014). What is the Text Encoding Initiative? : How to add intelligent markup to digital resources. OpenEdition Press. https://books.openedition.org/oep/426

Calderini, S., & Riccucci, M. (2020). L'ineffabilità della nefandezza: Dante «per dire» il Lager: un sondaggio preliminare nelle testimonianze non letterarie. *Italianistica*, *1*. https://doi.org/10.19272/202001301011

Cao, H. (2024). *Recent advances in text embedding: A Comprehensive Review of Top-Performing Methods on the MTEB Benchmark*. https://arxiv.org/abs/2406.01607

Del Grosso, A. M., Riccucci, M., & Mercatanti, E. (2024). The Impact of Digital Editing on the Study of Holocaust Survivors' Testimonies in the context of Voci dall'Inferno Project. In I. Anuradha, M. Wynne, F. Frontini, & A. Plum (A cura di), Proceedings of the First Workshop on Holocaust Testimonies as Language Resources (HTRes) @ LREC-COLING 2024 (pp. 1–9). ELRA and ICCL. https://aclanthology.org/2024.htres-1.1/

Del Grosso, A. M., Riccucci, M. & Mercatanti, E. (2024). Voci Dall'Inferno: Dante per Dire Il Lager - Digitalizzare e Studiare Le Testimonianze. In ME.TE. Digitali: Mediterraneo in Rete Tra Testi e Contesti. Proceedings of AIUCD 2024.

Hugging Face – The AI community building the future. (2025, gennaio 18). https://huggingface.co/

Mattingly, W. (2024). *Wjbmattingly/vulgata-spacy* [Python]. https://github.com/wjbmattingly/vulgata-spacy (Opera originale pubblicata 2022)

Microsoft/mpnet-base Hugging Face. (2025). Recuperato 20 gennaio 2025, da https://huggingface.co/microsoft/mpnet-base

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. https://arxiv.org/abs/1908.10084

Streamlit. (2025). Recuperato 20 gennaio 2025, da https://latin-vulgate.streamlit.app/

Streamlit: A faster way to build and share data apps. (2021, gennaio 14). https://streamlit.io/

Text Encoding Initiative. (2025). Recuperato 20 gennaio 2025, da https://tei-c.org/

The AI-native database developers love | Weaviate. (2025). Recuperato 20 gennaio 2025, da https://weaviate.io/

Il progetto RETI (REndering Texts and Images): metodologia e primi risultati[1]

Chiara Barbero[1], Matteo Di Franco[2], Federica Lazzerini[3], Annamaria Persia[4]
[1] Università del Molise, Italia - chiara.barbero@unimol.it
[2] Università di Napoli "Federico II", Italia - matteo.difranco@unina.it
[3] Università di Torino, Italia - federica.lazzerini@unito.it
[4] Università di Pisa, Italia - annamaria.persia@cfs.unipi.it

ABSTRACT (ITALIANO)

Il poster presenta i risultati cui è giunto il progetto PRIN 2022 RETI (Rendering Texts and Images, codice 2022HJW23K), tuttora in corso. Esso affronta le sfide poste dalla complessità di pubblicare testi antichi e medievali attraverso l'uso innovativo delle tecnologie digitali, allo scopo di creare, secondo un approccio fortemente interdisciplinare, l'edizione scientifica digitale di quattro opere: il *Lucullus* di Cicerone (45 a.C.), il *Panegirico a Cizico* di Elio Aristide (166 d.C.), un *corpus* di lettere diplomatiche tardomedievali (1461) e la *Cronaca Streghi* (c. 1470). Questi testi, diversi per la loro natura e le modalità in cui sono stati trasmessi, rappresentano casi studio ideali per mettere alla prova e sviluppare la terza versione di *Edition Visualization Technology* (EVT 3), un software open-source progettato per visualizzare edizioni scientifiche digitali da testi codificati in XML/TEI.
EVT supera i limiti delle edizioni tradizionali e di altri software, offrendo funzionalità avanzate come la gestione delle varianti testuali tramite apparati critici in stand-off, l'integrazione di immagini e annotazioni, e la creazione di indici semantici.
Parole chiave: edizioni digitali; EVT; interdisciplinarità; testi classici; testi medievali

ABSTRACT (ENGLISH)

Project RETI (REndering Texts and Images): Methodology and Early Results.
The poster aims to present the methodology and early results of project PRIN RETI 2022 (Rendering Texts and Images, code 2022HJW23K), currently ongoing. The project faces the challenges of publishing ancient and mediaeval texts through the innovative use of digital technologies. RETI aims to create, with a strongly interdisciplinary approach, the digital scientific edition of four works: Cicero's *Lucullus* (45 B.C.), Aelius Aristides' *Panegyricus to Cyzicus* (166 A.D.), a *corpus* of late-mediaeval diplomatic letters (1461), and the *Cronaca Streghi* (ca. 1470). These texts of different natures and transmission represent ideal case studies to test and develop the third version of *Edition Visualization Technology* (EVT 3), an open-source software used to visualize digital scientific editions from XML/TEI-encoded texts.
EVT overcomes the limitations of traditional editions and other softwares, offering advanced features such as supporting textual variants through critical stand-off apparatus, the integration of images and annotations, and the creation of semantic indexes.
Keywords: digital editions; EVT; interdisciplinarity; classical texts; mediaeval texts

1. INTRODUZIONE

I testi antichi e medievali sono stati trasmessi in un modo molto complesso che esige strumenti di analisi sempre più sofisticati per affrontare le sfide che essi pongono. Tali testi richiedono approcci interdisciplinari che spaziano dalla filologia alla paleografia, dall'ecdotica alla codicologia e agli studi storici. Le opportunità offerte dalle tecnologie digitali stimolano lo sviluppo di approcci e modelli innovativi, che facciano dialogare i testimoni tra loro o mettano in luce i rapporti tra apparati iconografici o supporti scrittori e testi.

Il progetto RETI (iniziato a ottobre 2022 e che si concluderà a ottobre 2025) riunisce quattro unità di ricerca (Università degli Studi di Pisa, Università degli Studi del Molise, Università degli Studi di Napoli "Federico II", Università degli Studi di Torino) che, congiuntamente, mirano a dare un contributo in questa direzione, testando e implementando una nuova versione del software *Edition Visualization Technology* (EVT)[2] per produrre edizioni scientifiche digitali di testi classici e medievali. Attraverso l'utilizzo integrato di

[1] Chiara Barbero è autrice dei §§ 3.3, 4; Matteo Di Franco dei §§ 2.1, 3.1; Federica Lazzerini dei §§ 2.2 e 5; Annamaria Persia dei §§ 1, 3.4, 6; il § 3.2 risulta dalla collaborazione di Matteo Di Franco e Federica Lazzerini. Il contributo nel suo complesso e il poster sono naturalmente l'esito di lavoro e riflessioni comuni.
[2] https://evt.labcd.unipi.it (cons. 07/04/2025). GitHub source code repository: https://github.com/evt-project/evt-viewer-angular/ (cons. 07/04/2025).

metodi tradizionali e strumenti digitali e con un approccio fortemente interdisciplinare, RETI affronta le complessità di testi di diversa natura e tradizione.

2. OBIETTIVI DEL PROGETTO

RETI ha iniziato il lavoro sui testi nella primavera 2024, perseguendo un obiettivo duplice (Gargiulo, 2023).

2.1. Il **primo obiettivo** è la creazione di edizioni scientifiche digitali di quattro testi classici e medievali:

- il *Lucullus* (45 a.C.), opera filosofica di Marco Tullio Cicerone sul dibattito epistemologico interno all'Accademia platonica nel I sec. a.C. (trasmesso da 74 manoscritti nella tradizione del cosiddetto *Corpus Leidense*), affidato a Federica Lazzerini (UniTo);
- il *Panegirico a Cizico* di Elio Aristide (166 d.C.; or. 27 Keil), un discorso pronunciato in occasione dell'inaugurazione del tempio di Zeus, ricostruito dopo il terremoto del 150/155 d.C. (il testo è trasmesso da 60 testimoni, i cui rapporti stemmatici sono indagati all'interno dello stesso progetto), ad opera di Matteo Di Franco (UniNa);
- un *corpus* di lettere diplomatiche tardomedievali scritte tra l'ottobre e il dicembre del 1461, oggi conservate nel fondo Archivio Gonzaga presso l'Archivio di Stato di Mantova, a cura di Chiara Barbero (UniMol);
- la *Cronaca Streghi* (Lucca, Biblioteca Statale, ms. 1661), una lunga opera storica in versi e dotata di un ricco e vivace corredo iconografico, che, a partire dalla leggenda della fondazione di Lucca, ripercorre le vicende più importanti della storia della città, sino al XV secolo, curata da Annamaria Persia (UniPi).

I testi, tra loro differenti dal punto di vista cronologico, linguistico e della storia della loro trasmissione, sono stati scelti in quanto l'edizione di ognuno di essi richiede differenti strumenti e metodi di ricerca. La loro selezione ha così permesso di testare la realizzazione di prodotti scientifici differenti.

I ricercatori di RETI stanno lavorando a diverse tipologie di edizioni in base alla natura dei testi. Per i due testi classici (il *Lucullus* e il *Panegirico a Cizico*) l'obiettivo è l'allestimento di edizioni critiche digitali. Nel caso del primo, il lavoro di edizione parte da una collazione già quasi completa (comunque sottoposta a revisione) e si basa sugli studi dello *stemma codicum* ad opera di Ermanno Malaspina e del suo gruppo di ricerca (Malaspina, 2018, 2020; Malaspina et al., 2014). Per il secondo, la ricostruzione del testo avviene seguendo le modalità tradizionali della filologia classica, collazionando i manoscritti *de visu*, tramite le immagini pubblicate on line liberamente dalle biblioteche di conservazione (come la Biblioteca Medicea Laurenziana, la Biblioteca Apostolica Vaticana o la Bibliothèque Nationale de France), o richiedendo le riproduzioni per fini di studio. Le due edizioni critiche digitali – le prime per i due testi – non prevedono la pubblicazione delle immagini ad accompagnamento dell'apparato critico.
Diversamente, le edizioni degli altri due casi di studio intendono offrire all'utente la visualizzazione delle immagini delle fonti originali: queste sono state ottenute mediante riproduzioni ad opera degli stessi istituti che conservano i documenti, fornite al gruppo di ricerca specificamente per questo progetto. Per il *corpus* delle lettere diplomatiche si vuole realizzare un'edizione con trascrizione sia diplomatica sia interpretativa dei testi, che offra allo stesso tempo strumenti di ricerca e possibilità di approfondimento storico. Per la *Cronaca Streghi*, si mira a un'edizione con trascrizione interpretativa, valorizzando la relazione tra testo e immagini.

2.2. Il **secondo obiettivo** del progetto è testare, sviluppare e migliorare la nuova versione del software EVT.

EVT (*Edition Visualization Technology*),[3] sviluppato da Roberto Rosselli Del Turco con la collaborazione e il supporto dell'Università di Pisa e del del centro interdipartimentale di ricerca DISH (*Digital Scholarship for the Humanities*)[4] dell'Università di Torino, è uno strumento agile e open-source progettato per visualizzare edizioni digitali da testi codificati in XML. L'applicazione è scritta in linguaggio TypeScript e basata sul framework Angular (Buzzoni & al. 2024). La sua interfaccia intuitiva consente agli utenti di navigare,

[3] http://evt.labcd.unipi.it (cons. 07/04/2025).

[4] https://www.dish.unito.it/it/il-centro (cons. 07/04/2025).

esplorare e studiare le edizioni digitali; inoltre, EVT è completamente open-source, web-based e compatibile con diversi sistemi operativi.
Il software, che può essere scaricato da un repository pubblico reperibile direttamente sul sito, permette la presentazione sincrona e interattiva di materiali complessi, come le immagini di un manoscritto o di un documento, la sua trascrizione diplomatica e interpretativa e la visualizzazione di note critiche (Rosselli Del Turco, 2016, 2019a, 2019b).
Attualmente il software è disponibile in due versioni (EVT 1 e EVT 2). Il progetto RETI collabora allo sviluppo della nuova versione (EVT 3) per l'introduzione di nuove funzionalità, tra cui:

- l'interazione con un apparato critico in stand-off, modellato sul DEPA, che dia conto in modo più ampio e accurato delle varianti testuali;
- la gestione avanzata di immagini e annotazioni testuali;
- l'integrazione di metadati e funzionalità semantiche.

Nel poster si presenterà anche lo stato dei lavori di questa nuova versione, in fase di sperimentazione. Gli aspetti a cui daremo rilievo si intersecano con quelli oggetto di altre presentazioni in questa stessa sede.[5]

3. METODOLOGIA

Il progetto adotta un approccio interdisciplinare, che combina analisi filologica, paleografica, storica e codicologica per lo studio delle fonti e delle tradizioni manoscritte con tecniche di markup digitale.

3.1. Nel caso di tradizioni in cui la distanza temporale tra la "pubblicazione" del testo da parte dell'autore e i primi testimoni che lo trasmettono non sia molto ampia, si rendono apprezzabili e ricercabili le caratteristiche materiali e codicologiche del testo, quali illustrazioni, legature e mani di copisti e illustratori; nel caso di testi antichi greco-latini, trasmessi da numerosi testimoni molto distanti cronologicamente dall'autore e redatti in più secoli, si testa la capacità di EVT 3 di gestire l'accumulo delle varianti e produrre un apparato critico in stand-off. Le edizioni digitali realizzate con EVT 3 sono supportate da un'analisi approfondita dei manoscritti, ivi compresi aspetti stemmatici, codicologici, paleografici, filologici e storico-artistici. Particolare attenzione sarà dedicata al rapporto testo-immagine e alla verifica dell'efficacia del software nel rispettare le specificità di ogni manoscritto.
Il modello adottato per la codifica dei testi segue le norme della TEI (*Text Encoding Initiative*)[6] poiché essa è considerata lo standard internazionale per la codifica dei testi (Ciotti, 2023). Le soluzioni ammesse dalle linee guida del TEI Consortium (P5 Version 4.9.0) sono selezionate e ampliate in funzione della correttezza formale della codifica stessa, dell'aderenza alle caratteristiche codicologiche e testuali dei singoli documenti, e della resa finale dei dati su EVT 3 in funzione di fruibilità e ricercabilità quanto più agevoli possibile.
Le edizioni realizzate nell'ambito del progetto RETI ottengono il notevole risultato di presentare testi corredati di un gran numero di informazioni di vario tipo (storia del testo, informazioni sull'autore, dati codicologici, notizie sui manoscritti, dati paleografici, analisi iconografiche, dati storici relativi a persone e luoghi citati), di cui gli utenti possono fruire secondo le proprie necessità. EVT consente anche una fruizione personalizzata da parte dell'utente, con la possibilità di creare indici semantici.

3.2. In particolare, per quanto riguarda i testi antichi dei quali si stanno realizzando le edizioni critiche (il *Lucullus* e il *Panegirico a Cizico*), il nostro lavoro entra nel merito del dibattito, ormai di lunga data, sulla fattibilità, l'efficacia e il valore scientifico delle edizioni critiche digitali (Stella 2015, Rosselli Del Turco, 2016, Monella, 2018). Fra gli esperti del settore, accanto a posizioni scettiche, si annoverano visioni più ottimistiche, che riconoscono che un'edizione critica digitale possa essere tanto valida e pregevole quanto una tradizionale edizione a stampa. Poche voci hanno sostenuto che la solidità scientifica di un'edizione critica digitale possa essere addirittura maggiore (Malaspina, 2019, Lazzerini, in corso di stampa); il nostro progetto offre risultati concreti che dimostrano che, con gli strumenti giusti – quali le nuove funzionalità implementate in EVT 3 – si possono ottenere dei vantaggi cospicui proprio dal punto di vista filologico. Mentre nelle edizioni critiche a stampa i vincoli intrinseci di impaginazione rendono necessario operare una netta selezione delle varianti da riportare nell'apparato critico, lo spazio di una pagina digitale è plasmabile

[5] "Embracing flexibility: new EVT features for critical editing, accessibility and inclusivity" (Roberto Rosselli Del Turco, Marina, Davide Cucurnia); "A Case Study of Linked Data and Open Science in Cultural Heritage: A Method Linking Three Open Data Tools – Digital Philology for Dummies (DPhD), Edition Visualization Technology (EVT), and a Relational Database" (Renato Caenaro, Chantal Pivetta).
[6] https://tei-c.org (cons. 07/04/2025).

a uso e preferenza dell'utente: i ricercatori che lavorano sui testi antichi greco-latini hanno dunque la possibilità di riportare tutte le varianti testuali tramandate per una certa lezione. Pertanto, edizioni così realizzate risultano a tutti gli effetti più complete rispetto a un'edizione critica tradizionale. Allo stesso tempo, l'editore del testo non rinuncia ad applicare i principi dell'ecdotica, valutando e classificando (con opportuna codifica XML) le varianti sulla base della loro posizione all'interno dei rami di trasmissione (`@type`) e anche sulla base della probabile origine dell'errore (`@cause`). Dunque, l'utente di un'edizione critica digitale così costituita ha a disposizione a tutti gli effetti più dati di quelli che avrebbe in un'edizione critica a stampa, seppur interpretati e valutati criticamente, e senza che ciò comprometta la leggibilità del testo.
L'adozione di un apparato critico in stand-off consente di gestire l'accumulo di dati di un apparato completo, collegando singole parole o più ampie stringhe testuali alle relative note di apparato: in questo modo si evita l'inclusione diretta (*in-line*) delle varianti nel testo, come avviene con i metodi *Location-referenced* e *Parallel Segmentation*. Il progetto RETI è il primo a implementare una variante del metodo *Double-End-Point Attachment* (DEPA), che permette la sovrapposizione (*overlap*) di varianti all'interno della stessa stringa, una funzionalità non supportata dagli altri due metodi. L'associazione visiva tra testo e note è gestita da un *parser* specificamente sviluppato per EVT 3 dalla software house SilentWave.[7]
L'introduzione del metodo DEPA in EVT 3, un'iniziativa innovativa nello sviluppo del software, ha richiesto la creazione e la costante ottimizzazione di un modello di codifica XML. Questo processo di adattamento è fondamentale per garantire una fruizione ottimale e di alta qualità delle edizioni, massimizzando l'efficacia dell'analisi e la correttezza delle informazioni presentate. In base a questo modello, a ogni parola, contenuta in un elemento `<w>`, viene assegnato un `@xml:id` univoco, richiamato successivamente negli elementi `<app>` tramite gli attributi `@from` e, se sono coinvolte più parole, `@to`.
Un esempio dal *Lucullus*:

```
<body>
...
  <w xml:id="Luc-002-21">Murena</w>
  <w xml:id="Luc-002-22">bellum</w>
  <w xml:id="Luc-002-23">in</w>
  <w xml:id="Luc-002-24">Ponto</w>
  <w xml:id="Luc-002-25">gerente</w>
...
</body>
```

L'apparato critico è codificato separatamente, all'interno della sezione `<back>` del documento TEI:

```
<back>
...
<app from="#Luc-002-22" to="#Luc-002-24">
  <rdg type="family-distinctive" cause="transposition" wit="#μ #Chis1">in ponto
bellum</rdg>
</app>

<app from="#Luc-002-24">
  <rdg type="polygenetic" cause="contextInduced" wit="#Bes #Cas #Erl #Pal
#Pan">ponte</rdg>
  <rdg type="lectio.sing" cause="palaeographicConfusion" wit="#Glas">ponco</rdg>
</app>

<app from="#Luc-002-23" to="#Luc-002-24">
  <rdg type="lectio.sing" cause="palaeographicConfusion" wit="#Parm">imposito</rdg>
</app>
...
</back>
```

[7] https://www.silentwave.eu/en/consulting-evt (cons. 07/04/2025).

In questo modo, è possibile creare riferimenti alla stessa parola in più note `<app>`, gestendo efficacemente le varianti testuali, anche in fase di collazione o di *constitutio textus*.

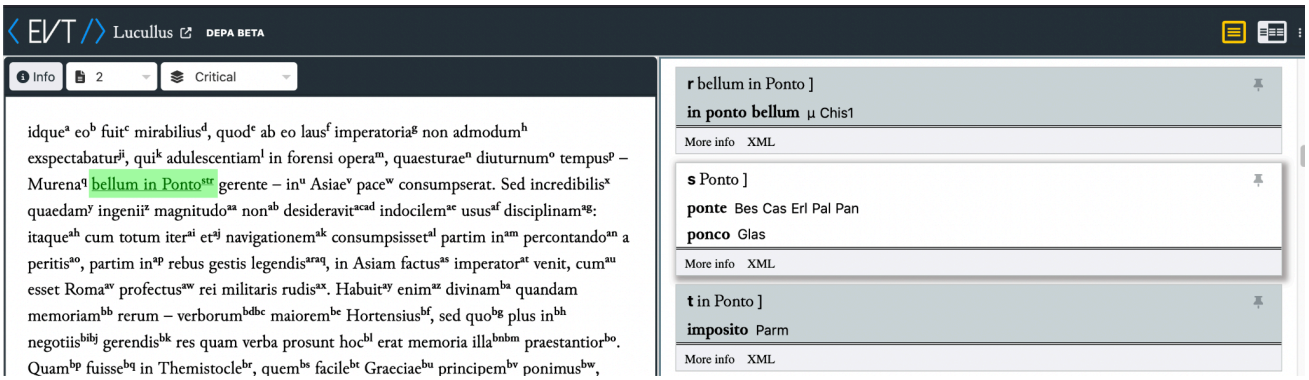


Figura 1. Visualizzazione del testo con apparato critico nella versione del software in corso di sviluppo

3.3. Per quanto riguarda il *corpus* di lettere tardomedievali, si vuole realizzare un'edizione digitale in grado di mettere a disposizione dell'utente maggiori strumenti di ricerca rispetto alle tradizionali edizioni cartacee e di fornire allo stesso tempo un collegamento diretto e immediato tra testo e immagine. Un importante esempio in tal senso, che utilizza il software EVT, è già offerto dall'edizione digitale delle lettere del compositore Vincenzo Bellini (1801-1835) curata da Angelo Mario Del Grosso e Daria Spampinato (Del Grosso & Spampinato, 2023). Tuttavia, il nostro caso si differenzia molto per la natura della documentazione selezionata: si tratta infatti di un *corpus* epistolare non solo più antico, ma anche costituito da diverse forme di redazione e trasmissione delle lettere. Il *corpus* di fonti scelto è costituito da un segmento della corrispondenza, relativo al periodo che va dall'ottobre al dicembre del 1461, dei marchesi di Mantova Ludovico Gonzaga e Barbara di Brandeburgo e del loro oratore a Roma Bartolomeo Bonatti, incaricato di portare a termine i difficili negoziati che videro l'elezione a cardinale del figlio dei marchesi, Francesco Gonzaga. Alle lettere scambiate tra questi tre principali interlocutori si è scelto di aggiungerne altre che coinvolgono terzi nella medesima questione, per un totale di circa 20 missive, conservate nel fondo Archivio Gonzaga presso l'Archivio di Stato di Mantova. Il campione selezionato, per quanto numericamente limitato, offre un caso di studio interessante e vario, che comprende testi latini e volgari, così come diverse forme documentarie: lettere in originale, minute e trascrizioni di lettere nei registri di copialettere. Di queste lettere si presentano le trascrizioni sia diplomatiche sia interpretative, affiancate dalla scansione del documento originale e arricchite da funzionalità importanti. La codifica di ciascuna missiva si propone, infatti, di utilizzare gli elementi messi a disposizione dallo standard XML/TEI per il markup, e per la conseguente individuazione, dei personaggi e dei luoghi citati, oltre che per la descrizione della corrispondenza e della struttura testuale delle lettere. Tale marcatura si traduce, grazie a EVT 3, in un'edizione digitale con un'interfaccia agile e intuitiva, in grado di mettere a disposizione del lettore testi facilmente leggibili e di fornire al tempo stesso possibilità di approfondimento ed elementi di ricerca utili.

3.4. Il progetto relativo alla *Cronaca Streghi* parte dallo studio del ms. BSL 1661, datato all'ultimo quarto del XV secolo e noto per il suo ricchissimo corredo iconografico. La scelta di questo particolare testimone per l'allestimento di un'edizione digitale di tipo interpretativo e per l'approfondimento e la valorizzazione del rapporto testo-immagine si giustifica con la sua complessità: si tratta, infatti, di un codice del tutto anonimo, frutto di una collaborazione così irregolare tra copisti e illustratori da rendere di difficile comprensione le dinamiche di copia e dell'allestimento del corredo iconografico. Le potenzialità di EVT trovano, infatti, uno stimolante banco di prova nell'ideazione di soluzioni visive atte alle peculiarità che contraddistinguono il codice. Il proposito è quello di realizzare una visualizzazione in simultanea di riferimenti topografici, onomastici ed evenemenziali, sfruttando le potenzialità della marcatura XML/TEI per rendere interattiva la consultazione del codice, superando così i limiti che anche il più dettagliato degli indici di un'edizione cartacea inevitabilmente comporta.

Questo tipo di edizione rientra nel filone di progetti curati precedentemente per versioni anteriori di EVT, come l'edizione del *Dream of the Rood*,[8] quella del Codice Pelavicino[9] e i lavori del progetto "Manoscritti preziosi del Piemonte".[10] Tuttavia, il progetto propone l'adeguamento e la personalizzazione del software sulla base di esigenze specifiche che possano rispondere *ad hoc* a particolari requisiti codicologici, paleografici, testuali e visivi, in particolare il raffronto con documenti iconografici esterni, quali, ad esempio, le illustrazioni dell'altro testimone illustrato dell'opera, il ms. BSL 2629 tramite *pop up*.

4. FUTURI SVILUPPI

La ricerca del PRIN RETI ha permesso anche di individuare obiettivi futuri su cui lavorare per accrescere l'efficacia e le potenzialità di questo software per testi di diversa natura. In particolare, sarà opportuno concentrare risorse verso lo sviluppo e miglioramento di nuove funzionalità, quali:

(a) Per le due edizioni critiche, la possibilità di usare gli attributi `@type` e `@cause` (che classificano le varianti in base alla loro rilevanza stemmatica e alla probabile origine dell'errore) come filtri di visualizzazione e ricerca; la gestione di lacune intenzionali nei manoscritti, diversa dalle omissioni erronee; la gestione dei testimoni acefali o mutili.

(b) Per l'edizione del *corpus* di lettere, la possibilità di ricercare e selezionare le diverse missive in base, ad esempio, al loro mittente.

(c) Per l'edizione della *Cronaca Streghi*, la possibilità di eseguire raffronti con altre testimonianze illustrative, tramite focus su dettagli dell'immagine, e meglio collegare la relativa descrizione didascalica.

5. INCLUSIVITÀ

In virtù della sua reperibilità, facilità di configurazione e delle sue minime esigenze di *maintenance*, EVT è accessibile a un'ampia platea di studiosi, indipendentemente dalle loro competenze in ambito informatico e senza richiedere l'acquisto di una licenza. Pertanto, esso è uno strumento che contribuisce all'abbattimento delle barriere poste dal diverso livello di competenze digitali e possibilità economiche che spesso separano potenziali utenti dal sapere, in ottemperanza ai principi FAIR (*Findability, Accessibility, Interoperability, Reusability*).[11] Inoltre, l'uso di EVT da parte del progetto RETI per produrre edizioni in open access permette tanto a studiosi quanto a utenti non specialisti di avere accesso a prodotti di solido valore scientifico che, tipicamente, richiedono un importante investimento economico o la possibilità di fruire di biblioteche dotate della strumentazione adeguata.
Il progetto RETI è improntato alla trasparenza e renderà pubblica e accessibile a tutti la metodologia seguita, in modo che ogni aspetto del nostro lavoro possa essere sottoposto alla considerazione e al vaglio della comunità scientifica.

6. RISULTATI ATTESI

Il progetto RETI avrà un impatto su più fronti.
Sul fronte filologico e storico, il progetto realizzerà edizioni digitali interattive e agilmente fruibili, capaci di evidenziare le peculiarità dei testi e immagini e di mettere a immediata disposizione dell'utente un gran numero di informazioni aggiuntive sul loro contenuto e contesto. Saranno inoltre disponibili per la ricerca dati testuali più ricchi e completi (meno soggetti alla necessaria sintesi delle edizioni a stampa).
Sul fronte tecnologico, il miglioramento di EVT rappresenterà un progresso significativo per le edizioni digitali, rendendo lo strumento più flessibile e accessibile a utenti con differenti gradi di specializzazione. Le edizioni realizzate e disponibili in open access sono di alta qualità, consentendo una visualizzazione integrata di testo, immagini e apparati critici e la possibilità di creare indici semantici per una fruizione interattiva.

[8] http://evt.labcd.unipi.it/demo/evt_v1-3/dotr (cons. 07/04/2025).
[9] https://pelavicino.labcd.unipi.it/evt (cons. 07/04/2025).
[10] https://bnuto.cultura.gov.it/biblioteca-digitale/manoscritti-in-evt (cons. 07/04/2025).
[11] https://www.go-fair.org/fair-principles (cons. 07/04/2025).

BIBLIOGRAFIA

Battaglia Ricci, L. (1994). *Parole e immagini nella letteratura italiana medievale: materiali e problemi*. Pisa: GEI.

Buzzoni, M., Cucurnia, D. Fenu, C., Rosselli Del Turco, R., & Tancredi, G. (2024). Progetto di edizione genetica digitale del Canzoniere manoscritto di U. Saba (1919-20). In A. Di Silvestro, D. Spampinato (a cura di), *Me.Te. Digitali. Mediterraneo in rete tra testi e contesti, Proceedings del XIII Convegno Annuale AIUCD,* Catania 28-30 maggio 2024. Catania: AIUCD. *Quaderni di Umanistica Digitale* (pp. 215–220). DOI: 10.6092/unibo/amsacta/7927.

Ciotti, F. (2023). *Digital Humanities. Metodi, strumenti, saperi*. Roma: Carocci.

Ciotti, F. (2023). La codifica del testo, XML e la TEI. In F. Ciotti (a cura di), *Digital Humanities. Metodi, strumenti, saperi* (pp. 66–90).

Del Grosso, A. M., & Spampinato, D. (a cura di) (2023), *Bellini Digital Correspondence*, Roma: CNR Edizioni, https://bellinicorrespondence.cnr.it [consultato il 07/04/2025].

Driscoll, M. J., & Pierazzo, E. (a cura di) (2016). Digital Scholarly Editing. Theories and Practices. *Digital Humanities*, *4*.

EVT – Edition Visualization Technology. (2014-). Home page: https://evt.labcd.unipi.it. GitHub source code repository: https://github.com/evt-project/evt-viewer-angular/.

Gargiulo, G. G. (2023). Tradizione e innovazione. Considerazioni in margine al Convegno inaugurale PRIN RETI (REndering Text and Images). *Quaderni dell'Archivio Storico - Fondazione banco di Napoli*, *N.S. 8*(1), 7–32.

Lazzerini, F. (in corso di stampa). Il valore aggiunto di un'edizione critica digitale di Cicerone filosofo. In F. Bellorio, V. Revello (a cura di), *'Nec ut interpres sed ut… philosophus': Pourquoi éditer et commenter encore les textes philosophiques cicéroniens? Actes du colloque du 4-6 novembre 2024,* Paris-Sorbonne, Cicero. Studies on Roman Thought and its Reception, Vol. 12. Berlin-New York: De Gruyter.

Malaspina, E. (2018). *Recentior non deterior*: Escorial R.I.2 e una nuova *recensio* del *Lucullus* di Cicerone, *Paideia 73*, 1969–85.

Malaspina, E. (2019). Il futuro dell'edizione critica (cioè lachmanniana), più o meno digitale. Riflessioni (in)attuali. In A. Chegai, M. Rossellini, E. Spangenberg (a cura di), Textual Philology Facing "Liquid Modernity": Identifying Objects, Evaluating Methods, Exploiting Media, *Storie e Linguaggi*, *5*, 35–60. Padova: libreriauniversitaria.it.

Malaspina, E. (2020). Lupo e "Adoardo" nel *Lucullus* di Cicerone: congetture carolinge e tradizioni perdute nel *Corpus Leidense*?, *Rationes rerum 16*, 251–88.

Malaspina, E., Borgna, A., Lucciano, M., Caso, D., & Senore, C. (2014). I manoscritti del *Lucullus* di Cicerone in Vaticana: valore filologico e collocazione stemmatica, *Miscellanea Bibliothecae Apostolicae Vaticanae*, *20*: 589–620.

Milanese, G. (2020). *Filologia, letteratura, computer. Idee e strumenti per l'informatica umanistica*. Milano: Vita e Pensiero.

Monella, P. (2018). Why Are There No Comprehensively Digital Scholarly Editions of Classical Texts? In A. Cipolla (a cura di), *Digital Philology: New Thoughts on Old Questions* (pp. 141–59). Padova: libreriauniversitaria.it.

Rosselli Del Turco, R. (2016). The Battle We Forgot to Fight: Should We Make a Case for Digital Editions? In M. J. Driscoll & E. Pierazzo (a cura di), *Digital Scholarly Editing. Theories and Practices* (Vol. 4, pp. 219–238). DOI: 10.11647/OBP.0095.12.

Rosselli Del Turco, R. (2019a). Designing an advanced software tool for Digital Scholarly Editions: The inception and development of EVT (Edition Visualization Technology). *Textual Cultures*, *12*(2), 91–111. DOI: 10.14434/textual.v12i2.27690.

Rosselli Del Turco, R. (2019b). La visualizzazione di edizioni digitali con EVT: una soluzione per edizioni diplomatiche e critiche. *Ecdotica*, *16*, 148–173.

Stella, F. (2015). *Il problema della codifica nelle edizioni filologiche digitali*. In L. Del Corso, F. De Vivo, A. Stramaglia (a cura di), Nel segno del testo. Edizioni, materiali e studi per O. Pecere, *Papyrologica Florentina*, *44* (pp. 347–57).

TEI Consortium (Eds.). (2025). TEI: Guidelines for Electronic Text Encoding and Interchange. 4.9.0. 24 January 2025. http://www.tei-c.org.

Edizione digitale ed autorialità plurima: quali sfide?

Stefania Tesser
Université Grenoble Alpes, Francia; Università Ca' Foscari Venezia, Italia – stefania@unive.it

ABSTRACT (ITALIANO)
Il numero crescente di studi dedicati al pensiero politico di Germaine de Staël è un fenomeno recente che ha comportato la riscoperta della sua più importate opera politica: le *Considérations sur les principaux événements de la Révolution française*. Progettare un'edizione digitale di quest'opera postuma presenta tuttavia diverse sfide perché l'ultimo manoscritto lasciato dall'autrice mostra un lavoro a più mani effettuato sotto la sua supervisione e anche dopo la sua morte. Quale testo o quali testi proporre? Quale sistema di codifica scegliere per le diverse tipologie di interventi? Il progetto è appena ai suoi inizi ma offre alcuni spunti di riflessione sulla scelta di un tipo di edizione che tenga conto del modo di lavorare di Germaine de Staël, modo che ci restituisce una nozione complessa di autorialità.
Parole chiave: edizione scientifica digitale; autorialità plurima; Germaine de Staël; codifica XML

ABSTRACT (ENGLISH)
Digital edition and multiple authorship: what challenges?
The increasing number of studies on the political thought of Germaine de Staël (1766-1817) is a recent phenomenon that has enabled us to rediscover her most important political work: the *Considérations sur les principaux événements de la Révolution française*. However, designing a digital edition of this posthumous work has to face several challenges because the last manuscript left by the author shows a multi-handed work carried out under her supervision and also after her death. What kind of text or texts should we propose? What kind of encoding system should we choose for the different types of interventions? The project has just started, but it provides some insights about the choice of an edition that takes into account Germaine de Staël's way of working, which shows a complex notion of authorship.
Keywords: digital scholarly editing; multiple authorship; Germaine de Staël; XML-TEI markup

1. LE *CONSIDÉRATIONS*: UN LAVORO A PIÙ MANI?

Germaine de Staël (1766-1817), conosciuta principalmente per la sua opera critica *De l'Allemagne* (Staël-Holstein, 2017b), ha scritto anche opere politiche rimaste in parte incompiute come le *Considérations sur les principaux événements de la Révolution française* (Staël-Holstein, 2017a). Nelle intenzioni originarie dell'autrice le *Considérations* dovevano descrivere la vita pubblica del padre Jacques Necker, ministro delle Finanze sotto Luigi XVI, ma finirono per narrare tutto il periodo rivoluzionario e quello successivo fino alla Restaurazione. L'interpretazione che viene data dei fatti rivoluzionari inaugurerà una nuova storiografia della Rivoluzione francese ed avrà un'influenza decisiva su alcune correnti liberali della prima metà del diciannovesimo secolo in Francia.
Germaine de Staël non riuscirà a pubblicare l'opera: morirà nel 1817 lasciando un lavoro molto avanzato ma non terminato di suo pugno (Usteri & Richter, 1903, p. 243). Come da testamento, le *Considérations* saranno riviste e pubblicate postume nel 1818 dal figlio Auguste, Victor de Broglie e August Wilhelm von Schlegel (Staël-Holstein, 1818). Auguste non nasconde l'imbarazzo di un tale incarico: i primi due volumi ed una parte del terzo erano pronti per la pubblicazione mentre la rimanente parte del terzo presentava capitoli non rivisti oppure allo stadio di una prima redazione autografa. Se l'autrice fosse rimasta in vita, queste parti avrebbero potuto probabilmente subire delle modifiche significative. Bisognava quindi decidere se intervenire per terminare il lavoro oppure lasciare tutto com'era (King, 1970, pp. 47–48).
Come per la maggior parte delle opere di Germaine de Staël, non è stato possibile per lungo tempo confrontare la versione pubblicata con i manoscritti lasciati dall'autrice perché la maggior parte di essi è conservata in archivi privati non accessibili. Nel 1963 la Biblioteca nazionale di Francia è riuscita tuttavia ad acquisire un manoscritto[1] eseguito da più copisti e che presenta correzioni della Staël. L'analisi di questo testimone e la storia della sua trasmissione hanno permesso di ipotizzare che si tratti probabilmente dell'ultimo manoscritto lasciato dall'autrice (Broglie, 1886, p. 16). La copia presenta anche importanti interventi da parte di Auguste de Staël che attenuano la vivacità polemica della madre su alcune questioni di attualità e che si configurano come vere e proprie censure. Anche la collazione tra manoscritto e prima edizione ha messo in evidenza notevoli differenze: capitoli interamente riscritti oppure lunghe aggiunte che inducono ad interrogarsi sulla natura di quegli interventi e su chi li avesse realmente

[1] Biblioteca nazionale di Francia, NAF 14606-14608.

effettuati. Tutte queste considerazioni hanno portato alla prima edizione critica che ha restituito il testo del manoscritto con le ultime revisioni della Staël (Staël-Holstein, 2017a). Tuttavia, quando si è cercato di individuare gli interventi di Auguste per escluderli dal testo finale, ne sono emersi degli altri di mano diversa. Alcuni di questi vengono a volte corretti dalla stessa autrice a dimostrazione del fatto che sono stati fatti mentre lei era ancora in vita. Il manoscritto rivela così un lavoro a più mani la cui peculiarità dovrebbe essere presa in considerazione per comprendere la genesi dell'opera. Nell'edizione critica su carta non è stato possibile mostrare il complesso intreccio di interventi, cosa che ha portato ad una riflessione sull'opportunità di altri strumenti editoriali in grado di facilitarne l'analisi.
La Fig. 1 mostra un intervento allografo precedente alle varie fasi correttorie della Staël (D):

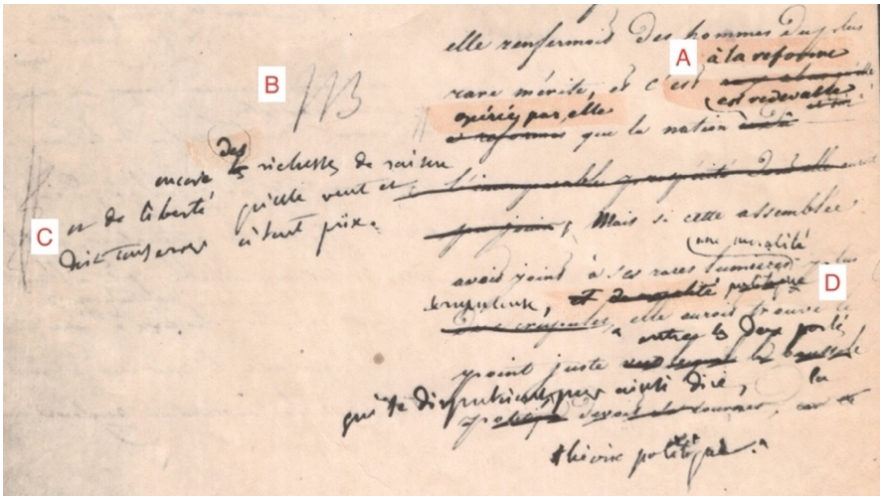


Figura 1. Interventi di diverse mani in uno stesso paragrafo: August Wilhelm Schlegel nell'interlinea (A); segno a matita non identificato (B); sostituzione di Germaine de Staël (C); aggiunta in interlinea di mano non identificata "et de moralité" dopo "scrupules" integrata dalla Staël con l'aggettivo "politique" (D); tutta la frase è poi sostituita dall'autrice con "une moralité plus scrupuleuse" in interlinea (Biblioteca nazionale di Francia, NAF 14606, f. 175v)

Quali scelte ecdotiche effettuare nei casi in cui questo tipo di interventi non risultano barrati dall'autrice e potrebbero quindi essere stati accettati da lei? Si può fare un'analoga considerazione per la sostituzione in interlinea di August Wilhelm Schlegel (A): i suoi interventi sono, anche se raramente, corretti dalla Staël in altri punti del testo. Del resto, la Staël era solita ricorrere a lui per rivedere le sue opere e questa pratica era molto comune all'interno del Gruppo di Coppet[2] i cui membri si scambiavano regolarmente i manoscritti per ricopiarli o correggerli (Genand & Rosset, 2022, p. 10; Saintes, 2022, p. 109). Schlegel era particolarmente attivo in questo scambio e quindi, quando i suoi interventi non sono stati barrati nel manoscritto, potrebbero essere stati accettati dall'autrice.
Anche gli interventi postumi sollevano questioni complesse dal punto di vista dell'autorialità. Quelli di Auguste nel manoscritto mirano, per esempio, ad attenuare le critiche della Staël a personaggi ancora in vita, alla dinastia dei Borboni appena insediata oppure alla politica estera dell'Inghilterra. Qualunque sia la motivazione alla base di queste modifiche, il lavoro di revisione di Auguste è fatto per eseguire una volontà testamentaria che gli chiedeva di terminare le *Considérations*. In un articolo intitolato "You Complete Me: Posthumous Works and Secondary Agency", Bacharach e Tollefsen (2015) affermano che il lavoro postumo di completamento di un'opera è attribuibile ancora all'autore se quest'ultimo ha autorizzato una determinata persona a farlo perché si tratterebbe di un lavoro fatto per suo conto. La questione dell'autorialità è affrontata da un punto di vista ontologico e non viene esaminata la natura degli interventi. L'articolo invita tuttavia a riflettere sulle molteplici sfaccettature del problema: quanto è plausibile la scelta di eliminare gli interventi di Auguste dal testo finale? È ancora possibile proporre una versione del testo fedele all'ultima "volontà dell'autore"?

[2] La denominazione "Gruppo di Coppet" è stata utilizzata per la prima volta da Carlo Pellegrini per indicare un gruppo di intellettuali europei che si riuniva nella residenza della Staël a Coppet (Svizzera). Era frequentato da vari esponenti della cultura e della politica in Europa. I membri più assidui, oltre alla Staël, erano Benjamin Constant, August Wilhelm Schlegel, Jean de Sismondi et Charles Victor de Bonstetten (Balayé, 1977; Pellegrini, 1974).

Far emergere il lavoro collaborativo come parte integrante del processo di scrittura contribuisce alla comprensione della genesi delle *Considérations* fornendo informazioni sul modo in cui interagiscono tutti i soggetti coinvolti nelle varie fasi di elaborazione dell'opera.
Il lavoro di revisione postumo non si è limitato all'ultimo manoscritto lasciato dall'autrice, ma ha comportato ulteriori cambiamenti visibili solo confrontandolo con la prima edizione. Le modifiche sono a volte di una tale ampiezza da arrivare a veri e propri rifacimenti effettuati probabilmente su un'altra copia. Si tratta di una copia postuma oppure di una copia almeno in parte rivista dall'autrice? Allo stato attuale delle ricerche non è possibile rispondere ma dare l'opportunità di mettere agevolmente a confronto la versione del manoscritto con quella pubblicata permetterebbe di esaminare la natura di questo lavoro. Pensare ad una edizione cartacea di entrambi i testi avrebbe significato fare i conti con i limiti di questo supporto come la sua bidimensionalità ed i costi di pubblicazione. Queste constatazioni hanno portato ad esplorare le potenzialità degli strumenti digitali per creare uno prodotto editoriale più flessibile.

2. UN'EDIZIONE DIGITALE PER L'AUTORIALITÀ PLURIMA

L'edizione digitale delle *Considérations* è un progetto ancora in una fase iniziale di sperimentazione e riflessione che ha come primo obiettivo quello di integrare nella codifica e nella visualizzazione tutti gli interventi presenti nel manoscritto trattandoli come parte del processo di scrittura. Il tipo di marcatura scelta permetterà comunque di distinguere le differenti mani, ma la natura incompiuta dell'opera ed il lavoro collaborativo che la caratterizza non ci permettono di sapere quali interventi allografi Germaine de Staël avrebbe sacrificato o accettato se avesse avuto il tempo di pubblicarla. È preferibile allora dare la possibilità di analizzare più versioni a seconda dei vari interventi sfruttando le potenzialità euristiche di questa scelta. L'edizione digitale può dare spazio ad una tale pluralità testuale (Pierazzo, 2018) basandosi su un sistema di codifica che tiene conto delle peculiarità dell'oggetto studiato.
Il secondo obiettivo è facilitare un confronto tra manoscritto e prima edizione mediante, per esempio, una presentazione sinottica che possa evidenziarne le divergenze senza presentare una versione come più autorevole rispetto all'altra. L'obiettivo è di ridurre l'interpretazione dell'editore scientifico nella ricostruzione del testo secondo "l'ultima volontà dell'autrice", ricostruzione che si è rivelata problematica. L'edizione digitale diventa in questo modo utile per la sua complementarità rispetto all'edizione critica del 2017: fornisce ulteriori strumenti per verificare il lavoro filologico precedentemente svolto aprendo la strada ad una nozione fluida di testo che incoraggia diverse ipotesi ermeneutiche.
Inoltre, l'edizione digitale mette a disposizione tutta una serie di dati utili all'analisi di questi e altri aspetti perché si basa su un sistema di codifica che consente il loro trattamento automatico. Ne sono un esempio i dati sulla materialità de manoscritto, aspetto ancora poco esplorato in Germaine de Staël (Balayé, 1994, p. 105) ma utile per comprendere la genesi delle *Considérations*. Anche i dati sui vari interventi si prestano ad analisi quantitative con analoghe finalità.
Infine, per consentire un ampio utilizzo dei dati messi a disposizione, il progetto dovrà garantire l'accesso aperto alla risorsa e delle scelte tecniche a favore dell'interoperabilità e della conservazione a lungo termine.

3. METODOLOGIA

L'edizione digitale delle *Considérations* prevede l'implementazione di una pagina web a partire da un prototipo realizzato su GitHub Pages[3] che dà accesso al testo del manoscritto,[4] di cui è già disponibile una trascrizione, a quello della prima edizione[5] ed ai rispettivi facsimili. Il prototipo prevede anche una sezione della pagina web con la documentazione sul progetto ed una sull'autrice e l'opera. In quest'ultima sezione verranno aggiunti una bibliografia e degli studi che potranno eventualmente basarsi sui dati messi a disposizione dall'edizione. Il testo del manoscritto verrà codificato in XML-TEI, scelta ormai d'obbligo ai fini dell'interoperabilità. Il modello di codifica è in fase di elaborazione ed è partito dall'individuazione di quattro categorie di interventi: 1) le autocorrezioni dei copisti; 2) un primo lavoro di revisione e correzione degli errori dei copisti effettuato da mano non identificata 3) le correzioni dell'autrice; 4) i suggerimenti di modifica a matita di mano non identificata; 5) gli interventi di Schlegel; 6) gli interventi di Auguste de Staël; 7) le modifiche che risultano dalla collazione tra manoscritto e prima edizione. L'identificazione delle mani e la cronologia degli interventi sono alcuni degli aspetti problematici da affrontare, soprattutto nel caso delle cassature di difficile attribuzione e degli eventuali interventi postumi di Schlegel effettuati su

[3] https://giada-collab.github.io/considerations/ (cons. 22/01/2025).
[4] Il manoscritto è composto da tre volumi (ff. 313, 326 e 390).
[5] https://catalogue.bnf.fr/ark:/12148/cb313971643 (cons. 24/01/2025).

richiesta degli editori. Del resto, la modellizzazione deve spesso confrontarsi con la difficoltà di creare limiti invalicabili tra una categoria e l'altra (Van Hulle, 2016, p. 49). La distinzione tra le varie mani è effettuata tramite l'attributo @hand all'interno degli elementi che indicano modifiche (<add>, <del>). Negli stessi elementi verranno aggiunte informazioni sulla posizione (@place) ed il modo in cui appaiono i vari interventi nel manoscritto (@rend) perché anche queste caratteristiche materiali possono fornire informazioni utili per interpretare il lavoro a più mani. La scelta di fornire una versione dell'opera, in cui l'ultima volontà dell'autrice sembra avere contorni sfumati, porta alla rinuncia a fornire un testo di base per mezzo dell'elemento <lem> (TEI Consortium, 2023). Verrà quindi utilizzato unicamente <rdg> all'interno di <app>.

La riflessione sullo schema di codifica ha considerato due esempi che presentano problematiche simili a quelle delle *Considérations*: il progetto *The Shelley-Godwin Archive*[6] e *VaSto*.[7] Il primo progetto tiene conto del lavoro collaborativo svolto da Mary e Percy Shelley in *Fankenstein e* sperimenta l'integrazione tra una codifica "document focused" ed una "text focused" (Muñoz & Viglianti, 2015) dando la possibilità all'utente di selezionare l'intervento che vuole visualizzare a partire dall'attributo @hand presente nella codifica. Il secondo progetto, relativo all'edizione digitale della *Storia fiorentina* di Benedetto Varchi, considera l'intreccio tra volontà d'autore, quella della committenza (l'*entourage* mediceo) e le correzioni di "revisione" del figlio Marcello (Brancato et al., 2021, p. 141) in un'opera rimasta incompiuta. Questa edizione dà l'opportunità di confrontare la trascrizione semi-diplomatica, il testo critico e la versione finale censurata creando due testimoni fittizi: uno secondo l'ultima volontà dell'autore e uno secondo la volontà della committenza. Tutte queste scelte hanno fornito un primo spunto per l'elaborazione del nostro modello di codifica. Entrambi gli esempi evidenziano inoltre una tendenza delle attuali pratiche ecdotiche verso il superamento della dicotomia tra "the topology of the document and the chronology of the text" (Van Hulle, 2016, p. 52), superamento a cui mira anche il nostro progetto con l'intento di arricchire le informazioni ricavabili dai due approcci.

Gli esempi di edizioni genetiche che considerano la questione dell'autorialità plurima non sono numerosi[8] forse perché il fenomeno, in realtà piuttosto diffuso, è stato per lungo tempo ignorato dal dibattito teorico (Margras, 2022). Quest'ultimo ha iniziato ad intensificarsi solo recentemente[9] inducendo a ripensare le pratiche ecdotiche in direzione di una maggiore attenzione verso il lavoro collaborativo (Stillinger, 1991). Inoltre, molti progetti sono costituiti da edizioni genetiche dove il processo correttorio è individuato dall'utente unicamente confrontando le varie versioni messe a disposizione senza che esso sia preso in considerazione nello schema di codifica (Christen & Spadini, 2019, p. 84).[10] In altri casi questo schema non è consultabile.[11] Il progetto Saba 2021 propone un modello di codifica che mira a colmare questa lacuna mediante l'utilizzo dell'attributo @varSeq all'interno di <rdg> o <lem> per numerare le varie fasi correttorie (Fenu & Tancredi, 2022, p. 219), soluzione che è stata adottata anche nel nostro progetto con l'aggiunta dell'attributo @cert per indicare il grado di certezza nella ricostruzione cronologica degli interventi.

I dati sulla materialità del manoscritto danno informazioni utili a definire gli stadi genetici delle varie parti di cui è composto il manoscritto (Bustarret & Linkès, 2008, p. 134). Sono informazioni che non possono essere ricavate unicamente dal facsimile perché alcune caratteristiche materiali (tipo di carta, ecc.) richiedono l'esame dell'originale. Si dovrà quindi partire dai dati codicologici precedentemente raccolti[12] e codificarli basandosi principalmente sui moduli 11[13] e 12[14] delle linee guida TEI.

---

[6] http://shelleygodwinarchive.org (cons. 24/01/2025).

[7] https://dharc-org.github.io/progetto-vasto/index.html (cons. 24/01/2025).

[8] Per lo meno questi esempi non sono emersi ricercando nei cataloghi di Patrick Sahle (https://www.digitale-edition.de/, cons. 08/04/2025) e Greta Franzini (https://dig-ed-cat.acdh.oeaw.ac.at/, cons. 08/04/2025).

[9] Per restare solo in ambito francofono, alcuni lavori hanno messo in evidenza l'importanza di questa problematica per la teoria genetica (Brancher et al., 2018; Donin & Ferrer, 2015; Goudinoux & Bridoux-Michel, 2019). Sempre più riviste vi dedicano dei numeri: nn. 41 e 53 di *Genesis*, n. 3 di *Recto/Verso*, n. 10 di *Crisol* e n. 25 di *Continents manuscrits.*

[10] Come nel ben documentato progetto *Digital Edition of Fernando Pessoa. Projects and Publications*: http://www.pessoadigital.pt (cons. 09/04/2025).

[11] È il caso dell'edizione *Caves du Vatican* di André Gide: https://www.dhi.ac.uk/gide-les-caves-du-vatican/data/sommaire/front.htm (cons. 08/04/2025) oppure *Friedrich Dürrenmatt - Das Stoffe-Projekt*: https://fd-stoffe-online.ch (cons. 09/04/2025).

[12] Questi dati sono stati raccolti basandosi su un protocollo elaborato dal gruppo "Techniques et pratiques de l'écrit" (ora "Supports et tracés") dell'ITEM (CNRS/ENS).

[13] "Manuscript Description". https://tei-c.org/release/doc/tei-p5-doc/en/html/MS.html (cons. 24/01/2025).

[14] "Representation of Primary Sources". https://tei-c.org/release/doc/tei-p5-doc/en/html/PH.html (cons. 24/01/2025).

La nozione di "fluid text"[15] (Bryant, 2007) è particolarmente adatta al trattamento dell'ultima categoria di fenomeni da codificare: le divergenze tra manoscritto e prima edizione. Manoscritto e prima edizione sono considerati come due testimoni che vengono messi a confronto senza considerare uno in particolare come più autorevole ma evidenziando, nella codifica e nella visualizzazione, i cambiamenti effettuati tra una versione e l'altra. L'edizione digitale delle *Considérations* è data da questo insieme inscindibile di versioni ("fluid text") che permette di analizzare e interpretare la natura di questi cambiamenti (Bryant, 2007, p. 18). Questo approccio comporta anche la scelta di non codificare le divergenze come aggiunte o cassature perché non vediamo traccia di questa operazione nei documenti esaminati. Per esempio, una lezione assente in uno dei due documenti sarà codificata all'interno di <app> con un elemento <rdg> vuoto (Burnard et al., 2010) contenente l'attributo @wit per identificare il documento a cui si riferisce.
A partire da queste prime considerazioni, il modello di codifica verrà sviluppato e probabilmente modificato in conseguenza di un'analisi più dettagliata delle problematiche che emergeranno nell'analisi dei documenti. Si dovranno, in particolare, affrontare dei problemi di *overlapping* perché in uno stesso segmento testuale possono essere contemporaneamente presenti fenomeni di tipo genetico ed altri risultanti dalla collazione. Un ulteriore questione sarà il trattamento dei *marginalia* la cui tipologia è talmente varia da non trovare un elemento che possa comprenderli tutti, facendo venir meno la stessa interoperabilità (Estill, 2016).
L'edizione prevede inoltre l'annotazione semantica delle entità nominate. Nel nostro caso riguarderà i nomi di persona, le opere citate, le date ed i luoghi per i quali è previsto un livello granularità fino a edifici, piazze, ecc.
Infine, nel prototipo sono state adottate alcune misure in linea con le raccomandazioni del W3C[16] per siti web *responsive*, lavoro che dovrà essere ulteriormente sviluppato. La pagina web contiene attualmente la trascrizione di alcune pagine del manoscritto, risultato della trasformazione del file XML-TEI in HTML5 attraverso i fogli di stile XSLT. È tuttavia allo studio la possibilità di utilizzare uno strumento di visualizzazione e navigazione per la presentazione sinottica del testo del manoscritto, della prima edizione ed il loro facsimile. Una scelta da valutare sarà il software *open source* EVT 3[17] che presenta alcuni vantaggi come la facilità di configurazione ed il modo in cui possono essere gestiti i dati della codifica genetica per consentire una certa flessibilità di visualizzazione dei vari strati correttori (Buzzoni et al., 2024, pp. 218–219).

4. CONCLUSIONI

Il progetto sarà ulteriormente sviluppato a partire dagli spunti elencati finora ma, al di là dei risultati che saranno realmente ottenuti, l'obiettivo è fare in modo che questo lavoro possa dare un contributo alla riflessione sulla questione dell'autorialità plurima e sui possibili modi per rappresentarla nella sua complessità.

BIBLIOGRAFIA

Bacharach, S., & Tollefsen, D. (2015). You Complete Me: Posthumous Works and Secondary Agency. *Journal of Aesthetic Education*, *49*(4), 71–86. https://doi.org/10.5406/jaesteduc.49.4.0071

Balayé, S. (1977). Le Groupe de Coppet: Conscience d'une mission commune. *Le Groupe de Coppet. Actes et documents du deuxième Colloque de Coppet. 10-13 juillet 1974*, 29–45.

Brancato, D., Corbellini, M., Italia, P., Pasqual, V., & Priore, R. (2021). *VaSto*: Un'edizione digitale interdisciplinare. *Magazén*, *2*(1), 139–169. https://doi.org/10.30687/mag/2724-3923/2021/03/006

Brancher, D., Burg, G., Berjola, G., Demonet, M. L., Mueggler, N., Charité, C. L., Menini, R., Furno, M., Goeury, J., Morel, N., Dord-Crouslé, S., Cureau, S., Maira, D., Piéjus, A., Boutoux, C., Réach-Ngô, A., Cartron, M., Bail, M. L., Gnocchi, M. C., … Chartier, R. (2018). *L'éditeur a l'oeuvre: Reconsidérer l'auctorialité*. emono. https://emono.unibas.ch/emono/catalog/book/61

Broglie, V. de. (1886). *Souvenirs, 1785-1870, du feu duc de Broglie de l'Académie française*. Calmann-Lévy.

Bryant, J. (2007). Witness and Access: The Uses of the Fluid Text. *Textual Cultures*, *2*(1), 16–42.

[15] Una realizzazione di "fluid text" è la *Herman Melville Electronic Library*: https://mel.netlify.app/ (cons. 09/04/2025). Questo tipo di edizione prende in considerazione le varie versioni di un'opera con revisioni d'autore e non, comprendendo anche gli adattamenti.
[16] https://www.w3.org/ (cons. 24/01/2025).
[17] https://github.com/evt-project/evt-viewer-angular/ (cons. 24/01/2025).

Burnard, L., Jannidis, F., Rehbein, M., & Pierazzo, E. (2010). *An Encoding Model for Genetic Editions*. https://tei-c.org/Vault/TC/tcw19.html

Bustarret, C., & Linkès, S. (2008). De MUSE en ARGOLIDE, ou la codicologie à l'ère du numérique. *Recherches & Travaux*, *72*, 119–136. https://doi.org/10.4000/recherchestravaux.97

Buzzoni, M., Cucurnia, D., Fenu, C., Rosselli del Turco, R., & Tancredi, G. (2024). Progetto di edizione genetica digitale del Canzoniere manoscritto di U. Saba (1919-20). *Me.Te. Digitali. Mediterraneo in rete tra testi e contesti*, 215–219. https://doi.org/10.6092/unibo%2Famsacta%2F7927

Christen, A., & Spadini, E. (2019). Modeling genetic networks. Gustave Roud's œuvre, from diary to poetry collections. *Umanistica Digitale*, *7*. https://doi.org/10.6092/issn.2532-8816/9063

Donin, N., & Ferrer, D. (2015). Auteur(s) et acteurs de la genèse. *Genesis*, *41*, 7–26. https://doi.org/10.4000/genesis.1440

Estill, L. (2016). Encoding the Edge: Manuscript Marginalia and the TEI. *Digital Literary Studies*, *1*(1). https://journals.psu.edu/dls/article/view/59715

Fenu, C., & Tancredi, G. (2022). XML-TEI: Un modello per la filologia d'autore. *AIUCD 2022 - Proceedings*, *XI*, 218–222. https://doi.org/10.6092/unibo%2Famsacta%2F6848

Genand, S., & Rosset, F. (2022). Coppet et l'auctorialité. *Cahiers staëliens*, *71–72*, 9–12.

Goudinoux, V., & Bridoux-Michel, S. (2019). Introduction. *Déméter. Théories et pratiques artistiques contemporaines. Œuvrer à plusieurs : enjeux d'aujourd'hui*, *2*. http://www.peren-revues.fr/demeter/278

King, N. (1970). Lettres de Madame de Staël à Sir James Mackintosh: Lettre du 22 août 1817. *Cahiers staëliens*, *10*, 27–54.

Margras, F. (2022). *Simone et André Schwarz-Bart: Analyse d'une collaboration littéraire* [Phd thesis, Université des Antilles]. https://theses.hal.science/tel-04004854

Muñoz, T., & Viglianti, R. (2015). Texts and Documents: New Challenges for TEI Interchange and Lessons from the Shelley-Godwin Archive. *Journal of the Text Encoding Initiative*, *8*. https://doi.org/10.4000/jtei.1270

Pellegrini, C. (1974). *Madame de Staël e il gruppo di Coppet* (2. ed. riveduta). Pàtron.

Pierazzo, E. (2018). Il testo è morto: Lunga vita ai testi. Pluralismo testuale e edizioni digitali. *Prassi Ecdotiche della Modernità Letteraria*, *3*, 321–344. https://doi.org/10.13130/2499-6637/10633

Saintes, L. (2022). «Il faut concourir au grand œuvre». La pratique collaborative de l'écriture polémique dans le Groupe de Coppet. *Cahiers staëliens*, *71–72*, 109–126.

Staël-Holstein, G. de. (1818). *Considérations sur les principaux événemens de la Révolution françoise, ouvrage posthume de madame la baronne de Staël, publié par M. le duc de Broglie et M. le baron de Staël* (V. de Broglie & A.-L. de Staël-Holstein, A c. Di). Delaunay.

Staël-Holstein, G. de. (2017a). *Considérations sur les principaux événements de la Révolution française* (L. Omacini, A c. Di). Champion.

Staël-Holstein, G. de. (2017b). *De l'Allemagne* (A. Blaeschke, A c. Di). Honoré Champion.

Stillinger, J. (1991). *Multiple Authorship and the Myth of Solitary Genius*. Oxford University Press.

TEI Consortium. (2023). *12.1.1 «The Apparatus Entry». TEI P5: Guidelines for Electronic Text Encoding and Interchange. 4.7.0*. https://www.tei-c.org/release/doc/tei-p5-doc/en/html/TC.html#TCAPLL

Usteri, P., & Richter, E. (1903). *Lettres inédites de Madame de Staël à Henri Meister*. Hachette.

Van Hulle, D. (2016). Modelling a Digital Scholarly Edition for Genetic Criticism: A Rapprochement. *Variants*, *12–13*, 34–56. https://doi.org/10.4000/variants.293

Il corpus di prosa letteraria del progetto RIND (1830-1930). Assunti teorici e vincoli pratici

Stefano Ondelli[1], Pietro Mazzarisi[2]
[1] Università degli studi di Trieste, Italia - sondelli@units.it
[2] Università degli studi di Trieste, Italia - pietro.mazzarisi@units.it

ABSTRACT[1]
Questo contributo descrive la componente prodotta in Italia di un corpus di 1.000 testi comprendenti romanzi e raccolte di racconti scritti da autori italiani e tradotti da altre lingue e pubblicati tra il 1830 e il 1930. In particolare, si sofferma sui criteri ipotizzati per il bilanciamento del corpus atti a garantirne la rappresentatività e la possibilità di indagarne le caratteristiche con metodi quantitativi: per es. distribuzione temporale, lunghezza, genere e grado di appartenenza al canone letterario dei testi da una parte, genere e provenienza degli autori dall'altra. Si illustrano anche le difficoltà derivanti dalla scarsa conoscenza della popolazione di riferimento e i problemi pratici legati al reperimento dei testi.
Il corpus qui illustrato è oggetto di studio del progetto di ricerca "Reading the Italian Novel at a Distance (1830-1930)" (RIND). Scopo finale sarà la verifica empirica della periodizzazione tradizionale delle fasi dello sviluppo della narrativa in Italia (dal romanzo storico e verista al modernismo) e la valutazione del contributo a tale sviluppo apportato dai modelli esogeni accessibili tramite le traduzioni.
Parole chiave: discorso riportato; linguistica dei corpora; romanzo italiano; storia della letteratura italiana; traduzioni.

ABSTRACT
The corpus of literary prose of the RIND project (1830-1930). Theoretical assumptions and practical constraints. This paper describes the non-translated component of a corpus of 1,000 texts comprising novels and short story collections written by Italian authors and translated from other languages, published between 1830 and 1930. In particular, the paper discusses possible criteria for balancing the corpus in order to ensure it is representative and can be investigated through quantitative methods: e.g. the chronological distribution, size, genre and canonicity of texts on the one hand, and the gender and origin of the authors on the other. The difficulties arising from the lack of knowledge of the overall reference population, compounded by practical problems in finding the texts, are also illustrated.
The corpus illustrated here provides the material for the research project 'Reading the Italian Novel at a Distance (1830-1930)' (RIND). Its final objectives include the empirical verification of the traditional periodization of literary prose in Italy (from historical and realist novels to Modernism) and the evaluation of the contribution provided by foreign models made accessible to the general public through translations.
Keywords: reported speech; corpus linguistics; Italian novel; history of Italian literature; translations.

1. IL PROGETTO RIND (1830-1930)

Nel progetto interdisciplinare "Reading the Italian Novel at a Distance (1830-1930) - RIND" (PRIN 2022JAYFJH) confluiscono studi linguistici computazionali e dei corpora, teoria letteraria e narratologia cognitiva, scienza dei dati e statistica. La cornice metodologica è fornita dal concetto di "distant reading" (Moretti, 2013), atto a valutare la presenza, la distribuzione e il ciclo di vita dei fenomeni letterari attraverso l'analisi quantitativa di grandi collezioni di testi. I metodi quantitativi di analisi dei dati testuali (Lebart et al. 1998) sono stati applicati a numerosi campi, dalla storiografia alla giurisprudenza. Negli studi letterari, questi approcci hanno preso le mosse da stilometria classica e attribuzione dell'autorialità (Burrows, 1987; Hoover, 2013), per poi ampliarne il campo di applicazione (Jockers, 2013; Piper, 2018; Underwood, 2019).
Il progetto RIND analizza un corpus di 1.000 romanzi e raccolte di racconti di autori italiani o tradotti da originali stranieri e pubblicati in Italia tra il 1830 e il 1930, ovvero un periodo che, approssimativamente, va dalla nascita e affermazione del genere romanzesco (storico e non), passa dal Realismo del XIX secolo e giunge fino al Modernismo del primo Novecento. Le fasi e i compiti, distribuiti tra le unità di ricerca partecipanti, comprendono: la creazione del corpus; la definizione formale del Discorso Riportato (DR) a livello linguistico; la compilazione di repertori lessicali relativi all'ambientazione delle narrazioni; lo sviluppo

[1] Pur essendo questo contributo il risultato di uno sforzo congiunto, specifichiamo che Ondelli è responsabile dei §§ 1 e 2 e Mazzarisi dei §§ 3 e 4.

di metodi per estrarre i dati testuali, identificarne e rappresentarne l'evoluzione diacronica (serie temporali) e analizzarne il ciclo di vita; la valutazione delle tradizionali periodizzazioni dei movimenti letterari e la classificazione di generi e autori attraverso approcci quantitativi che incrocino l'illustrazione degli sviluppi sociali (l'ambientazione) e lo sviluppo delle tecniche narrative (DR); lo studio dell'interazione tra modelli stranieri e produzione italiana.
I dati linguistici che il progetto intende estrarre dal corpus appartengono a due ambiti principali: il DR e l'ambientazione delle vicende. Nel primo caso, l'obiettivo è rilevare i cambiamenti diacronici nella resa della "coscienza" dei personaggi di finzione, la cui istanziazione formale è identificabile nel DR in quanto insieme di strutture morfo-sintattiche che eccedono i confini frastici e investono il livello pragmatico-testuale. L'ipotesi di partenza è che l'evoluzione delle strategie di resa del DR rifletta la transizione tra diversi periodi letterari, come tradizionalmente riconosciuto dalla critica. A livello teorico, il DR può essere considerato parte della "polifonia" narrativa, introdotta da Bachtin (1997 [1974]) come dialogismo e ampiamente adottata nella teoria letteraria (Ducrot, 1984; Roulet, 2011). Finora la ricerca linguistica sul DR si è concentrata principalmente sulla tipologia, portando alla classificazione delle strutture dirette rispetto a quelle indirette e all'identificazione dei marcatori formali (Mortara Garavelli, 1995; Calaresu, 2004). Sono però pochi i tentativi di rilevare il DR con metodi quantitativi (Byszuk et al., 2020).
Anche nel caso dell'"ambientazione narrativa", il dato linguistico contribuisce a cogliere quantitativamente la nozione di *setting* come definita dalla geo-critica e dalle *spatial humanities* (per es., Cooper et al., 2016; Bushell, 2020). Pur riconoscendo il valore della vasta letteratura sull'impiego del *Named Entity Recognition* (NER) nella ricerca letteraria (Nouvel et al., 2016; Archana et al., 2018; Frontini et al., 2020), occorre notare che non è possibile limitarsi ai nomi propri (in particolare toponimi), ma è necessario includere qualsiasi luogo (naturale e architettonico, reale e fittizio) che contribuisce all'ambientazione in senso lato: città vs. campagna, interno vs. esterno, pubblico vs. privato ecc., con tutte le possibili sottocategorizzazioni (per es. edifici residenziali, commerciali, industriali, privati, pubblici) e sovrapposizioni (per es. commerciali vs. pubblici).
Alle ambientazioni contribuiscono inoltre i ruoli sociali dei personaggi, come figure professionali, gradi militari, cariche istituzionali e titoli religiosi. Naturalmente, i repertori lessicali si scontrano con la difficoltà di delimitare elenchi potenzialmente aperti e diversificati in base al periodo storico, mentre l'estrazione dei dati deve affrontare la questione della polisemia lessicale e della variabilità ortografica. Tuttavia, anche in questo caso, l'ipotesi di ricerca prevede che i cambiamenti nelle ambientazioni (per es. dalla campagna alla città) e nei ruoli sociali (per es. dai contadini agli operai) siano in grado di riflettere l'evoluzione delle tematiche attraverso le diverse fasi dello sviluppo sociale, nonché il passaggio tra diversi periodi letterari in Italia e all'estero (nelle traduzioni).
Infine, nel confronto tra le opere di autori italiani e le traduzioni di autori stranieri, si ipotizza che l'evoluzione diacronica dei dati testuali relativi ai due ambiti di indagine illustrati sopra getti luce anche sul contributo all'evoluzione del sistema letterario italiano da parte di modelli esogeni accessibili al grande pubblico (Even-Zohar, 1990; Toury, 1995).

2. IL CORPUS: PREMESSE TEORICHE

L'arco temporale (1830-1930) è stato scelto poiché copre il periodo che parte all'incirca dai primi esempi di romanzo in Italia fino all'ascesa del Modernismo e perché i testi possono essere liberamente consultati, manipolati e diffusi senza violare i diritti d'autore (Legge 22 aprile 1941, n. 633). Il corpus, una volta completato e bilanciato, sarà reso liberamente accessibile, scaricabile e consultabile per ulteriori ricerche sul sito del progetto (https://rind.units.it/home/) o su *Github*, con licenza CC BY-NC-ND 4.0.
La compilazione del corpus, atta a garantirne autenticità, rappresentatività e bilanciamento, è una delle fasi critiche dal punto di vista teorico e pratico (Bode, 2018; Ondelli, 2018; Odebrecht, Burnard & Schöch, 2021). Il progetto prevede il reperimento di 1.000 romanzi e raccolte di racconti e novelle, suddivisi in due subcorpora di 500 opere di autori italiani e 500 traduzioni da lingue straniere. I criteri di bilanciamento ipotizzati sono i seguenti: per gli autori, genere, origine geografica e numero di opere; per i testi, anno di pubblicazione, lunghezza, appartenenza al canone letterario o alla paraletteratura; nel caso delle traduzioni si aggiunge la lingua di partenza. I criteri applicati agli autori rendono conto delle variazioni diastratiche e diatopiche e sono mirati a evitare la sovrarappresentazione di singole personalità. I criteri relativi ai testi riguardano la copertura in diacronia e la presenza di generi letterari diversi, oltre a evitare la sovrarappresentazione di determinate opere a causa delle loro dimensioni.
Il problema principale ai fini del bilanciamento risiede nella mancata conoscenza della popolazione di riferimento, non esistendo un repertorio completo delle opere letterarie in prosa pubblicate in Italia (come

in altri paesi) nel periodo considerato, tantomeno corredato dei metadati qui presi in considerazione. Oltre a ciò, sorgono difficoltà teoriche in relazione ai seguenti aspetti: definizione dei fattori di possibile impatto linguistico in diatopia (luogo di nascita dell'autore vs. luogo di residenza vs. luogo di pubblicazione del testo); attribuzione delle singole opere al canone o alla paraletteratura; definizione delle fasce e dei limiti massimi e minimi di lunghezza dei testi; trattamento dei racconti pubblicati in momenti diversi e raccolti successivamente in volume; certezza della lingua di partenza della traduzioni. Poiché tali criteri sono subordinati alla fattibilità del corpus in relazione alla disponibilità di tempo e risorse, la loro valutazione in dettaglio può essere condotta più agevolmente dopo una prima ricognizione della disponibilità e reperibilità effettiva dei testi a partire da archivi digitali e biblioteche on-line come *Biblioteca Italiana, Progetto Manuzio, Project Gutenberg, Wikisource, Internet Archive, Hathi Trust* ecc., oltre a corpora costituiti in occasione di studi precedenti e messi gentilmente a disposizione da altri gruppi di ricerca.
Infine va sottolineato che i criteri adottati contraddistinguono il corpus RIND da altri corpora di (o con) prosa narrativa in lingua italiana. Per es., l'ELTeC (Odebrecht, Burnard & Schöch, 2021), nella sua sezione italiana, conta 100 romanzi, mentre il corpus RIND dieci volte tanti (se consideriamo anche le traduzioni); il periodo storico coperto dall'ELTeC è minore (1840-1920); ogni sua sezione è destinata a una singola nazione e non prevede la presenza di traduzioni. Altre differenze salienti riguardano il bilanciamento, che nell'ELTeC condiziona la compilazione *ab origine*. Infatti il progetto europeo prefissava per ciascun subcorpus nazionale almeno il 10%-50% di donne; la presenza di almeno 9-11 autori con 3 romanzi; almeno il 30% di opere appartenenti al canone e almeno il 30% di opere non canoniche (con la discriminante "canone/non canone" stabilita in base al numero di ristampe realizzate tra il 1970 e il 2009); almeno il 20% costituito da novelle (identificate come testi di lunghezza compresa tra le 10.000 e le 50.000 parole) e almeno il 20% composto di romanzi di ampio respiro (intendendo i testi di 200.000 e oltre parole); per ultimo, la lunghezza dei testi veniva distribuita in tre fasce: "brevi" (tra le 10.000 e le 50.000 parole), "medi" (tra le 50.000 e le 100.000 parole) e "lunghi" (oltre le 100.000 parole). Questi criteri hanno consentito al progetto europeo di ottenere risultati tendenzialmente omogenei nella creazione dei vari subcorpora nazionali, ma hanno limitato, e sicuramente condizionato, la rappresentatività del campione in relazione alla popolazione di riferimento (peraltro non nota).
Anche il progetto RIND ha affrontato il problema della insufficiente conoscenza della popolazione di riferimento, soprattutto nel caso della componente relativa alla paraletteratura e alle traduzioni. Tuttavia, piuttosto che impostare il bilanciamento del corpus secondo criteri predeterminati e arbitrari, si è optato per un approccio più spiccatamente – per così dire – *bottom up* che, tra l'altro, tenesse in considerazione gli aspetti legati alla fattibilità del progetto. In altre parole, dopo aver stabilito un numero congruo ma realistico di testi che renda possibile l'indagine statistica delle serie temporali dei fenomeni linguistici presi in considerazione (i 500 testi di ciascun subcorpus si concretizzano in poco meno di 5 testi per ognuno dei 101 anni considerati, con una media presunta di circa 250.000 parole per anno, stante la lunghezza media dei romanzi moderni), si è proceduto a un sondaggio preliminare teso a recuperare il maggior numero possibile di testi adatti disponibili in formato digitale e caratterizzati da un grado di "pulizia" sufficiente a ipotizzarne l'impiego nelle ricerche previste. Solo al termine di questa prima raccolta sono state prese decisioni ponderate alla luce della distribuzione interna dei testi secondo i criteri già considerati (distribuzione temporale, lunghezza, genere letterario, genere dell'autorialità, canonicità ecc.), basandosi sull'esistente ed eventualmente prevedendo integrazioni ad hoc per risolvere gli squilibri più evidenti. In altre parole, per quanto la raccolta dei testi già disponibili in formato elettronico non possa in alcun modo essere considerata un'immagine della popolazione esistente, alcune distribuzioni (per es. la maggiore o minore numerosità dei testi in alcune decadi, la minore o maggiore presenza delle autrici nel corso del tempo, l'andamento della lunghezza media dei testi, la distribuzione dei generi letterari ecc.) possono fornire indicazioni utili per il bilanciamento, soprattutto alla luce di eventuali conferme provenienti dalla storia della letteratura e dell'editoria italiana (e dei loro contatti con l'estero).

3. IL SUBCORPUS DI OPERE DI AUTORI ITALIANI: UN CENSIMENTO PRELIMINARE

Per quanto riguarda il subcorpus di opere di autori italiani, il sondaggio delle fonti già disponibili ha portato alla raccolta di 1.198 testi caratterizzati da un grado di "pulizia" accettabile ai fini delle analisi previste dal progetto. In assenza di dati certi sulla popolazione reale di riferimento (cioè le opere effettivamente pubblicate nel periodo considerato), si può assumere come riferimento l'andamento cronologico dei testi reperiti, illustrato in Fig. 1, possibilmente messo in relazione ai principali snodi della storia della letteratura e dell'editoria in Italia (Cadioli & Vigini, 2018; Ragone, 1999; Turi, 1997).

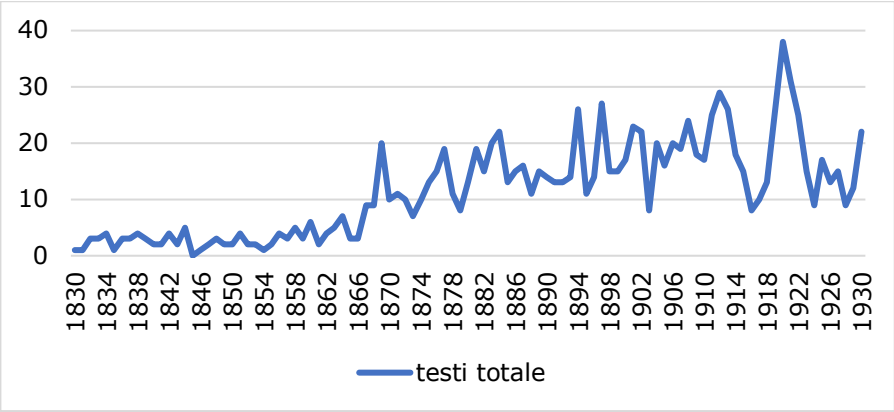


Figura 1. Distribuzione cronologica delle opere di autori italiani reperite

Per esemplificare gli assunti illustrati sopra (par. 2) per quanto concerne la significatività della distribuzione delle opere reperite, l'evidente penuria nel periodo precedente l'Unità può essere messa in relazione alla scarsa reattività del panorama nazionale alle novità letterarie almeno fino alla Scapigliatura e al Verismo, come testimoniano il notissimo saggio di Madame de Staël (1816) all'origine del dibattito che segna l'inizio del Romanticismo nella penisola a scapito di Arcadia e Neoclassicismo, ma anche il successivo contributo a firma di Niccolò Tommaseo e Giuseppe Bianchetti (1832) che, proprio nell'anno della morte di Walter Scott, illustra opinioni critiche e favorevoli in relazione all'introduzione del romanzo storico nella letteratura italiana. Un altro fattore determinante è da ricercarsi nell'iniziale mancanza di un'editoria attiva a livello nazionale, quando invece, nel decennio successivo all'Unità, l'ampliamento del pubblico dei lettori e delle attività editoriali determina un netto incremento delle pubblicazioni che si dimostrerà altalenante nei primi decenni del XX secolo, con battute d'arresto (probabilmente dovute a questioni di tipo economico e censorio) in corrispondenza della Grande Guerra e dell'avvento del Fascismo.
Prima di affrontare vari aspetti legati alle variabili considerate per il bilanciamento del corpus, occorre notare che la decisione di dimensionare il corpus in base al numero di testi è arbitraria: perché 1.000 testi e non 100 o 10.000? Perché non considerare piuttosto il numero di tokens? Benché gli esperimenti di Distant Reading previsti dal progetto non risentano particolarmente dell'accuratezza filologica, sulla scorta di un'ampia messe di studi precedenti (Jockers, 2013; Piper, 2018; Underwood, 2019) è senza dubbio preferibile includere lo stesso numero di testi completi nei due subcorpora, nonostante la possibilità che emergano differenze in termini di tokens. Considerate sia la scarsa disponibilità di testi in italiano digitalizzati con una certa cura sia le risorse economiche del progetto eventualmente destinabili a ulteriori reperimenti e digitalizzazioni, la soglia di 1.000 testi totali rappresenta un traguardo realistico e comunque ben superiore a progetti analoghi giù realizzati (cfr. Odebrecht, Burnard & Schöch, 2021).
La scelta di raccolte di racconti e novelle – oltre ai romanzi – quale format narrativo ammesso discende dall'ipotesi secondo cui le raccolte presentino una certa unitarietà di stile (DR) e contenuti (agentivi e ambientazioni), mentre la brevità dei singoli racconti condurrebbe a squilibri nel bilanciamento del corpus. Sempre per evitare tali squilibri, le soglie minima e massima di lunghezza dei testi sono state ipotizzate tra le 50.000 e 150.000 parole. All'atto pratico, la distinzione tra "romanzo", "racconto" e "novella" in base a semplici considerazioni dimensionali non risulta agevole; per es. è definita novella *Maestra* di Clarice Tartufari (1887) con 15.392 parole, ma anche *Gli americani di Rabbato* di Luigi Capuana (1912) con 31.366 parole. È invece definito "racconto" *Storia della Tonia* di Giovanni Battista Cereseto (1856) con 17.046 parole, ma anche *Catene* di Cordelia o Virginia Tedeschi Treves (1882) con 46.387 parole. Inoltre, l'etichetta "racconto" viene talvolta attribuita anche a testi divisi in più volumi come *La montanara* di Anton Giulio Barrili (1886), con 117.700 parole. Anche la letteratura per l'infanzia presenta grandi oscillazioni: *Il principino* di Ida Baccini (1881) con 16.996 parole è definito "racconto", così come *Uccelletto nero* di Guido Fabiani (1908) con il triplo delle parole (50.571). D'altra parte, sono chiamati "romanzi" testi assai brevi come *Amori moderni* di Grazia Deledda (1907) con 12.709 parole o *La colonia felice* di Carlo Dossi (1874) con 16.228 parole. Infine, per quanto riguarda le raccolte, bisogna prestare attenzione ristampe o riedizioni per non includere nel corpus due copie dello stesso testo. Capuana è un caso emblematico: il testo intitolato *Il dottor Ficicchia* compare con questo titolo nella raccolta *Nostra gente* del 1915, mentre una precedente versione – stavolta intitolata *Il dottore dei poveri* – era già stata inclusa nel 1894 nella raccolta di racconti dal titolo *Le paesane*.

Ancora in relazione alla lunghezza, sono state scartate le opere reperite che eccedevano il limite massimo previsto (150.000 parole); tuttavia, quando permettevano di colmare una certa lacuna temporale nel corpus (specialmente nei primi quarant'anni considerati), si è preferito mantenerle, come nel caso di *Cent'anni* di Giuseppe Rovani (1859) con 413.000 parole e i due testi di Nievo *Le confessioni di un italiano* (1858) con 332.000 parole e *Le confessioni di un ottuagenario* (1867) con 329.000 parole.
Per quanto concerne la datazione delle opere, in precedenza Bertinetto (2003) si era basato sulle date di nascita e morte degli autori per risolvere il problema delle pubblicazioni a puntate, riedizioni, revisioni ecc. Ai nostri fini si è, invece, preferito optare per la data dell'edizione princeps in volume sia dei romanzi che delle raccolte di novelle e racconti, salvo nei casi noti di forti interventi anche sull'assetto linguistico, per es. *I promessi sposi* di Alessandro Manzoni, di cui è stata inclusa la "Quarantana". Degno di menzione *Il castello di Trezzo* di Giambattista Bazzoni, pubblicato per la prima volta nel 1828 (quindi al di fuori del periodo considerato nella ricerca) e, dopo revisioni e correzioni, riedito nel 1838, data con la quale appare nel nostro corpus. Inoltre, abbiamo deciso di includere edizioni successive quando non è stata reperita la princeps in formato digitale (per es. *Il diavolo al Pontelungo* di Riccardo Bacchelli, edizione 1951, princeps 1927), poiché si è ritenuto che eventuali interventi non riguardanti l'assetto linguistico del testo (impaginazione, grafia ecc.) non abbiano alcun impatto sulle analisi che intendiamo realizzare, dato lo scarso peso del rigore filologico.
L'appartenenza o meno di un'opera al canone letterario dipende non solo dal periodo storico (il canone odierno è diverso da quello di un secolo fa; cfr. Ricci, 2013) ma anche dal contesto culturale: la distinzione tra letteratura alta e di consumo è relativamente netta in Italia, molto più sfumata nei Paesi anglofoni, con le ovvie ricadute in termini di classificazione delle traduzioni. Tuttavia, per verificare se la tradizionale periodizzazione delle correnti letterarie sia in qualche modo collegata all'evoluzione del DR e delle ambientazioni, pare necessario includere opere di ampia circolazione, seppur limitata nel tempo, per evitare di riproporre una storia della letteratura basata esclusivamente sul canone. I problemi di attribuzione sono molteplici, a cominciare dai possibili incroci tra letteratura di genere e valore letterario. Analogamente, poiché si è deciso di includere la letteratura per l'infanzia, è ovvio che, accanto a Collodi, anche autori come Salgari hanno pieno diritto a comparire nel corpus, e considerazioni analoghe sono valide nel caso della letteratura rosa o poliziesca. Infine, un altro problema riguarda se assegnare o meno al canone l'intera opera di un autore oppure solo alcune opere (e quali i criteri di eventuali esclusioni?). Per es., se Luigi Capuana può essere considerato senza discussioni un autore canonico, lo è di conseguenza anche la sua novella per l'infanzia *Scurpiddu* (1898)? Per prendere decisioni dotate di una certa oggettività, per il subcorpus qui considerato si è deciso di adottare come riferimento quattro storie della letteratura pubblicate negli ultimi 50 anni (Asor Rosa, 1985; Contini, 1992; Guglielmino & Grosser, 2001; Ferroni, 2012). Definiamo quindi canonici titoli e autori che compaiono in almeno due di esse, fatta salva ovviamente l'eventuale classificazione degli stessi come esempi di paraletteratura (per es. Carolina Invernizio e Francesco Mastriani in Gugliemino & Grosser, 2001).
Anche il genere degli autori incide sulla distribuzione in base al canone. In una storia letteraria come quella italiana, in cui l'affermazione della scrittura femminile giunge tardi rispetto al panorama europeo, autrici come Virginia Tedeschi Treves (Cordelia) e Anna Gentile Vertua non sono considerate canoniche, ma la prima è imprescindibile nella letteratura per l'infanzia, la seconda, con oltre 150 romanzi di discreto successo, è fautrice di una maggiore indipendenza femminile malgrado il sentimentalismo.
A questo proposito, il genere degli autori è un'altra variabile interessante di cui ci sfugge l'effettiva incidenza nel periodo considerato, anche in considerazione dell'incrocio con i generi letterari. Ancor più problematica è la ricaduta sul bilanciamento del corpus della provenienza geografica di opere e autori, stante la forte influenza della diatopia nella storia dell'italiano. A questo proposito, data la difficoltà nello scegliere un criterio di selezione univoco (luogo di nascita, a prescindere dai confini nazionali? residenza ufficiale o vissuto personale? sede della casa editrice?), pur registrando il metadato del luogo di nascita degli autori, il bilanciamento del corpus non ha tenuto conto della diatopia.

4. IL SUBCORPUS DI OPERE DI AUTORI ITALIANI: UNA SELEZIONE PLAUSIBILE

Tenendo conto della compresenza simultanea e degli incroci di tutti i criteri esposti sopra, sulla base dei testi reperiti in questa prima fase di compilazione la cernita delle 500 opere componenti il subcorpus di narrativa scritta direttamente in italiano è stata mirata a conservare il maggior numero possibile di opere di scrittrici (per permettere eventuali campionamenti) e a distribuire in maniera bilanciata i testi in base all'autorialità, onde evitare l'eccessiva incidenza di singoli idioletti. Al momento (ma il corpus è ancora da considerarsi in fieri e perfezionabile), le 500 opere sono riconducibili a 214 autorialità, di cui 45 donne

(21%) e 171 uomini (79%). Le autrici sono riconducibili a 121 opere (24%) mentre gli autori a 379 opere (76%). La Fig. 2 ne illustra la distribuzione cronologica.

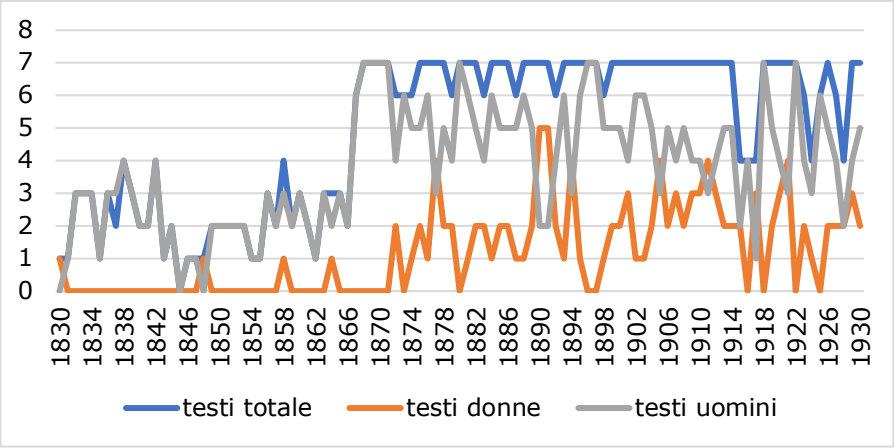


Figura 2. Distribuzione cronologica delle opere per anno, donne e uomini

Si tratta di una distribuzione realistica rispetto alla popolazione di riferimento? Sull'incremento del numero di opere edite nel tempo, valgono le considerazioni svolte sopra a commento della Fig. 1. Sul lato della distribuzione del genere, anche se in un periodo storico diverso, la selezione è confortata dal confronto con i vincitori dei maggiori premi letterari italiani: dal 1947 a oggi il premio *Strega* è stato attribuito a 13 donne (16,7%) e dal 1963 il premio *Campiello* è stato appannaggio di 17 donne (27%). Per quanto riguarda, invece, la lunghezza delle opere inserite nel subcorpus qui considerato, 10 non superano le 15.000, mentre 17 eccedono le 150.000 parole. La Fig. 3 illustra la distribuzione cronologica della lunghezza media e dimostra come, dopo una fase iniziale di definizione del genere "romanzo" e di penuria di testi disponibili, che registra differenze notevoli nel numero di opere selezionate per anno (fino alla completa assenza) e obbliga a mantenere testi molto brevi o molto lunghi, successivamente all'Unità si verifica una certa stabilizzazione.

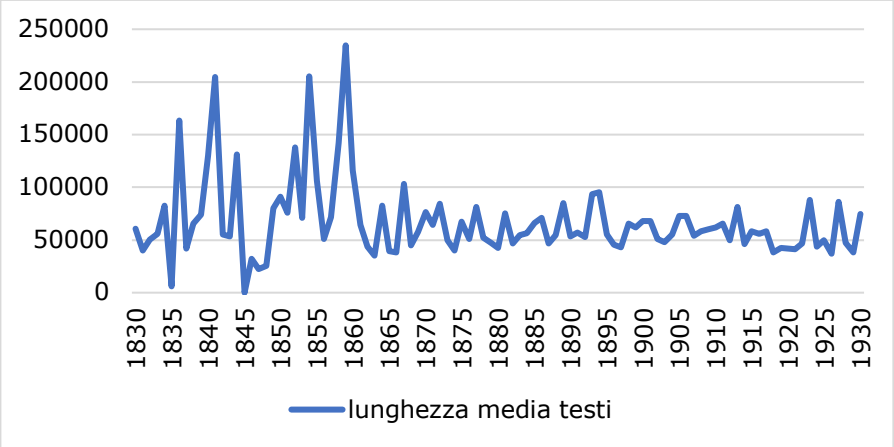


Figura 3. Distribuzione cronologica della lunghezza media dei testi in parole

Per motivi di spazio, i dati qui illustrati forniscono un'immagine solo parziale del subcorpus raccolto finora; peraltro una valutazione esauriente della sua rappresentatività dovrà tener conto della compresenza e dell'incrocio di tutti i criteri adottati, al fine di stabilire quali lacune colmare e che miglioramenti apportare in termini di bilanciamento e rappresentatività tramite il reperimento e la digitalizzazione ad hoc di testi non ancora disponibili in formato elettronico. Al momento quest'ultimo passaggio pare doversi indirizzare alla sostituzione di 27 opere per un migliore bilanciamento della distribuzione esistente. Pare invece più complessa, alla luce della scarsa "pulizia" dei testi disponibili e dell'onerosità di scansioni "a tappeto", l'eventuale integrazione delle opere che coprono le prime tre decadi considerate; non da ultimo, occorrerà attendere almeno una selezione parziale del subcorpus comprendente le traduzioni per ipotizzare interventi mirati a un bilanciamento complessivo ottimale.

BIBLIOGRAFIA

Archana, G., Gupta, V. & Kumar, M. (2018). Recent Named Entity Recognition and Classification techniques: A systematic review. Computer Science Review, 29, 21-43.

Asor Rosa, (1985). Storia della letteratura italiana. Firenze: La nuova Italia.

Bachtin, M. (1997). Estetica e romanzo. Torino: Einaudi.

Bertinetto, P. M. (2003). Tempi verbali e narrativa italiana dell'Otto/Novecento. Alessandria: Edizioni dell'Orso.

Bode, K. (2018). A World of Fiction: Digital Collections and the Future of Literary History. Ann Arbor: University of Michigan Press. https://doi.org/10.3998/mpub.8784777.

Burrows, J. F. (1987). Computation into Criticism: A Study of Jane Austen's Novels and an Experiment in Method. Oxford: Clarendon Press.

Bushell, S. (2020). Reading and Mapping Fiction: Spatialising the Literary Text. Cambridge: Cambridge University Press.

Byszuk, J., Woźniak, M., Kestemont, M., Leśniak, A., Łukasik, W., Šeļa, A. & Eder M. (2020). Detecting Direct Speech in Multilingual Collection of 19th-century Novels. Proceedings of LT4HALA 2020. 1st Workshop on Language Technologies for Historical and Ancient Languages. Sprugnoli R. & Passarotti M. (Eds.). 100-104. Marseille: European Language Resources Association (ELRA).

Cadioli, A. & Vigini, G. (2018). Storia dell'editoria in Italia dall'Unità a oggi. Milano: Editrice Bibliografica.

Calaresu, E. (2004). Testuali parole. La dimensione pragmatica e testuale del discorso riportato. Milano: Franco Angeli.

Contini, G. (1992). La letteratura italiana Otto-Novecento. Milano: BUR.

Cooper, D., Donaldson, C. & Murrieta-Flores, P. (Eds.). 2016. Literary Mapping in the Digital Age. London: Routledge.

Ducrot, O. (1984). Le dire et le dit. Paris: Les éditions de Minuit.

Even-Zohar, I. (1990). The Position of Translated Literature within the Literary Polysystem. Poetics Today 11/1, 45-51.

Ferroni, G. (2012). Storia della letteratura italiana. 4 voll. Milano: Mondadori università.

Frontini, F., Brando, C., Byszuk, J., Galleron, I., Santos, D. & Stanković, R. (2020). Named Entity Recognition for Distant Reading in ELTeC. CLARIN Annual Conference 2020 (5-7 October). Navarretta, C. & Eskevich, M. (Eds.). 37-41. Madrid: Virtual edition. https://hal.science/hal-03160438/document.

Gugliemino, S. & Grosser, H. (2001). Il Sistema letterario 2000. 12 voll. Milano: Principato.

Hoover, D. L. (2013). Quantative analysis and literary studies. A Companion to Digital Literary Studies. Siemens R. & Schreibman S. (Eds.). 517-533. Hoboken: Wiley & Sons.

Jockers, M. L. (2013). Macroanalysis: Digital Methods and Literary History. Champaign: University of Illinois Press.

Lebart, L., Salem, A. & Lisette, B. (1998). Exploring Textual Data. Dordrecht: Kluwer Academic Publication.

Moretti, F. (2013). Distant Reading. London-New York: Verso Books.

Mortara Garavelli, B. (1995). Il discorso riportato. Grande grammatica italiana di consultazione. Vol. III. Renzi, L., Salvi, G. & Cardinaletti, A. (Eds.). 426-468. Bologna: Il Mulino.

Necker, A.-L.-G. (1816). Sulla maniera e la utilità delle Traduzioni. Biblioteca italiana, 1, Gennaio, pp. 9-18.

Nouvel, D., Ehrmann, M. & Rosset, S. (2016) Named Entities for Computational Linguistics. Hoboken: Wiley & Sons.

Odebrecht, C., Burnard, L. & Schöch, C. (Eds.). (2021). European Literary Text Collection (ELTeC), version 1.1.0, April 2021. COST Action (CA16204), Distant Reading for European Literary History. https://doi.org/10.5281/zenodo.4662444).

Ondelli, S. (2018). Treat Texts as Data but Remember They Are Made of Words: Compiling and Processing Corpora. Tracing the Life Cycle of Ideas in the Humanities and Social Sciences. Tuzzi, A. (Ed.). 133-150. Cham: Springer.

Piper, A. (2018). Enumerations: Data and Literary Study. Chicago: The University of Chicago Press.

Ragone, G. (1999). Un secolo di libri. Storia dell'editoria dall'Unità d'Italia al post-moderno. Torino: Einaudi.

Ricci, L. (2013). Paraletteratura. Lingua e stile dei generi di consumo. Roma: Carocci.

Roulet, E. (2011). Polyphony. Discursive Pragmatics. Zienkowski, J., Östman, J-O. & Verschueren, J. (Eds.). 208-222. Amsterdam: John Benjamins.
Tommaseo, N. & Bianchetti, G. (1832). Discorsi critici intorno alla questione se giovi di ammettere o no nella letteratura italiana il romanzo storico. Treviso: Tipi di Gio. Paluello del fu Antonio.
Toury, G. (1995). Descriptive Translation Studies and Beyond. Amsterdam: John Benjamins.
Turi, G. (Ed.). 1997. Storia dell'editoria nell'Italia contemporanea. Firenze: Giunti.
Underwood, T. (2019). Distant Horizons: Digital Evidence and Literary Change. Chicago: The University of Chicago Press.

ZoneRW: verso un'integrazione con Kraken ed eScriptorium per il riconoscimento e la gestione avanzata delle regioni di interesse

Pietro Sichera[1], Angelo Mario Del Grosso[2], Laura Mazzagufo[3], Daria Spampinato[4]
[1]CNR ILIESI, Italia pietro.sichera@cnr.it
[2] CNR ILC, Italia - angelomario.delgrosso@cnr.it
[3]CNR ISTC, Italia - laura.mazzagufo@istc.cnr.it
[4]CNR ISTC, Italia - daria.spampinato@cnr.it

ABSTRACT (ITALIANO)

L'identificazione delle regioni di interesse (*region of interest*, ROI) nei documenti facsimilari è essenziale per un corretto processo di digitalizzazione e per lo studio sia di testi manoscritti sia di testi a stampa. Questo contributo presenta un'estensione di funzionalità del software ZoneRW (nato nel contesto dell'edizione scientifica digitale *Bellini Digital Correspondence*), per integrare il tool Kraken, dedicato al rilevamento automatico delle zone e al riconoscimento automatico del testo, e l'ambiente digitale eScriptorium per la gestione avanzata del processo di digitalizzazione a partire da repertori di immagini di documenti testuali. Il workflow proposto utilizza il formato XML-PAGE per garantire interoperabilità e scalabilità, permettendo contestualmente di definire, modificare ed esportare le regioni di interesse verso eScriptorium. Inoltre, le nuove funzionalità di ZoneRW consentono di collegare i documenti di descrizione di immagini mediante protocollo IIIF (manifest IIIF). Il contributo evidenzia e discute le differenze tra i formati XML-PAGE e XML-ALTO nonché le prospettive per futuri sviluppi dello strumento ZoneRW nel contesto delle Digital Humanities.
Parole chiave: ZoneRW; regioni di interesse; Kraken; eScriptorium; digitalizzazione.

ABSTRACT (ENGLISH)

ZoneRW: Towards Integration with Kraken and eScriptorium for Advanced Recognition and Management of Regions of Interest

The detection of regions of interest (ROI) in facsimile documents is essential for a proper digitization process and for the study of both manuscript and printed texts. This contribution presents a feature extension of the ZoneRW software (created in the context of the *Bellini Digital Correspondence* digital scholarly edition), to integrate the Kraken tool, dedicated to automatic zone detection and automatic text recognition, and the eScriptorium digital environment for advanced management of the digitization process from image repositories of textual documents. The proposed workflow uses the XML-PAGE format to ensure interoperability and scalability, while simultaneously allowing regions of interest to be defined, edited, and exported to eScriptorium. In addition, the new ZoneRW capabilities allow image description documents to be linked via IIIF protocol (manifest IIIF). The paper highlights and discusses the differences between XML-PAGE and XML-ALTO formats as well as prospects for future developments of the ZoneRW tool in the Digital Humanities context.
Keywords: ZoneRW; regions of interest; Kraken; eScriptorium; digitization.

1. INTRODUZIONE

Il riconoscimento delle regioni di interesse consente di delimitare aree rilevanti presenti all'interno di documenti. Il riconoscimento e la classificazione in automatico di tali regioni, insieme al relativo contenuto, facilita la trascrizione e l'interpretazione delle fonti digitalizzate.

ZoneRW,[1] (Sichera P., 2023) strumento sviluppato originariamente nel contesto delle attività volte alla realizzazione dell'edizione scientifica digitale *Bellini Digital Correspondence*,[2] è una applicazione stand-alone progettata per la gestione CRUD (Create, Read, Update, Delete) delle zone codificate nei file XML-TEI dell'edizione, mediante un'interfaccia user-friendly e intuitiva. La caratteristica principale di ZoneRW prevedeva inizialmente il trasferimento delle coordinate delle zone di interesse in un file XML-TEI; successivamente l'applicativo si è dimostrato uno strumento estremamente flessibile, che ha consentito di estendere il suo utilizzo verso nuove funzionalità. Grazie a questa flessibilità, ZoneRW si è evoluto in un

[1] La versione di ZoneRW è in fase di test. La versione stabile attualmente pubblicata è disponibile al link https://github.com/pierpaolosichera/ZoneRW (cons. 26/01/2025)

[2] *Bellini Digital Correspondence* - https://bellinicorrespondence.cnr.it/ (cons. 26/01/2025)

componente in grado di gestire l'interoperabilità tra strumenti diversi. In particolare, la prima estensione è rivolta all'integrazione con Kraken,[3] un sistema open source per il rilevamento automatico delle zone di interesse e il riconoscimento ottico del testo, e, in secondo luogo, con eScriptorium,[4] una piattaforma per la gestione avanzata dei dati provenienti da campagne di digitalizzazione e il supporto di workflow collaborativi. Queste nuove caratteristiche hanno ampliato notevolmente le potenzialità del software.
L'estensione proposta consente di collegare i suddetti strumenti grazie al formato XML-PAGE,[5] uno standard per la descrizione del layout, della struttura e dei contenuti testuali presenti in documenti digitalizzati. Questo approccio garantisce interoperabilità e scalabilità nonché la possibilità di definire, modificare ed esportare zone di interesse verso l'ambiente eScriptorium. Inoltre, ZoneRW permette di associare ad ogni immagine il relativo documento manifest IIIF (International Image Interoperability Framework),[6] migliorando l'accessibilità e la condivisione delle risorse digitali.
Il beneficio di questa integrazione è duplice: da un lato, semplificare il workflow di digitalizzazione riducendo l'intervento manuale; dall'altro, migliorare l'accuratezza del riconoscimento automatico del testo grazie all'uso combinato di ZoneRW e Kraken. Il presente contributo discute inoltre le differenze tra i due formati di riferimento per la rappresentazione di risorse testuali in ambienti OCR/HTR, vale a dire XML-PAGE e XML-ALTO,[7] evidenziando i vantaggi offerti dal primo formato rispetto al secondo nell'ambito delle Digital Humanities.

2. STATO DELL'ARTE

Diverse soluzioni software dedicate al rilevamento automatico delle zone di interesse in documenti scritti si sono affermate nel tempo (Pinche A., 2023), con approcci che spaziano dall'elaborazione manuale alla completa automazione (Clérice T., 2022). Tra le più note si annovera TEIZoner (Dumont B., 2022),[8] uno strumento online che permette di segmentare i documenti in zone e generare file XML-TEI (Burnard L., 2014). Tuttavia, TEIZoner non consente di memorizzare i risultati della segmentazione in modo permanente per eventuali ulteriori modifiche: questo limite riduce l'efficienza dei workflow che richiedono interazioni ripetute, attività collaborative o l'integrazione con altri strumenti. In questo contesto, ZoneRW si distingue offrendo funzionalità CRUD per la gestione avanzata delle zone, permettendo agli utenti di apportare modifiche dinamiche e persistenti alle informazioni. Un altro strumento online di annotazione è Recogito,[9] che consente di creare un account dove memorizzare i propri progetti. Lo spazio di memorizzazione però è molto limitato (200MB), e rende necessario, nel caso di progetti di grandi dimensioni, l'impiego di un server IIIF per il caricamento delle immagini da annotare. Un ulteriore tool da segnalare è IMA[10] (Image annotation tool), un sistema di annotazione per singole immagini. Il tool esporta in formato XML-TEI, ma non gestisce le <surface> multiple.
Le piattaforme digitali per la trascrizione, l'annotazione e l'analisi di testi, come Transkribus[11] ed eScriptorium, sono focalizzati sull'attività di trascrizione automatica di testi (ATR). Pur eccellendo nell'analisi testuale, queste piattaforme richiedono, almeno inizialmente, un intervento manuale per la configurazione delle regioni di interesse, specialmente in contesti complessi. Strumenti come Kraken e Tesseract[12] hanno mostrato notevoli progressi nel rilevamento automatico delle zone e nel riconoscimento testuale (Kay A., 2007), ma non offrono nativamente funzionalità grafiche moderne[13] per la modifica delle zone di interesse, così come non supportano nativamente workflow completamente integrati con formati di rappresentazione standard come XML-PAGE o XML-ALTO.

[3] https://kraken.re/main/index.html (cons. 26/01/2025)

[4] https://escriptorium.readthedocs.io/en/latest/ (cons. 26/01/2025)

[5] https://github.com/PRImA-Research-Lab/PAGE-XML (cons. 26/01/2025)

[6] https://iiif.io/guides/using_iiif_resources/ (cons. 26/01/2025)

[7] https://loc.gov/standards/alto/ (cons. 26/01/2025)

[8] http://teicat.huma-num.fr/zoner.php (cons. 26/01/2025)

[9] https://recogito.pelagios.org/ (cons. 10/04/2025)

[10] https://ima.coders.tools/ (cons. 26/01/2025)

[11] https://www.transkribus.org/ (cons. 26/01/2025)

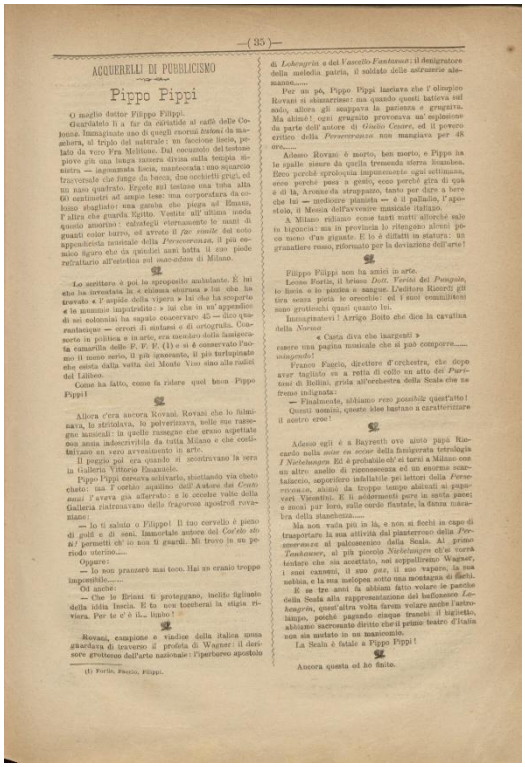
[12] https://github.com/tesseract-ocr/tesseract (cons. 26/01/2025)

[13] Ad esempio, Tesseract-OCR QT4 è una semplice GUI per Tesseract ma non viene aggiornata dal 2011 https://github.com/zdenop/tesseract-ocr-qt4gui (cons. 26/01/2025)

L'estensione di ZoneRW supera tali limiti grazie alla sua flessibilità e alla capacità di integrarsi con sistemi come Kraken ed eScriptorium. La possibilità di generare file XML-PAGE o XML-ALTO, entrambi standard consolidati nel panorama della digitalizzazione, consente a ZoneRW di garantire una piena interoperabilità nell'ambito di progetti complessi di ambito DH e perseguire al meglio i principi FAIR per la scienza aperta (Dumouchel, S & al., 2020).

3. IMPLEMENTAZIONE DEL WORKFLOW

L'implementazione del workflow proposto si basa sull'integrazione di ZoneRW con Kraken ed eScriptorium. ZoneRW è stato sviluppato utilizzando la piattaforma RAD per database 4th Dimension (4D),[14] che offre tra gli altri un supporto nativo per XML tramite la libreria Xerces della Apache Foundation. La gestione delle strutture XML avviene tramite lo standard DOM (Document Object Model), che consente di accedere e manipolare in maniera nativa ed efficiente ogni elemento della struttura XML. Questo approccio permette non solo di navigare l'albero XML, ma anche di aggiornare, riorganizzare o aggiungere elementi in tempo reale senza necessità di operazioni intermedie complesse. La gestione dinamica dei nodi <surface>, <graphic> e <zone> avviene con estrema precisione, ottimizzando i tempi di elaborazione e garantendo la piena coerenza dei dati. Le regioni di interesse vengono rappresentate utilizzando il formato SVG (Scalable Vector Graphics), che garantisce un'accurata visualizzazione e modifica delle coordinate direttamente sull'immagine. L'interfaccia grafica di ZoneRW è progettata per essere user-friendly e reattiva, permettendo all'utente di lavorare agevolmente su immagini ad alta risoluzione. ZoneRW accetta due modalità di input: immagini accoppiate a file XML-TEI, oppure cartelle contenenti esclusivamente immagini. Nel primo caso, ZoneRW utilizza il tag <surface> per rappresentare la pagina e i relativi tag <graphic> e <zone> per identificare visivamente le regioni di interesse. Ogni zona può essere creata o modificata manualmente tramite un'interfaccia grafica intuitiva, che consente di aggiungere o correggere coordinate e identificativi delle zone senza necessità di intervento esterno. Nel secondo caso, quando vengono caricate cartelle di immagini senza XML associato, l'utente ha la possibilità di lanciare da ZoneRW un'istanza di Kraken per il rilevamento automatico delle zone di testo e il riconoscimento OCR. Kraken, un potente strumento open source basato su Python, utilizza modelli di deep learning per segmentare il layout delle immagini e riconoscere il testo, anche su documenti complessi.



ACQUERELLI DI PUBBLICISMO

Pippo Pippi

Fig.1 Riproduzione fotografica di una pagina della rivista «La Farfalla» (anno 1876, n.2, fascicolo 5, p. 35)

[14] https://www.4d.com (cons. 26/01/2025)

I file generati includono le coordinate delle zone e i dati di riconoscimento testuale, pronti per essere importati in ZoneRW per ulteriori modifiche o validazioni.

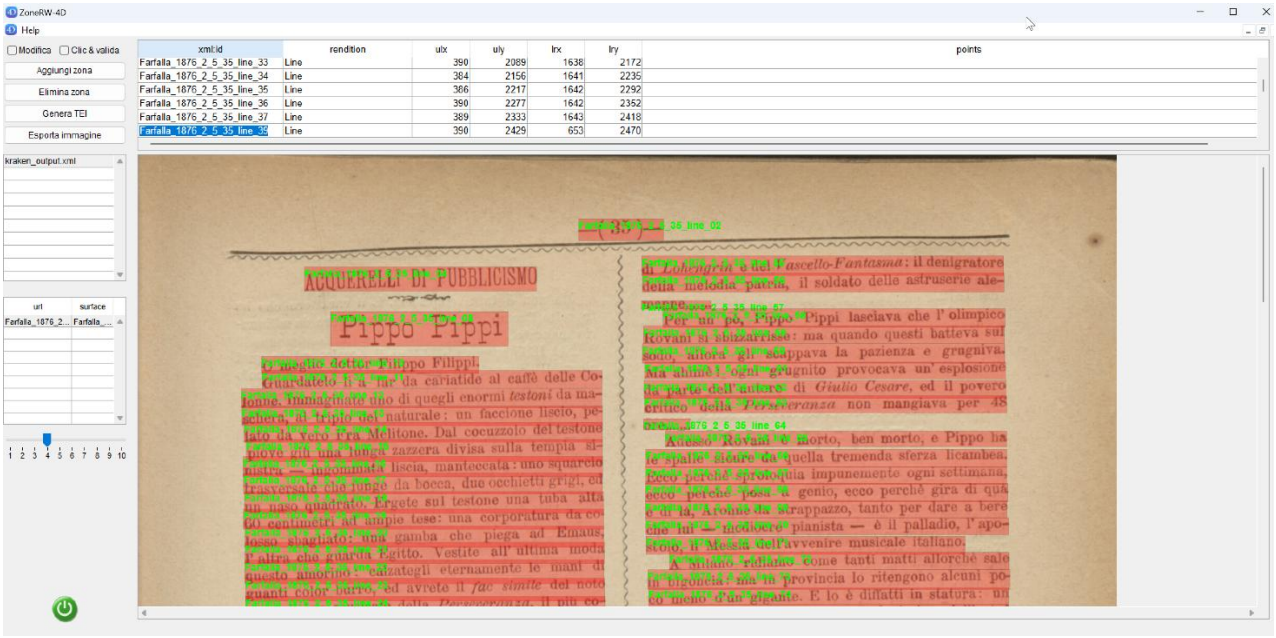


Fig.2 Le regioni di interesse della stessa pagina de «La Farfalla» importate in ZoneRW

ZoneRW consente di modificare, arricchire e correggere le zone rilevate e importate da Kraken. L'utente può inoltre associare ogni immagine a un manifest IIIF, un file in formato JSON standardizzato per descrivere e organizzare risorse digitali ospitate su server compatibili con il protocollo IIIF. Questo protocollo è progettato per facilitare la gestione e la condivisione di immagini ad alta risoluzione, fornendo metadati dettagliati e percorsi di accesso strutturati alle risorse. L'adozione dei manifest IIIF elimina la necessità di caricare manualmente le immagini in eScriptorium, consentendo invece un accesso diretto e referenziato alle immagini già ospitate su server IIIF. Tale approccio non solo migliora l'accessibilità e la tracciabilità delle risorse, ma ottimizza anche il workflow riducendo i tempi e i costi associati alla duplicazione dei dati.
Una volta completata l'elaborazione, ZoneRW esporta le informazioni sulle zone in formato XML-PAGE che, grazie alla sua capacità di rappresentare layout complessi, è il principale standard supportato da eScriptorium, permettendo un'integrazione diretta con quest'ultimo. eScriptorium consente di esportare i dati anche nel formato XML-ALTO, che è però maggiormente orientato alla descrizione del testo.
Ogni file XML-PAGE (vd. Fig.3) descrive una pagina digitale, tipicamente un'immagine derivata dalla digitalizzazione di un documento fisico, attraverso una struttura gerarchica ricca di informazioni.

```xml
<?xml version="1.0" encoding="UTF-8"?>
<PcGts xmlns="http://schema.primaresearch.org/PAGE/gts/pagecontent/2019-07-15">
  <Metadata>
    <Creator>ZoneRWProcessor</Creator>
    <Created>2025-01-01T12:00:00</Created>
    <LastChange>2025-01-01T12:30:00</LastChange>
  </Metadata>
  <Page imageWidth="2480" imageHeight="3508" imageFilename="ilmomento.png">
    <TextRegion id="region_1">
      <Coords points="100,100 500,100 500,200 100,200"/>
      <TextEquiv>
        <Unicode>Primo esempio di testo.</Unicode>
      </TextEquiv>
    </TextRegion>
    <TextRegion id="region_2">
      <Coords points="600,300 1000,300 1000,400 600,400"/>
      <TextEquiv>
        <Unicode>Secondo esempio di testo.</Unicode>
      </TextEquiv>
    </TextRegion>
  </Page>
</PcGts>
```

Fig.3 File in formato XML-PAGE

Il file contiene metadata che documentano il processo di creazione, tra cui informazioni sulla provenienza dell'immagine, i dettagli tecnici del processo di digitalizzazione (ad esempio, la risoluzione o il software utilizzato) e altri dati descrittivi.
La struttura del tag <Page> rappresenta l'elemento principale, che descrive l'immagine della pagina stessa con informazioni come dimensioni (gli attributi @imageWidth e @imageHeight) e un riferimento all'immagine sorgente (attributo @imageFilename). Questo elemento contiene una successione ordinata di <TextRegion>, che rappresentano le diverse zone di testo presenti nella pagina. Ogni <TextRegion> è ulteriormente arricchito da elementi figli come <Coords>, che definisce le coordinate geometriche delle zone di testo mediante poligoni o rettangoli, e <TextEquiv>, che include il contenuto testuale associato ad ogni zona. Allo stesso livello di <TextRegion> il tagset PAGE supporta una rappresentazione di altre tipologie di regioni, come <ImageRegion> per le immagini contenute nella pagina, <SeparatorRegion> per le linee di separazione, e <TableRegion> per le tabelle, ognuno sempre con le relative coordinate.
XML-ALTO (vd. Fig.4), invece, è maggiormente focalizzato sulla descrizione testuale e sullo stile dei caratteri, risultando ridondante rispetto alle necessità di esportazione di ZoneRW.

```xml
<?xml version="1.0" encoding="UTF-8"?>
<alto xmlns="http://www.loc.gov/standards/alto/ns-v4#">
  <Description>
    <MeasurementUnit>pixel</MeasurementUnit>
    <sourceImageInformation>
      <fileName>ilmomento.png</fileName>
    </sourceImageInformation>
    <OCRProcessing ID="ocr_1">
      <ocrProcessingStep>
        <processingSoftware>
          <softwareName>ZoneRWProcessor</softwareName>
          <softwareVersion>1.0</softwareVersion>
        </processingSoftware>
      </ocrProcessingStep>
    </OCRProcessing>
  </Description>
  <Layout>
    <Page WIDTH="2480" HEIGHT="3508" PHYSICAL_IMG_NR="1" ID="page_1">
      <PrintSpace>
        <TextBlock ID="block_1" HPOS="100" VPOS="100" WIDTH="400" HEIGHT="100">
          <TextLine>
            <String CONTENT="Primo esempio di testo."/>
          </TextLine>
        </TextBlock>
        <TextBlock ID="block_2" HPOS="600" VPOS="300" WIDTH="400" HEIGHT="100">
          <TextLine>
            <String CONTENT="Secondo esempio di testo."/>
          </TextLine>
        </TextBlock>
      </PrintSpace>
    </Page>
  </Layout>
</alto>
```

Fig.4 Lo stesso file mostrato in Fig.3, in formato XML-ALTO

4. CONCLUSIONI E SVILUPPI FUTURI

ZoneRW si è dimostrato uno strumento versatile e adattabile per la gestione delle zone di interesse nei documenti digitalizzati. La sua capacità di integrare funzionalità di rilevamento automatico con quelle di modifica manuale dei dati lo rende uno strumento molto efficace all'interno di progetti complessi di ambito DH. L'integrazione con strumenti quali Kraken ed eScriptorium, attraverso il formato XML-PAGE, garantisce un flusso di lavoro fluido e interoperabile, migliorando l'efficienza e riducendo il carico operativo per gli utenti, seguendo principi di scienza aperta e FAIR. Il supporto per i manifest IIIF rappresenta un ulteriore avanzamento nella gestione e accessibilità delle risorse digitali.
Come per qualsiasi prodotto software, esistono sono ampie opportunità per migliorare ulteriormente ZoneRW, ampliandone le potenzialità.
Un primo sviluppo previsto è il supporto del formato XML-ALTO, sia in fase di importazione delle ROI che in esportazione delle stesse.
Uno sviluppo futuro importante consiste nella realizzazione di una versione web-based di ZoneRW, che permetterebbe agli utenti di accedere alle sue funzionalità direttamente da un browser senza necessità di

installazioni locali. Questo approccio aumenterebbe notevolmente la diffusione dello strumento, rendendolo accessibile a una comunità di utenti più ampia e facilitando la collaborazione in progetti distribuiti.
Un'altra possibile direzione di sviluppo riguarda l'esposizione di interfacce di programmazione (API) compatibili con il formato richiesto da sistemi di integrazione e orchestrazione di servizi, quale ad esempio il MarketPlace in corso di sviluppo all'interno di H2IOSC,[15] il progetto che federa i nodi italiani delle infrastrutture europee CLARIN, DARIAH, E-RIHS e OPERAS. Il MarketPlace di H2IOSC avrà la capacità di orchestrare API (Sichera & al. 2024) provenienti da sorgenti e provider (endpoint) differenti, facilitando la condivisione e l'interoperabilità delle risorse digitali tra diversi sistemi e progetti. Questa implementazione consentirebbe a ZoneRW di integrarsi agevolmente con altre piattaforme e applicazioni disponibili sul MarketPlace.
In conclusione, ZoneRW si configura non solo come uno strumento per la gestione delle regioni di interesse, ma come un utile componente in un ecosistema di strumenti interoperabili, contribuendo alla valorizzazione e alla fruizione delle risorse digitalizzate in contesti accademici, archivistici e museali. Il suo sviluppo continuo risponde alle esigenze in evoluzione delle Digital Humanities, promuovendo la sostenibilità e l'innovazione nei progetti di trascrizione, codifica e valorizzazione di risorse testuali.

BIBLIOGRAFIA

Burnard, L. (2014). What is the Text Encoding Initiative? (1-). OpenEdition Press. https://doi.org/10.4000/books.oep.426

Clérice, T. (2022). You Actually Look Twice At it (YALTAi): Using an object detection approach instead of region segmentation within the Kraken engine. ArXiv, abs/2207.11230. https://api.semanticscholar.org/CorpusID:251018572

Cristofaro, S., Del Grosso, A.M., Mazzagufo, L., Sichera, P., & Spampinato D. (2023). "Bellini Digital Correspondence: A Model for Making Collaborative Digital Scholarly Editions." In *2023 7th IEEE Congress on Information Science and Technology (CiSt)*, 615–20. IEEE. https://doi.org/10.1109/CiSt56084.2023.10409920

Del Grosso, A. M., Spampinato, D. (Eds.). (2023). Bellini Digital Correspondence. CNR Edizioni, <https://bellinicorrespondence.cnr.it>. ISBN: 978-88-8080-562-5.

Dumont, B. (2022). "Review of 'TEI Critical Apparatus Toolbox: Web-based tools for ongoing XML-TEI editions'." RIDE 15. doi: 10.18716/ride.a.15.4.

Dumouchel, S., Blotière, E., Breitfuss, G., Chen, Y., Donato, F. D., Eskevich, M., Forbes, P., Georgiadis, H., Gingold, A., Gorgaini, E., Moranville, Y., Pohle, S., Paoli, S. de, Petitfils, C., & Toth-Czifra, E. (2020). GOTRIPLE: A User-Centric Process to Develop a Discovery Platform. Information, 11(12), 563. https://doi.org/10.3390/info11120563

Kay, A. (2007). Tesseract: An open-source optical character recognition engine. Linux J., 2007(159), 2.

Pinche, A. (2023). Generic HTR Models for Medieval Manuscripts. The CREMMALab Project. J. Data Min. Digit. Humanit., 2023. https://api.semanticscholar.org/CorpusID:264424158

Sichera P. (2023). pierpaolosichera/ZoneRW: v0.5-beta (7th IEEE CiSt'23) (v0.5-beta). Zenodo. https://doi.org/10.5281/zenodo.5599509

Sichera, P., Marras, C., & Pasini, E. (2024). Orchestrazione API per workflow applicativi nell'ambito delle Digital humanities. Zenodo. https://doi.org/10.5281/zenodo.14187534

[15] https://www.h2iosc.cnr.it/ (cons. 26/01/2025)

L’edizione digitale dei primi 16 Taccuini di Paolo Orsi

Giuseppina Monterosso[1], Andrea Bolioli[2], Elisa Bonacini[3], Gianmario Cattaneo[4], Dario Gonella, Anna Maria Marras[5], Salvatore Spina[6], Paola Venuti[7]

[1]Parco Archeologico e Paesaggistico di Siracusa, Eloro, Villa del Tellaro e Akrai - giuseppina.monterosso@regione.sicilia.it

[2]Ricercatore e consulente - andrea.bolioli@gmail.com

[3]Università di Bari «Aldo Moro», Dipartimento di Ricerca e Innovazione Umanistica - elisa.bonacini@uniba.it

[4]Università degli Studi del Piemonte Orientale - gianmario.cattaneo@uniupo.it

[5]Università di Torino, Dipartimento di Studi Storici - annamaria.marras@unito.it

[6]Università degli Studi di Catania, Dipartimento di Scienze Umanistiche - salvatore.spina@unict.it

[7]Ce.Di.Pa., Università degli Studi di Perugia - paola.venuti@unipg.it

ABSTRACT (ITALIANO)

In questo articolo presentiamo l’edizione digitale dei primi 16 taccuini dell’archeologo Paolo Orsi, conservati presso il Museo Archeologico Regionale “Paolo Orsi” di Siracusa. Il progetto è stato realizzato da un team interdisciplinare utilizzando la piattaforma Transkribus. Nell’articolo descriviamo brevemente la storia dei taccuini dell’archeologo, le attività di trascrizione e codifica delle pagine, segmentazione delle immagini, annotazione dei testi, pubblicazione dei risultati, dopo una sintesi dello stato dell’arte delle edizioni digitali di diari e manoscritti.

Parole chiave: edizioni digitali; Transkribus; manoscritti; archeologia

ABSTRACT (ENGLISH)

The Digital Edition of the First 16 Notebooks of Paolo Orsi.

In this article we present the digital edition of the first 16 notebooks of the archaeologist Paolo Orsi, preserved at the Regional Archaeological Museum “Paolo Orsi” in Syracuse. The project was carried out by an interdisciplinary team using the Transkribus platform. We briefly present the history of the archaeologist's notebooks, the activities of transcription and encoding of the pages, image segmentation, text annotation, and the publication of the results, following a summary of the current state of digital editions of diaries and manuscripts.

Keywords: digital scholarly editing; Transkribus; manuscripts; archaeology

1. INTRODUZIONE

In questo articolo presentiamo il progetto di creazione dell’edizione digitale dei primi 16 taccuini di scavo dell’archeologo Paolo Orsi. Gli originali dei taccuini sono conservati presso il Museo Archeologico Regionale “Paolo Orsi” di Siracusa, che fa parte del Parco Archeologico e Paesaggistico di Siracusa, Eloro, Villa del Tellaro e Akrai.

Il progetto dell’edizione digitale dei primi 16 taccuini è stato realizzato nell’autunno del 2023 e ha fatto parte di un progetto più ampio della Regione Siciliana denominato “Realizzazione spazio virtuale del Parco Archeologico di Siracusa” (PON Cultura e Sviluppo FESR 2014-2020).

Dopo l’analisi dei requisiti e la valutazione degli esempi e degli strumenti esistenti, abbiamo deciso di utilizzare la piattaforma Transkribus (https://www.transkribus.org/) sia per la correzione, l’allineamento e l’annotazione delle trascrizioni, sia per la pubblicazione online dell’edizione digitale, che si trova all’indirizzo https://app.transkribus.org/sites/taccuini-paolo-orsi. L’edizione digitale contiene il facsimile e la trascrizione allineata di circa 2000 pagine singole, accessibili tramite un sistema di navigazione, un motore di ricerca full text, e le annotazioni delle citazioni dei luoghi e delle date. Il lavoro è stato effettuato da un team interdisciplinare di 6 persone.

2. STATO DELL’ARTE DELLE EDIZIONI DIGITALI DI DIARI E LETTERE MANOSCRITTE

L’analisi dello stato dell’arte è stata effettuata nei mesi di settembre e ottobre 2023 e si riferisce quindi alla situazione delle edizioni digitali di manoscritti antecedenti a quella data. I 3 esempi principali che abbiamo analizzato sono stati i seguenti:

- "Bellini Digital Correspondence (BDC)", https://bellinicorrespondence.cnr.it/ : progetto di sviluppo e fruizione dell'edizione scientifica digitale delle lettere manoscritte del compositore Vincenzo Bellini conservate presso il Museo Civico Belliniano di Catania realizzato da ricercatori di CNR-ILC, CNR-ISTC, Università di Catania, in collaborazione con studenti dell'Università di Pisa, tra il 2017 e il 2022 (Del Grosso & Spampinato, 2023; Del Grosso et al., 2019). "Il carteggio comprende 40 unità testuali distrubuite su un alto numero di carte, riprodotte in oltre 100 immagini digitali" (Del Grosso & Spampinato, 2023). La piattaforma software open source utilizzata per la pubblicazione dell'edizione digitale è EVT - Edition Visualization Technology (https://github.com/evt-project/evt-viewer/), sviluppata a partire dal 2014 (Rosselli Del Turco, 2014; Cacioli et al., 2022).
- "Il quaderno di Paolo Bufalini", https://projects.dharc.unibo.it/bufalini-notebook/ : un'edizione scientifica digitale semantica del quaderno di Paolo Bufalini intitolato "Appunti 1981-1991", che si pone l'obiettivo di valorizzare la complessa rete di relazioni intratestuali ed intertestuali che caratterizza il quaderno. Realizzata dal Dipartimento di Filologia Classica e Italianistica (FICLIT) e del Digital Humanities Advanced Research Center (/DH.ARC) dell'Università di Bologna (D'Aquino et al., 2020). Il quaderno si compone di 145 pagine rilegate e due carte sciolte. L'applicazione web è stata sviluppata ad hoc e si basa su dati RDF archiviati in una database a grafo (https://projects.dharc.unibo.it/bufalini-notebook/specifications).
- il "Codice Pelavicino Edizione Digitale" (https://pelavicino.labcd.unipi.it/evt/index.html): edizione digitale creata dal 2014 al 2020 da un team interdisciplinare guidato da Enrica Salvatori come "un lavoro i corso", nel senso che "documenti, dati e note sono stati aggiunti mano a mano che procedeva nella trascrizione, annotazione, codifica e pubblicazione. Una parte rilevante del progetto ha previsto infatti – e ancora prevede in parte – la partecipazione attiva di studiosi e appassionati nel correggere interpretazioni e aggiungere dati alle note" (Salvatori et al. 2020). Piattaforma software open source utilizzata per la pubblicazione dell'edizione digitale: EVT - Edition Visualization Technology.

La letteratura scientifica rilevante per il progetto dei Taccuini di Paolo Orsi ha incluso i lavori di trascrizione e annotazione di manoscritti epigrafici (tra i quali Calvelli et al., 2023), oltre ovviamente alla letteratura recente relativa a metodi, strumenti ed esempi di edizioni scientifiche digitali (a partire da Pierazzo, 2015; Franzini et al., 2016). Un'analisi recente della situazione delle edizioni digitali di manoscritti in Italia si trova in (Senatore et al., 2024).

3. I TACCUINI DI PAOLO ORSI

Paolo Orsi, archeologo roveretano vissuto dal 1859 al 1935, responsabile della Soprintendenza e direttore del Museo Archeologico Nazionale di Siracusa dal 1891 al 1934, è unanimemente considerato uno dei padri fondatori dell'Archeologia in Italia, e il pioniere di una nuova metodologia che iniziò ad affermarsi in Italia proprio grazie a lui. Le sue meticolose indagini dei terreni, lo studio del rapporto tra gli strati archeologici e il metodo rigoroso, con cui condusse vasti scavi di necropoli e santuari, furono presi a modello da numerosi altri studiosi, e, ancora oggi, gli esperti sottolineano l'incidenza profonda che il lavoro di Orsi ebbe nello sviluppo della Scienza archeologica. Della sua attività di archeologo, che si concentrò in particolare in territorio siciliano e magnogreco, danno soprattutto testimonianza i taccuini, che Orsi ha redatto da poco prima del suo arrivo a Siracusa, fino al suo collocamento a riposo (1888-1934). La maggior parte di questi documenti rimase di proprietà della famiglia, finché, nel 1964, furono acquistati dalla Soprintendenza Archeologica di Siracusa e destinati ad essere conservati presso l'allora Museo Archeologico Nazionale di Siracusa - oggi Museo Archeologico Regionale "Paolo Orsi". Inizia quindi il lavoro per realizzare le prime modalità di fruizione dei preziosi documenti, che vennero microfilmati, in modo da renderne consultabili le copie presso la Soprintendenza. Nello stesso tempo si avviano le prime trascrizioni, la redazione di indici e si avvia l'iter finalizzato alla edizione cartacea dei taccuini. I contatti con l'Accademia dei Lincei (determinante il ruolo della Professoressa Pelagatti, che per anni ha curato i rapporti con l'Accademia) individuata come la sede più adatta alla loro pubblicazione (Orsi appartenne all'Accademia fin dal 1896, pubblicando gran parte del suo lavoro in "Monumenti antichi dei Lincei e Notizie scavi") furono avviati da Concetta Ciurcina, direttrice del Museo archeologico di Siracusa, e proseguiti da Bice Basile. Con Gioconda Lamagna e Giuseppina Monterosso si arriva alla edizione del primo volume nel 2018 (taccuini 1-4) dopo un iter molto complesso; Giuseppina Monterosso prosegue il lavoro con la pubblicazione del secondo volume (taccuini 5-16) nel 2022 (Orsi, 2018; Orsi, 2022).

Durante la sua attività di ricerca, Orsi redige 151 taccuini. Di questi, 137 rimasero in possesso degli eredi, ed entrarono a far parte del patrimonio del Museo siracusano nel 1964. Altri 11, mai posseduti dalla famiglia e depositati presso la Soprintendenza Archeologica di Reggio Calabria, furono restituiti nel 1985. I taccuini n. 41 e 49 risultano invece dispersi. Si tratta di piccoli quaderni di formato tascabile (le dimensioni sono di circa 150/160 mm. di altezza e poco meno di 100 di larghezza), con pagine a quadretti tenute insieme da una cucitura a macchina e con coperta di carta telata nera, rivestita di carta più leggera, nei vari toni del beige o del rosa.
I primi 16 taccuini coprono un arco cronologico che va dal 1888 agli inizi del 1893. In quel periodo, Orsi si occupò soprattutto degli scavi archeologici a Siracusa e Megara Hyblea, in Sicilia, e Locri Epizefiri, in Calabria. Nei taccuini vengono registrate le sue giornate di lavoro, i frequenti spostamenti da uno scavo e all'altro, descrivendo e disegnando i manufatti ritrovati, o quelli già conservati nei musei da lui visitati. Orsi realizza schizzi topografici delle aree di scavo, mette a confronto le informazioni ricavabili dalle fonti greche e latine con i suoi ritrovamenti, dà notizia di pubblicazioni coeve di interesse archeologico; non mancano appunti più coloriti riguardanti la sua quotidianità.

4. L'EDIZIONE DIGITALE IN TRANSKRIBUS

Per il progetto dell'edizione digitale dei primi 16 Taccuini di Paolo Orsi abbiamo scelto di utilizzare la piattaforma Transkribus (https://www.transkribus.org/): sviluppata a partire dal 2016, Transkribus è una piattaforma che consente la trascrizione e l'annotazione manuale collaborativa dei documenti (Kahle et al, 2017), il riconoscimento automatico delle grafie manoscritte e dei caratteri a stampa. Negli ultimi anni, inoltre, è stata aggiunta la possibilità di pubblicare online l'edizione digitale, creando il sito web per la consultazione e l'accesso alle informazioni (https://www.transkribus.org/sites), ed aggiornala in tempo reale.

Abbiamo utilizzato le fotografie delle pagine dei taccuini realizzate dai fotografi del Museo nel corso degli anni, evitando quindi una nuova digitalizzazione. Abbiamo effettuato un editing delle fotografie, creando una immagine per ogni pagina dei taccuini, in modo uniforme. Abbiamo utilizzato anche le trascrizioni della riproduzione anastatica, effettuandone una revisione completa. Abbiamo quindi proceduto alla segmentazione delle immagini, all'allineamento dei testi, al tagging manuale delle citazioni di luoghi e date. La codifica del testo è in formato TEI-XML ed è esportabile dalla piattaforma in vari formati. Contemporaneamente alla codifica dei testi, abbiamo costruito il sito web attraverso l'app "Transkribus Sites" (https://app.transkribus.org/sites/taccuini-paolo-orsi). Nella figura seguente (Fig. 1) si vede un esempio di pagina segmentata in regioni e linee a sinistra, e a destra la trascrizione e le annotazioni (pag 40 del Taccuino 4).

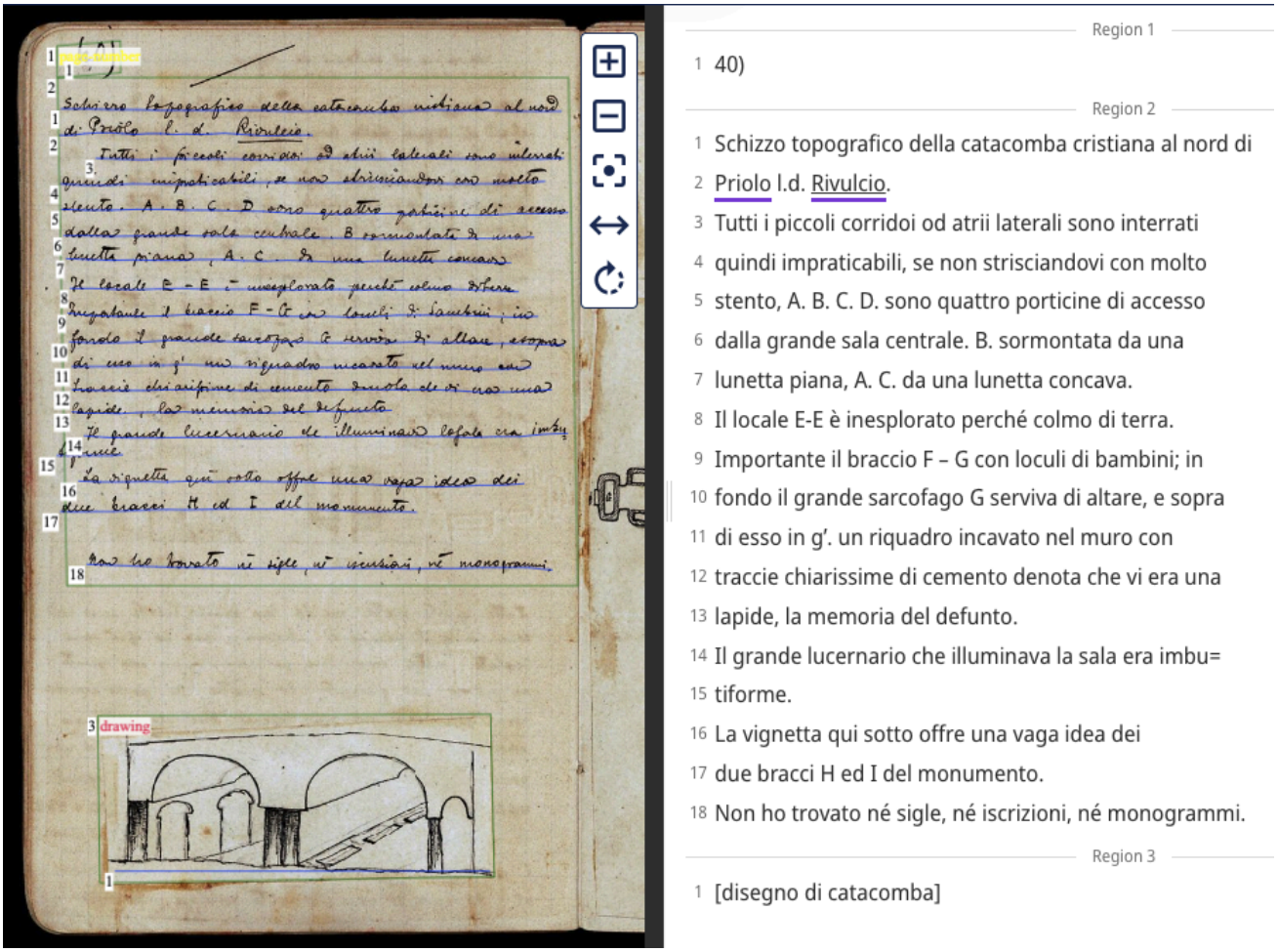


Figura 1. Esempio di pagina (pag 40 del Taccuino 4) segmentata in regioni e linee a sinistra, con a destra la trascrizione e le annotazioni, nella piattaforma Transkribus.

Il motivo principale della scelta della piattaforma Transkribus è stato la necessità di lavorare collaborativamente e velocemente alla trascrizione e codifica delle pagine, segmentazione delle immagini, annotazione dei testi e contemporaneamente alla pubblicazione dei risultati, per consentire la visione e il controllo da parte delle curatrici del museo e le responsabili del progetto. I dati annotati in modo standard potranno essere riutilizzati in altri sistemi di visualizzazione di TEI-XML ed eventualmente integrati o corretti.

5. SVILUPPI FUTURI

La creazione dell'edizione digitale dei primi 16 taccuini di Paolo Orsi vuole essere la prima fase di un progetto collaborativo più ampio di pubblicazione online di tutti i taccuini manoscritti dell'archeologo. L'esigenza di una edizione digitale è stata manifestata da molti ricercatori nel corso degli anni; tra questi citiamo ad es. Jonathan Prag in occasione della pubblicazione cartacea dei primi 4 taccuini: "Given the existence already of high quality digital scans, and the desirability, over time, of providing supporting information or annotation (rather than requiring any scholar who approaches the volumes to repeat the work of identification and bibliographic investigation which others may already have undertaken), a digital solution, perhaps using the increasingly widely adopted International Image Interoperability Framework, which would facilitate viewing, comparison and annotation of the pages, might ultimately be of greater value" (Prag, 2021).

Negli ultimi anni, la svolta digitale e le tecnologie di Intelligenza Artificiale hanno ampliato le frontiere della ricerca umanistica, che può avvalersi di tali tecnologie in diversi momenti della ricerca, quali l'analisi dei testi, la ricostruzione di network, la visualizzazione delle informazioni in mappe e timeline. Tra le varie tecnologie, quelle di Handwritten Text Recognition (HTR) rappresentano un'innovazione utile perché

consentono la trascrizione semi-automatica dei documenti d'archivio manoscritti, e facilitano quindi la realizzazione di edizioni digitali (Spina, 2023).
In questo progetto sui primi 16 taccuini di Paolo Orsi abbiamo provato ad addestrare un modello di HTR sulla scrittura di Paolo Orsi, che potrà essere perfezionato e quindi utilizzato nella trascrizione semi-automatica dei successivi taccuini.

RINGRAZIAMENTI

Il progetto dell'edizione digitale dei primi 16 taccuini ha fatto parte di un progetto più ampio della Regione Siciliana denominato "Realizzazione spazio virtuale del Parco Archeologico di Siracusa" (PON Cultura e Sviluppo FESR 2014-2020). Si ringrazia la Dott.ssa Rosa Lanteri e lo staff del Parco Archeologico e Paesaggistico di Siracusa, Eloro, Villa del Tellaro e Akrai.

BIBLIOGRAFIA

Cacioli, G., Cerretini, G., Chiara, D. P., Maenza, S., Rosselli Del Turco, R., & Zenzaro, S. (2022). There and back again: what to expect in the next evt version. *In AIUCD 2022-Culture digitali. Intersezioni: filosofia, arti, media. Preceedings della 11a conferenza nazionale*, Lecce, 2022 (pp. 212-217). Associazione per l'Informatica Umanistica e la Cultura Digitale.

Calvelli, L., Boschetti, F., & Tommasi, T. (2023). EpiSearch. Identifying Ancient Inscriptions in Epigraphic Manuscripts. *Journal of Data Mining & Digital Humanities*, (Sciences of Antiquity and digital humanities).

Codice Pelavicino. Edizione digitale, a cura di E. Salvatori, E. Riccardini, R. Rosselli del Turco, L. Balletto, C. Alzetta, C. Di Pietro, C. Mannari, R. Masotti, A. Miaschi, 2a ed., 2020 <http://pelavicino.labcd.unipi.it> ISBN 978-88-944430-2-8 DOI: https://doi.org/10.13131/978-88-944430-2-8

Daquino, M., Dello Buono, M., Giovannetti, F., & Tomasi, F. (2020). Paolo Bufalini, Appunti (1981-1991) [Edizione scientifica digitale semantica]. Digital Humanities Advanced Research Centre (/DH.arc), Bologna, https://doi.org/10.6092/unibo/amsacta/6415

Del Grosso, A. M., & Spampinato, D. (Eds.). 2023. *Bellini Digital Correspondence*. CNR Edizioni, 2023. ISBN: 978-88-8080-562-5.

Del Grosso, A. M., Capizzi, E., Cristofaro, S., De Luca, M. R., Giovannetti, E., Marchi, S., Seminara, G. & Spampinato, D. (2019, dicembre). Bellini's Correspondence: A Digital Scholarly Edition for a Multimedia Museum. *Umanistica Digitale*, 7, 23-47. http://doi.org/10.6092/issn.2532-8816/9162

Franzini, G., Mahony, S., and Terras, M. (2016), 'A Catalogue of Digital Editions', In: Pierazzo, E. and Driscoll, M. J. (eds) Digital Scholarly Editing: Theories and Practices. Cambridge: Open Book Publishers. DOI: 10.11647/OBP.0095.09

Kahle, P., Colutto, S., Hackl, G., & Mühlberger, G. (2017, November). Transkribus-a service platform for transcription, recognition and retrieval of historical documents. In *2017 14th iapr international conference on document analysis and recognition (icdar)* (Vol. 4, pp. 19-24). IEEE.

Orsi, P. (2018). I Taccuini, I. Riproduzione anastatica e trascrizione dei Taccuini 1-4. A cura di Gioconda Lamagna, Giuseppina Monterosso. *Monumenti Antichi Accademia Nazionale dei Lincei,* 75, Giorgio Bretschneider Editore, ISBN: 9788876892998

Orsi, P. (2022). I Taccuini, II. Riproduzione anastatica e trascrizione dei Taccuini 5-16. *Monumenti Antichi Accademia Nazionale dei Lincei,* 84, Giorgio Bretschneider Editore, ISBN: 9788876893407

Pierazzo, E. (2015). Digital Scholarly Editing: Theories, Models and Methods. Farnham, Surrey: Ashgate.

Prag, J. R. (2021). Paolo Orsi,«un modestissimo cultore della epigrafia»: Paolo Orsi's contribution to Sicilian epigraphy as illustrated by the newly published 'taccuini'. *Atti della Accademia Nazionale del Lincei. Rendiconti della Classe di Scienze Morali, Storiche, Filologiche*, 32(1-2)

Rosselli Del Turco, R., Buomprisco, G., Pietro, C. D., Kenny, J., Masotti, R., & Pugliese, J. (2014). Edition visualization technology: A simple tool to visualize TEI-based digital editions. *Journal of the Text encoding initiative*, (8).

Senatore, C., Del Turco, R. R., & Pisano, P. (2024). Digital Resources for Manuscripts: Between Fragmentation and Development Prospects. *Mimesis Journal*, 13(2), 104-118

Spina, S. (2023). Handwritten Text Recognition as a digital perspective of Archival Science. Aidainformazioni, (1-2), 115-132.

Artificial intelligence vs human handwriting: annotating damaged manuscripts

Dumitru Scutelnic[1], Laura Gazzani[2], Paolo Pellegrini[3], Claudia Daffara[4]
[1]Dept. of Computer Sciences, University of Verona, Italy - dumitru.scutelnic@univr.it
[2] Dept. of Computer Sciences, University of Verona, Italy - laura.gazzani@univr.it
[3] Dept. of Cultures and Civilizations, University of Verona, Italy - paolo.pellegrini@univr.it
[4]Dept. of Computer Sciences, University of Verona, Italy - claudia.daffara@univr.it

ABSTRACT (ENGLISH)

To decipher the text of damaged manuscripts using state-of-the-art AI methods such as Deep Learning, a training phase with an annotated dataset is required. We have developed an optimized dataset to assess the potential of AI in different book heritage applications, ranging from text recovery to diagnostic analysis, using high-resolution multispectral imaging. This dataset is continuously expanded with image stacks acquired on ongoing case studies; here, as exemplary cases, a damaged notebook by a 20th-century Italian writer and a manuscript by a 19th-century Italian intellectual. The annotation process must yield precise and detailed labels for the text and for each individual handwritten character, thereby providing AI models with suitable input for training. Unlike standard annotation approaches, which rely primarily on transcribing the text, we propose a different method based on instance segmentation of each character in the manuscript. Masks are traced to follow the exact shape of each character, each assigned to its own distinct class. Specific annotation criteria were defined in cross-disciplinary collaboration with philologists, physicists, and AI experts to maximize the system's potential for accurate handwriting recognition in degraded materials.
Keywords: multispectral imaging; damaged manuscripts; characters annotation; semantic segmentation, deep learning.

ABSTRACT (ITALIANO)

Intelligenza artificiale vs scrittura umana: annotare manoscritti danneggiati.
Per decifrare il testo di manoscritti danneggiati usando lo stato dell'arte dei metodi AI, come il Deep Learning, è necessaria una fase di training su dataset annotato. Abbiamo sviluppato un dataset ottimale per testare il potenziale dell'AI in varie applicazioni che riguardano il patrimonio librario, dal recupero del testo alla diagnostica, usando l'imaging multispettrale ad alta risoluzione. Il dataset viene continuamente ampliato con serie di immagini acquisite su casi studio; qui, come casi esemplari, un taccuino danneggiato di uno scrittore italiano del XX sec. e un manoscritto di un intellettuale italiano del XIX sec. Per fornire all'AI un modello adeguato, il processo di annotazione deve produrre "label" precise e dettagliate del testo e di ogni carattere della scrittura a mano. A differenza del normale metodo di annotazione, basato sulla trascrizione del testo, proponiamo un approccio diverso basato sulla segmentazione di istanze per ciascun carattere del manoscritto. Le maschere vengono tracciate seguendo ogni carattere, ognuno dei quali ha una sua classe distinta. I criteri specifici di annotazione sono stati determinati in collaborazione interdisciplinare, tra filologi, fisici, ed esperti di AI, in modo da massimizzare il potenziale per il riconoscimento della scrittura manoscritta anche in materiali degradati.
Parole chiave: imaging multispettrale; manoscritti danneggiati; annotazione dei caratteri; segmentazione semantica e di istanza, deep learning.

1. INTRODUCTION

Text recognition using artificial intelligence (AI) has evolved rapidly in recent years, thanks to the advancement of Deep Learning (DL) methods (Wang et al., 2021). The methods initially were based on network architectures such as Convolutional Neural Networks (CNNs) for recognizing single characters or words for different handwriting and alphabets (Liu et al., 2011). The introduction of Transformers (e.g., TrOCR and HTR-VT) (Minghao Li et al.,2021; Yuting Li, 2024), transformed traditional optical character recognition (OCR) into sophisticated systems capable of handling complex text formats such as a whole line of text. The solutions analyzed offer excellent performance when the text is clear and legible. However, when manuscript materials are degraded and are characterized by partially or completely illegible areas, such techniques show significant difficulties in character recognition. This has been verified

on our datasets using established OCR tools, such as Transkribus, eScritporium, and OCR4All. Undamaged manuscripts (such as Del Bene's case study) show better results when the text is still legible, the ink is uniform, on a uniform and clear background, and the contrast is high. Traditional OCR algorithms fail to provide a transcript when the ink is faded, and reflectance variations alter the contrast against an uneven background due to damaged materials. Consequently, there is a need for a robust strategy to separate ink from noise before text recognition. Similar studies have focused mainly on Asian languages (Maheswari, Maheswari & Aakaash, 2024; Shikha, Sonu & Vivek, 2020), but not as many have examined highly degraded manuscripts, the Latin alphabet or the Italian language. It is worth mentioning the Vesuvius Challenge concerning the decipherment of Greek text in the carbonized papyrus scrolls known as the Herculaneum papyri (Lourenco et al. 2023).

In this regard, the AIPAD Project (Artificial Intelligence and Physics for Art Diagnostics) focuses on the analysis and diagnostics of damaged historical manuscripts with non-invasive techniques and advanced data processing. Among the tasks are: 1) to exploit multispectral imaging in the UV-VIS-IR range to detect manuscript materials; 2) to annotate the multispectral dataset and train AI for handwritten text recognition on degraded manuscripts.

The dual nature of the manuscript – intangible text and physical support – is the key point. Optical imaging allows the different layers to be inspected, providing additional information (Perino et al., 2024; Tonazzini et al., 2019). It is known that UV imaging can reveal surface features of the manuscript (faded ink, erased text), while IR imaging can detect subsurface features (underwriting, carbon-based inks on blackened parchment). Recently, an integrated protocol with multiband imaging up to the mid-IR range has proven effective in the study of damaged manuscripts (Cimino et al., 2022).

In this study, we employ a multispectral imaging dataset acquired in the UV-VIS-NIR (up to 1 micron), spatially registered (Mazzocato et al., 2024). The dataset, called MADAM (Carcagnì et al., 2024) consists of 90 patches, each one a multispectral stack of 19 images. MADAM is continually populated with new images acquired on ongoing case studies; here, we consider a multispectral stack of a flood-damaged notebook of a 20th-century Italian writer, where the writing is degraded and not clearly visible to the naked eye. Multispectral imaging maximized the information in the 660 nm band, where the ink is more detectable, as shown in Fig. 1. An additional example includes a set of VIS images from a manuscript of Benedetto Del Bene (1749-1825), family annals, which are clearly legible (Fig. 2). They were included to enable predictive models to effectively learn and generalize character recognition. The challenges for handwritten character recognition are the diversity of writing styles, inconsistent spacing between letters and lines, and uncertainty about the number of lines and characters in a patch. Thus, we evaluated different strategies of annotation from simple transcription to more complex ones with masks that include from character level to the even non-uniform background.

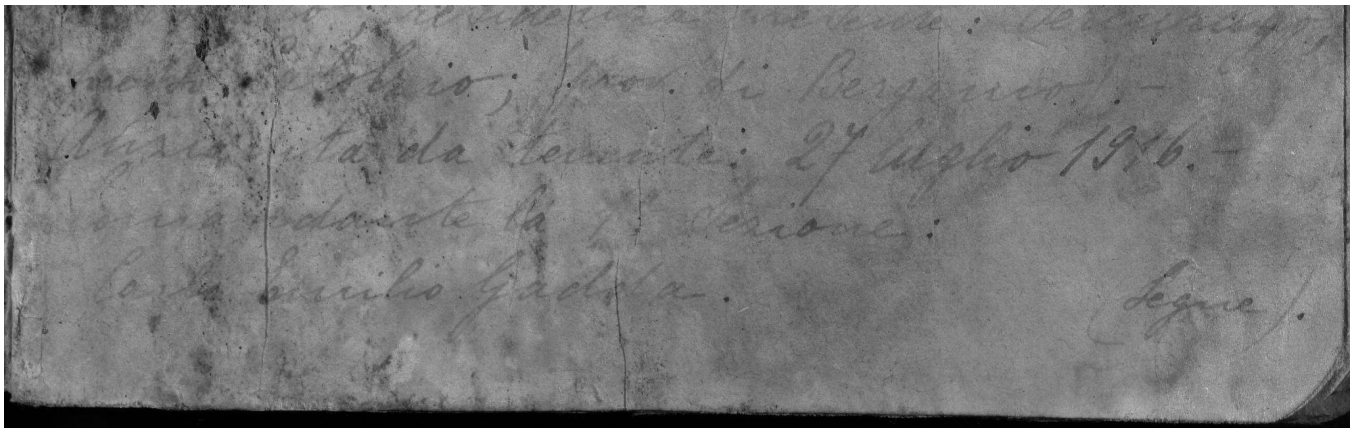


Figure 1: Image at 660 nm of an exemplary patch from the highly degraded manuscript.

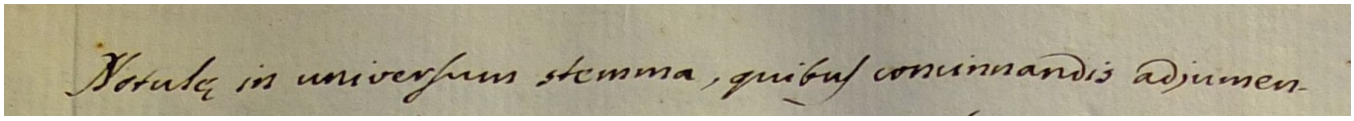


Figure 2: VIS image of an exemplary patch of Benedetto Del Bene's manuscript, page 1.

2. CRITERIA FOR ANNOTATION

Annotation must be as detailed and precise as possible to achieve accurate and scalable recognition results. Image binarization techniques are adopted to overcome problems of color gradients and different levels of ink transparency. In case of degraded materials, the key point is to consider only the traits of characters and exclude the background, i.e., the spurious signs of the damage, which cause OCR to fail. As annotation technique, we adopted the instance segmentation of each character, by creating masks that provide pixel by pixel the shape of the object, its position, and its class. This method replicates the exact shape of each character on the label by tracing it manually with a drawing pad. The definition of the classes to be annotated was done in synergy with philologists. It was decided to dedicate a separate class to every single character that can be found on the manuscript: one class for every letter (majuscule and minuscule), one for every number and one for every sign of punctuation (period, comma, hyphen, parentheses) – with an average of 70 classes per manuscript (Maheswari et al., 2024). There may be small variations from one manuscript to another, because annotation is language-specific (Maheswari et al., 2024). For example, the Del Bene manuscript (used for testing, written in Latin) presents the character "æ", which is absent in the other dataset. Any illegible characters or any characters that do not have a corresponding key are annotated as the "_" label. Such gaps will be filled by AI in the transcription, using a statistical approach based on the semantic content of the text, and possibly by the hypotheses of philologists, who would act as the ultimate quality check for the annotations (to fine-tune the algorithm) and the AI predictions and transcriptions. At the request of the philologists who are lending their expertise to this project, the annotation is carried out with particular attention to the ligatures between the letters to make the characters recognizable in any possible context. This is aimed at facilitating a successful recognition and transcription for an AI algorithm. This is also done to support any studies regarding the script itself, since it would be easier to pick and compare all the different ways of writing the same letter. Some examples of annotated patches can be found in Fig. 3.



(a)



(b)

Figure 3: VIS image of Del Bene manuscript (a) and image at 660 nm (b) of patches annotated by using the method of semantic segmentation with multiple classes (multiple colours of the characters).

This annotation procedure is very laborious and requires a lot of time and resources, however, the advantages are in scalability and possibility to use new DL techniques with more complex architectures. The annotation includes also bounding boxes that can be exploited for object detection, not fully explored in the context of text recognition. One aspect to consider concerns bounding boxes that may include parts of the background with signs of degradation that can confuse the DL model. Overcoming this problem would require huge datasets and resources for the AI training.

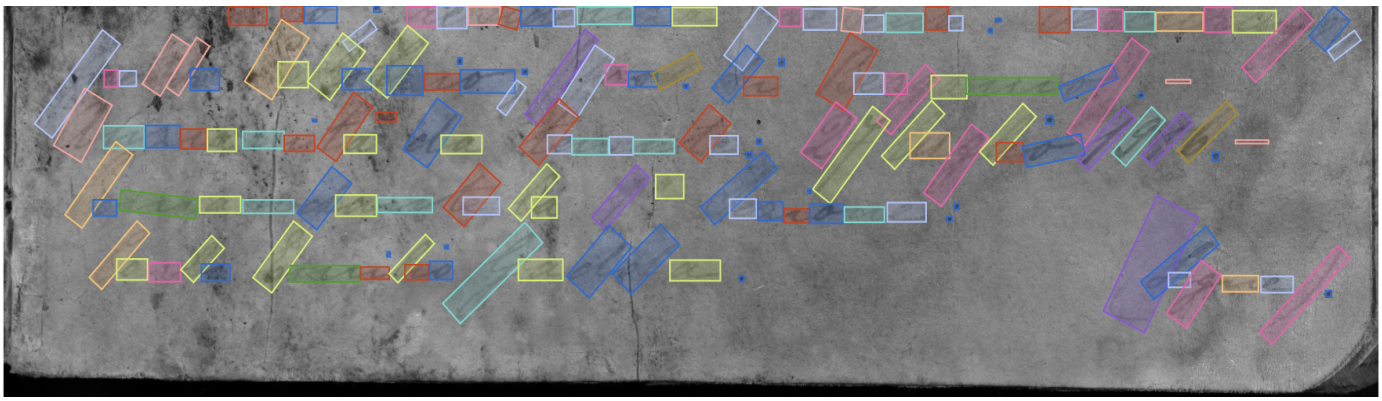


Figure 3: Image at 660 nm of the patch annotated by using the method of object detection with bounding boxes.

3. ARCHITECTURE AND METRICS

As DL models for instance segmentation, various architectures such as Mask R-CNN, U-Net, PointRend, HTC and DeepLabv3+, etc. (Chen et al., 2017; Singh et al., 2024), never used in text recognition, will be evaluated. While, for object detection, architectures such as Faster-RCNN, YOLO, Single Shot MultiBox Detector (SSD), RetinaNet, etc. (Srivastava et al., 2021), which are little explored for text recognition, will be evaluated. In addition, the procedure will be integrated with robust and established OCR methods such as Convolutional Recurrent Neural Network (CRNN) (Puarungroj 2020; Singh et al., 2024) with CTC Loss and Transformer OCR (e.g., TrOCR and HTR-VT) (Minghao Li et al.,2021; Yuting Li, 2024) to further improve text recognition.

Quality control is ensured by the expert philologists who accurately annotate each letter and background. While, the mean Average Precision (mAP), Intersection over Union (IoU), Recall, Precision and F1-score were used as metrics to estimate the learning performance of the models. For text recognition, Character Error Rate (CER), Word Error Rate (WER), Word Accuracy (correct words / total words) and CTC Loss (Yuting Li, 2024; Retsinas et al. 2024) will be used.

4. (EXPECTED) RESULTS

Fig. 4 represents the groundtruth obtained through the annotation of masks for each character. This is the expected result using the innovative approach of instance segmentation, where, in addition to the recognition of the text traced by the segmentation, its location is also identified together with its class. The approach aims to improve character recognition capability over available OCR tools, which do not perform well in the case of damaged manuscripts.

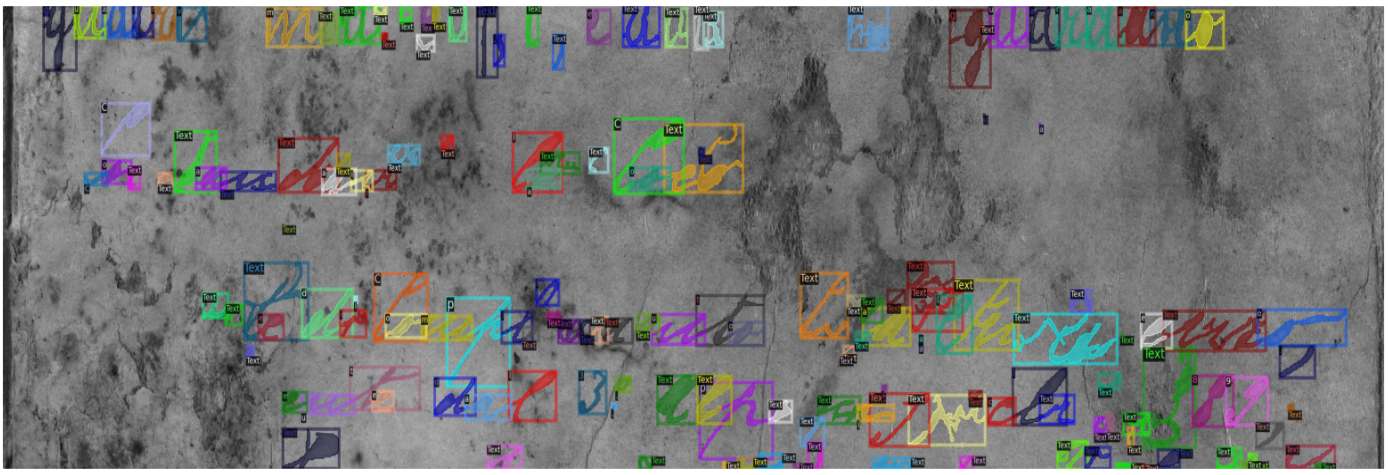


Figure 4: Annotations superimposed on a patch taken from the same page; the corresponding letters are indicated on any annotation, in the corner of the polygons.

5. FUTURE WORK

Future developments of the project include the expansion of the multispectral dataset with further historical damaged manuscripts from libraries in Verona. Another interesting direction for further research is the annotation of the background for diagnostic purposes by mapping the manuscript degradation features. Another task, which is underway, is the analysis and annotation of the physical aspects of the manuscripts that are intentional, such as watermarks and ruling. More expansions can also be directed towards new alphabets, such as the Cyrillic and the Greek alphabet.
In addition, the dataset will be made accessible to the community with impact in the field of text-assisted restoration using advanced optical and artificial intelligence techniques and in the development of new predictive methods.

ACKNOWLEDGEMENTS

This research is funded by the European Union – NextGenerationEU, M4C2 component, investment 1.1, PRIN PNRR 2022 Project: "Artificial Intelligence and Physics for Art Diagnostics (AIPAD): the killer application of the historical manuscripts and palimpsests", P2022SFEPN - CUP B53D23029240001 - Grant Assignment Decree No. 1372 (1/9/2023) by the Italian MUR.

REFERENCES

Yintong Wang, Wenjie Xiao and Shuo Li, Offline Handwritten Text Recognition Using Deep Learning: A Review, ICAACE 2021, Journal of Physics: Conference Series, 1848 (2021) 012015 https://doi.org/10.1088/1742-6596/1848/1/012015.

Minghao Li, Tengchao Lv, Jingye Chen, Lei Cui, Yijuan Lu, Dinei Florencio2, Cha Zhang, Zhoujun Li, Furu Wei, TrOCR: Transformer-based Optical Character Recognition with Pre-trained Models, Computation and Language, 2021, https://doi.org/10.48550/arXiv.2109.10282

Yuting Li, Dexiong Chenb, Tinglong Tanga, Xi Shenc, HTR-VT: Handwritten Text Recognition with Vision Transformer, Computer Vision and Pattern Recognition, 2024, https://doi.org/10.48550/arXiv.2409.08573

Maheswari, S.U., Maheswari, P.U. & Aakaash, G.R.S. (2024). An intelligent character segmentation system coupled with deep learning based recognition for the digitization of ancient Tamil palm leaf manuscritps. *Heritage Science 12*, 342. https://doi.org/10.1186/s40494-024-01438-4.

Shikha, C., Sonu, M., Vivek, S. (2020). Ancient text character recognition using deep learning. *International Journal of Engineering Research Technology, 13(9)*, 2177-2184. https://doi.org/10.37624/IJERT/13.9.2020.2177-2184.

Alex Lourenco, Brent Seales, Christy Chapman, Daniel Havir, Ian Janicki, JP Posma, Nat Friedman, Ryan Holbrook, Seth P., Stephen Parsons, and Will Cukierski. Vesuvius Challenge - Ink Detection. https://kaggle.com/competitions/vesuvius-challenge-ink-detection, 2023. Kaggle.

Michela Perino, Lucilla Pronti, Candida Moffa, Michela Rosellini and Anna Candida Felici, New Frontiers in the Digital Restoration of Hidden Texts in Manuscripts: A Review of the Technical Approaches, Heritage 2024, 7, 683–696. https://doi.org/10.3390/heritage7020034

Liu, C., Yin, F., Wang, D., & Wang, Q. (2011). CASIA Online and Offline Chinese Handwriting Databases. *International Conference on Document Analysis and Recognition*, 37-41. https://nlpr.ia.ac.cn/2011papers/gjhy/gh39.pdf.

Tonazzini, A., Salerno, E., Abdel-Salam, Z.A., Harith, M.A., Marras, L., Botto, A., Campanella, B., Legnaioli, S., Pagnotta, S., Poggialini, F. & Vincenzo Palleschi (2019). Analytical and mathematical methods for revealing hidden details in ancient manuscripts and paintings: A review. *Journal of Advanced Research*, 17, 31-42. https://doi.org/10.1016/j.jare.2019.01.003.

Cimino, D., Marchioro, G., De Paolis, P., Daffara, C. (2023). Evaluating the integration of Thermal Quasi-Reflectography in manuscript imaging diagnostic protocols to improve non-invasive materials investigation, *Journal of Cultural Heritage 62*, 72-77. https://doi.org/10.1016/j.culher.2023.04.009.

S. Mazzocato, D. Cimino, C. Daffara, Integrated microprofilometry and multispectral imaging for full-field analysis of ancient manuscripts, Journal of Cultural Heritage, Volume 66, 2024, Pages 110-116, ISSN 1296-2074, https://doi.org/10.1016/j.culher.2023.11.014.

Carcagnì, P., Del Coco, M., Leo, M., Gazzani, L., Malagodi, M., Paturzo, M., Daffara, C. (2024). MADAM: Manuscript Annotated Dataset based on Multispectral Imaging for Handwritten Text Enhancement and Restoration. Proc. of the International Conference Florence HeriTech, 2024, Firenze. (in press)

Chen, L., Papandreou, G., Kokkinos, I., Murphy, K. & Yuille, A.L. (2017). Deep Lab: Semantic Image Segmentations with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. https://arxiv.org/pdf/1606.00915.

Puarungroj, W., Boonsirisumpun, N., Kulna, P., Soontarawirat, T., Puarungroj, N. (2020). Using Deep Learning to Recognize Handwritten Thai Noi Characters in Ancient Palm Leaf Manuscripts. Ishita, E., Pang, N.L.S., Zhou, L. (Eds.) *Digital Libraries at Times of Massive Societal Transition. ICADL 2020. Lecture Notes in Computer Science*, Vol 12504. https://doi.org/10.1007/978-3-030-64452-9_20.

Srivastava, S., Divekar, A.V., Anilkumar, C. *et al.* Comparative analysis of deep learning image detection algorithms. *J Big Data* **8**, 66 (2021) https://doi.org/10.1186/s40537-021-00434-w

Sukhdeep Singh, Sudhir Rohilla, Anuj Sharma, An inclusive review on deep learning techniques and their scope in handwriting recognition, Computer Vision and Pattern Recognition, 2024, https://doi.org/10.48550/arXiv.2404.08011

George Retsinas, Giorgos Sfikas, Basilis Gatos, Christophoros Nikou, Best Practices for a Handwritten Text Recognition System, arXiv, 2024, https://arxiv.org/pdf/2404.11339v1

Modellazione, interoperabilità e riuso in DiScEPT

Tiziana Mancinelli[1], Hansmichael Hohenegger[2], Federico Boschetti[3], Angelo Mario Del Grosso[4], Eleonora De Longis[5], Gloria Mugelli[6]

[1] Istituto Italiano di Studi Germanici, Italia - mancinelli@studigermanici.it
[2] Istituto Italiano di Studi Germanici, Italia - hohenegger@studigermanici.it
[3] Istituto di Linguistica Computazionale CNR, Italia – federico.boschetti@unive.it
[4] Istituto di Linguistica Computazionale CNR, Italia - angelomario.delgrosso@cnr.it
[5] Istituto Italiano di Studi Germanici, Italia – delongis@studigermanici.it
[6] Istituto di Linguistica Computazionale CNR, Italia - gloria.mugelli@gmail.com

ABSTRACT (ITALIANO)

Il progetto DiScEPT (Digital Scholarly Editions and Parallel Translations Platform) ha l'obiettivo di creare un ambiente per la produzione e la pubblicazione di edizioni digitali. Il multilinguismo è il punctum saliens di questo progetto per una scelta scientifica, ma anche perché si propone di favorire l'accessibilità, l'inclusione e il dialogo interculturale almeno nell'ambito della comunità inerente alla filologia digitale nonché della traduttologia. Le traduzioni non sono solo il superamento delle barriere linguistiche, ma strumenti di arricchimento critico, opportunità per lo studio semantico, storico e culturale dei testi. Integrando strumenti avanzati, come il riconoscimento automatico dei caratteri (HTR e OCR) e l'allineamento linguistico basato su BERT-Align, DiScEPT propone flussi di lavoro dinamici per la creazione di edizioni digitali con traduzioni allineate. La metodologia si fonda su modelli multipli che considerano la materialità dei documenti, la critica testuale e soprattutto l'analisi dei corpora paralleli. Superando l'idea di una piattaforma monoblocco, il progetto adotta un ambiente distribuito, basato sull'integrazione di piattaforme esistenti tramite API.

Parole chiave: edizione scientifica digitale; TEI-XML; Allineamento di corpora; Linked Open Data.

ABSTRACT (ENGLISH)

Modeling, interoperability, and reuse in DiScEPT. The DiScEPT (Digital Scholarly Editions and Parallel Translations Platform) project aims to create an environment for the production and publication of digital editions. Multilingualism is the key focus of this project, not only as a scientific choice but also to promote accessibility, inclusion, and intercultural dialogue, at least within the community of publishers and translators. Translations are not merely a way to overcome language barriers but are critical tools for enriching textual analysis, offering opportunities for semantic, historical, and cultural studies. By integrating advanced tools such as Handwritten Text Recognition (HTR, OCR) and linguistic alignment using BERT-Align, DiScEPT offers dynamic workflows for creating digital editions with aligned translations. The methodology is based on multiple models that consider document materiality, textual criticism, and, particularly, the analysis of parallel corpora. Moving beyond the idea of a monolithic platform, the project adopts a distributed environment, relying on the integration of existing platforms through APIs.

Keywords: digital scholarly editing; linked open data; corpora alignment, TEI-XML;

1. INTRODUZIONE

Il progetto DiScEPT parte dall'esigenza di creare edizioni scientifiche digitali multilingue in grado di mettere in relazione testi originali e loro traduzioni già edite attraverso l'integrazione di strumenti e pratiche digitali già consolidate, limitando le aggiunte di componenti soltanto ai casi in cui ci siano specifiche esigenze di studio. L'obiettivo principale è costruire un ecosistema flessibile e aperto che consenta di analizzare e rappresentare le connessioni linguistiche e culturali, migliorando così la qualità editoriale e promuovendo una maggiore consapevolezza delle dinamiche traduttive.

Il focus sul multilinguismo è un elemento distintivo di DiScEPT, che non solo affronta la necessità di superare le barriere linguistiche, ma enfatizza anche il valore critico delle traduzioni e di trasmissione della conoscenza. Le traduzioni non vengono considerate come semplici strumenti per comunicare un contenuto, ma come veri e propri strumenti di arricchimento culturale, di diffusione del sapere e analisi comparativa. Questo approccio inoltre favorisce una comprensione più profonda delle dinamiche linguistiche e delle scelte traduttive, offrendo nuove opportunità per la ricerca e l'editoria scientifica.

Il progetto rappresenta anche uno spazio per ripensare i flussi di lavoro e i modelli editoriali, considerando i diversi formati e materiali che possono essere coinvolti, dalle immagini facsimilari ai testi strutturati come XML-TEI. DiScEPT si distingue per il suo approccio centrato sul riuso e sull'interoperabilità, evitando la creazione di silos informatici e incoraggiando la condivisione e la sostenibilità dei dati.

Un elemento chiave è l'utilizzo di tecnologie avanzate come il riconoscimento automatico dei caratteri (HTR) tramite strumenti come eScriptorium e l'allineamento linguistico automatizzato con BERT-Align. Questi strumenti permettono di confrontare testi e traduzioni, evidenziando scelte semantiche, storiche e culturali sia tramite analisi da parte dell'utente/editor, sia attraverso strumenti automatici di allineamento linguistico come BERT-Align, che supportano l'individuazione delle corrispondenze testuali.
DiScEPT non è solo un progetto tecnologico, ma anche un'iniziativa interdisciplinare che valorizza il multilinguismo per ampliare l'accesso a fonti e dati. Inoltre, contribuisce alla conservazione del patrimonio culturale globale, traducendo e interpretando testi in diverse lingue, e promuove l'uso di tecnologie linguistiche avanzate per affrontare sfide complesse.

2. MODELLI E WORKFLOW PER LE EDIZIONI DIGITALI: UN LABORATORIO DI EDIZIONI MULTILINGUE.

DiScEPT è progettato per consolidare metodologie e pratiche derivate da diversi progetti nel campo dell'ecdotica digitale e dei corpora multilingue allineati. L'obiettivo principale è proporre e integrare strumenti e modelli già affermati nella filologia digitale in un ecosistema che copra tutte le fasi del processo editoriale, dalla produzione alla diffusione.
Il progetto sviluppa diversi tipi di workflow, adattandosi al materiale di partenza: il lavoro può iniziare da immagini facsimilari, testi digitali in XML-TEI o trascrizioni critiche. Nel caso in cui il flusso di lavoro parta da immagini facsimilari prive di trascrizione testuale, vengono impiegate tecnologie avanzate di riconoscimento automatico dei caratteri (HTR), come eScriptorium, garantendo trascrizioni digitali precise e coerenti.
Infine, DiScEPT si distingue per tre modelli principali di descrizione delle edizioni:

- Materialità del documento: Analisi e valorizzazione del documento, della mise en page e della trascrizione del testo.
- Tipologie di edizioni e critica testuale: modello di edizione di ogni singolo documento, tenendo presente la tradizione del testo e le metodologie provenienti dalla filologia e dall'ecdotica.
- Modello linguistico per corpora paralleli: Creazione di corpora per studi comparativi, favorendo analisi approfondite delle dinamiche linguistiche e culturali.

Un focus particolare è dedicato alla materialità del documento, attraverso la possibilità di includere, rappresentare e analizzare immagini digitali facsimilari. La trascrizione non ha il solo fine di digitalizzare il testo, ma viene arricchita con annotazioni e dettagli sulla struttura fisica del testo come sezioni del testo, layout della pagina e anomalie grafiche. Le immagini, elaborate tramite tecnologie HTR, diventano quindi parte integrante del workflow, garantendo che il legame tra contenuto e forma sia preservato e valorizzato in tutte le fasi del processo editoriale. La peculiarità e la sfida di DiScEPT consistono nel collegare i diversi modelli di rappresentazione testuale (materialità, critica testuale, corpora paralleli) e, allo stesso tempo, nel renderli formalmente espliciti e interoperabili, garantendo la tracciabilità tra struttura editoriale, la relazione fra testi originali e traduzioni, contenuto e metadati.

Un altro elemento distintivo di DiScEPT è il sistema di allineamento linguistico, che combina metodi manuali e automatici per analizzare le relazioni tra testi originali e traduzioni. Basato su tecnologie come BERT-Align, l'allineamento automatico individua corrispondenze tra segmenti testuali, evidenziando scelte semantiche, stilistiche e culturali. Quando necessario, nei casi più complessi, è previsto un intervento manuale per affinare i risultati.
L'allineamento considera diverse configurazioni relazionali tra i segmenti di testo. Nel caso di uno a uno, ogni segmento del testo originale corrisponde direttamente a uno nella traduzione, comune nei testi tecnici o strutturati. La configurazione molti a uno si verifica quando più segmenti originali confluiscono in uno tradotto, tipico nelle traduzioni sintetiche o nei testi semplificati. Al contrario, la relazione molti a molti rappresenta un livello di complessità maggiore, in cui più segmenti originali e tradotti sono interconnessi, riflettendo riformulazioni stilistiche, aggiunte o omissioni.

Questo consente di tracciare le dinamiche storiche e culturali che influenzano il processo traduttivo, offrendo una comprensione più profonda delle trasformazioni testuali. La combinazione di strumenti automatici e interventi manuali migliora non solo la qualità dell'allineamento, ma anche l'analisi comparativa e gli studi filologici avanzati.
Grazie a questo approccio modulare e integrato, DiScEPT promuove un ambiente di ricerca multidisciplinare, valorizzando la collaborazione internazionale e l'accesso inclusivo alle risorse digitali.

3. CHE COS'È DISCEPT E LA SUA EVOLUZIONE

DiScEPT è un ecosistema per la creazione di edizioni digitali basato sull'integrazione di progetti open source e tecnologie aperte, in grado di gestire dati e metadati complessi. Questo processo ha permesso di sviluppare un'architettura modulare composta da diversi componenti:

- Editor testuale basato su DSL (Domain-specific language): Facilita la codifica tramite convenzioni familiari ai filologi, convertibili in XML-TEI attraverso l'uso di un parser ANTLR e XSLT. Può ricevere i dati della trascrizione tramite API e permetterne così una codifica ulteriore (per esempio per la costruzione di un apparato critico);
- Dialogo fra software HTR e sistema di editing basato su DSL;
- Allineamento testuale: Utilizza metodologie stand-off per creare corrispondenze tra testi originali e traduzioni a diversi livelli di granularità, promuovendo lettura sinottica e analisi linguistica. Allineamento di corpora multilingue automatico basato su tecnologie come Bert-Align.

Pubblicazione e visualizzazione: Basata su TEI-Publisher e tecnologie come ODD e WebComponent, offre funzionalità end-to-end dalla creazione alla pubblicazione delle edizioni digitali, garantendo riuso e interoperabilità con comunità come e-editiones.

4. INTEROPERABILITÀ TRA PIATTAFORME

Posizionare Facendo tesoro del concetto di Virtual Research Environment (VRE) distribuito, DiScEPT abbandona l'idea ambiziosa di nuova piattaforma (o, per meglio dire, "just another platform") integrata per le edizioni scientifiche digitali e abbraccia invece il paradigma di integrazione tra piattaforme esistenti e comunicanti tramite API.
Il modello di VRE distribuito è stato teorizzato nei primi lustri degli anni Duemila all'interno di progetti pioneristici come Bamboo (https://academic.oup.com/dsh/article/29/3/326/2938127), che purtroppo "failed to move to the implementation phase" (https://digital.humanities.ox.ac.uk/project/project-bamboo), in parte per motivi contingenti e in parte perché troppo avanzato rispetto ai limiti tecnologici dell'epoca, sia hardware, per le connessioni di rete, sia software, per gli inefficienti protocolli di comunicazione.
L'evoluzione delle tecnologie di comunicazione macchina a macchina senza l'intermediazione di interfacce utente, ci permette oggi di poter gestire in modo molto più efficiente flussi di dati tra piattaforme altamente specializzate e fra loro indipendenti.

È il caso, in DiScEPT, della connessione fra eScriptorium, piattaforma sviluppata da INRIA (https://gitlab.inria.fr/scripta/escriptorium) per l'Handwritten Text Recognition (HTR) e TEIPublisher (https://teipublisher.com), un toolbox di instant publishing di documenti XML-TEI per la piattaforma eXist-DB (https://exist-db.org).
Benché in DiScEPT sia prevista la comunicazione bidirezionale, attualmente è stata sviluppata la parte di comunicazione da eScriptorium a eXist-DB, per trasferire tramite API i dati testuali relativi alle trascrizioni (semi)automatiche dall'ambiente di editing focalizzato sulla correzione degli errori di HTR all'ambiente di editing sviluppato appositamente per DiScEPT su eXist-DB. Tale ambiente è basato su Domain-Specific Languages (DSL) con esportazione in XML-TEI per rendere i contenuti finali fruibili tramite TEIPublisher (la demo è in corso di sviluppo; quando sarà pronta il codice sorgente si troverà all'indirizzo: https://github.com/CoPhi/cophi-discept).

Il componente riguardante l'allineamento dei testi in DiScEPT (https://istituto-italiano-di-studi-germanici.github.io/DiScEPT/) si distingue per due approcci complementari: manuale e automatico, ciascuno mirato a garantire la massima precisione nell'identificazione delle relazioni tra testo originale e traduzioni.
Nel caso dell'allineamento manuale, il flusso di lavoro inizia con una fase di tokenizzazione, ancora in fase di perfezionamento, che suddivide il testo in unità linguistiche fondamentali. L'editor offre un'interfaccia utente intuitiva, che consente per chi la utilizza di selezionare manualmente segmenti di testo e abbinarli alle traduzioni corrispondenti. Un aspetto importante di questo processo è la possibilità di espandere il modello di riferimento in TEI, in modo da includere anche RDF, per migliorare la capacità di citazione e l'interoperabilità tra risorse digitali.
Per quanto riguarda l'allineamento automatico, DiScEPT si avvale di modelli linguistici avanzati, come Bert-Align, applicato specificamente alle lingue italiana e tedesca. Questo sistema è in fase di sviluppo e mira a sfruttare l'intelligenza artificiale per riconoscere e allineare automaticamente i segmenti di testo tra versione originale e traduzione.

Il modello Bert-Align, basato su reti neurali, consente un allineamento molto affidabile, con potenziali applicazioni per una vasta gamma di lingue. Alla luce delle riflessioni proposte da Hämmerl, Libovický e Fraser (2024), l'integrazione di modelli come BERT-Align si colloca all'interno di un quadro teorico in cui l'allineamento cross-linguistico viene inteso non solo come corrispondenza formale, ma come costruzione di un ponte semantico tra sistemi linguistici diversi. In questa prospettiva si inseriscono anche gli sviluppi proposti da Reimers e Gurevych (2020), che mostrano come modelli monolingui per il confronto semantico di frasi possano essere adattati con successo a contesti multilingue, attraverso strategie che permettono di trasferire le capacità apprese da una lingua all'altra. A queste soluzioni si affiancano strumenti recenti che rendono disponibili, in modo modulare, diversi algoritmi per il confronto tra segmenti testuali, permettendo di combinare in modo flessibile la somiglianza semantica con il calcolo della distanza formale. La capacità di bilanciare informazioni condivise e specificità linguistiche, emersa come centrale in tale studio, rappresenta una direzione promettente per lo sviluppo futuro di DiScEPT, ad esempio attraverso il fine-tuning controllato per domini o generi testuali differenti.
La capacità di bilanciare informazioni condivise e specificità linguistiche, emersa come centrale in tale studio, rappresenta una direzione promettente per lo sviluppo futuro di DiScEPT, ad esempio attraverso il fine-tuning controllato per domini o generi testuali differenti. L'allineamento, sia automatico che manuale, genera un unico file TEI in cui i testi sono collegati tramite un markup stand-off, strutturato all'interno del tag <tei:corpus>. I due approcci sono complementari e flessibili, consentendo agli utenti di scegliere il metodo più adatto al materiale da analizzare. La loro combinazione unisce la precisione manuale all'efficienza automatica, migliorando l'accuratezza e la coerenza nell'analisi dei testi tradotti. Questo sistema promuove una comprensione più approfondita delle scelte traduttive e delle dinamiche linguistiche.



Figura 1. Flusso di lavoro

Il progetto DiScEPT dunque rappresenta un passo avanti nella creazione di edizioni digitali multilingue, l'attenzione sul multilinguismo non è solo un valore scientifico, ma anche un obiettivo per favorire l'accesso globale ai testi e promuovere l'inclusione interculturale e di diffusione del patrimonio culturale.

RINGRAZIAMENTI

L'attuale prototipo di DiScEPT si avvale di servizi offerti da H2IOSC Project - Humanities and cultural Heritage Italian Open Science Cloud funded by the European Union – NextGenerationEU – NRRP M4C2 - Project code IR0000029 - CUP B63C22000730005.
ILC4CLARIN, nodo B di CLARIN-IT, ospiterà i dati e il software del progetto per la long-term preservation. Il CLARIN Knowledge Centre DiPText-KC mette inoltre a disposizione del progetto DiScEPT il proprio Helpdesk (https://diptext-kc.clarin-it.it/helpdesk).

BIBLIOGRAFIA

Boschetti, F. (2008). Alignment of Variant Readings for Linkage of Multiple Annotations. *Proceedings of the ECAL 2007 Electronic Corpora of Ancient Languages*, Prague 16–17 November 2007. Zemánek, P. (Ed.). 11-24. http://usj.ff.cuni.cz/system/files/Boschetti-Ch-2007.pdf.

Boschetti, Federico; Taddei Andrea, Bambaci, Luigi; Del Grosso Angelo Mario; Mugelli, Gloria; Khan, Fahad; Bellandi, Andrea. "Collaborative and Multidisciplinary Annotations of Ancient Texts: the Euporia System". In The Ancient World Goes Digital Case Studies on Archaeology, Texts, Online Publishing, Digital Archiving, and Preservation, Edited by: Bigot Juloux, Vanessa; di Ludovico, Alessandro; Matskevich, Sveta. Leiden - Boston: Brill, 2023

Boschetti, Federico; Rigobianco, Luca; Quochi, Valeria. "Domain-Specific Languages for Epigraphy: The Case of ItAnt". In CLARIN Annual Conference Proceedings. (2023): pp. 80-84

Buzzetti, D., & McGann, J. (2006). *Electronic Textual Editing: Critical Editing in a Digital Horizon*. *Electronic Textual Editing*. Burnard L. et al. (Eds.), Modern Language Association of America. http://www.tei-c.org/About/Archive_new/ETE/Preview/mcgann.xml.

Cushman, Ellen. "Supporting Manuscript Translation in Library and Archival Collections: Toward Decolonial Translation Methods". In Libraries and Archives in the Digital Age, edited by Susan L. Mizruchi. London: Palgrave Macmillan, 2020.

Daquino, Marilena; Giovannetti, Francesca; Tomasi, Francesca. "Linked Data Per Le Edizioni Scientifiche Digitali. Il Workflow Di Pubblicazione dell'edizione Semantica Del Quaderno Di Appunti Di Paolo Bufalini". In *Umanistica Digitale* 3 (7). https://doi.org/10.6092/issn.2532-8816/9091

Del Grosso, A. M., Capizzi, E., Cristofaro, S., Seminara, G., & Spampinato, D. (2019). *Promoting Bellini's Legacy and the Italian Opera by Scholarly Digital Editing His Own Correspondence*. Poster presented at the TEI Conference and Member's Meeting. What is Text, really? TEI and beyond. Graz, Austria. https://zenodo.org/badge/DOI/10.5281/zenodo.3461673.svg.

Del Grosso, A. M., & Spampinato, D. (Eds.). 2023. *Bellini Digital Correspondence*. CNR Edizioni, 2023. ISBN: 978-88-8080-562-5.

Dombrowski, Q. (2014). What Ever Happened to Project Bamboo? *Literary and Linguistic Computing*, 29(3), 326–339. https://doi.org/10.1093/llc/fqu026

Hämmerl, Johanna; Libovický, Jindřich; Fraser, Alexander. "Understanding Cross-Lingual Alignment—A Survey." *Findings of the Association for Computational Linguistics*: ACL 2024. https://aclanthology.org/2024.findings-acl.649

Martignano, Chiara. 2021. "A Conceptual Model to Encourage the Development and Reuse of Apps for Digital Editions". *Umanistica Digitale* 5 (10):71-88. https://doi.org/10.6092/issn.2532-8816/12620.

Pierazzo, E. (2015). Digital Scholarly Editing: Theories, Models and Methods. Farnham, Surrey:Ashgate. Pozzo, Riccardo; Gatta Timon; Hohenegger, Hansmichael; Kuhn, Jonas; Pichler, Axel; Turchi, Marco and

Genabith, Josef van. "Aligning Immanuel Kant's Work and its Translations". In *CLARIN: The*

Infrastructure for Language Resources. Edited by Fišer, D. and Witt, A. Berlin, Boston: De Gruyter, 2022, pp. 727-746. https://doi.org/10.1515/9783110767377-029

Reimers, Nils, and Iryna Gurevych. 2020. "Making Monolingual Sentence Embeddings Multilingual Using Knowledge Distillation." In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics*. https://arxiv.org/abs/2004.09813.

Viglianti, Raffaele. Why TEI Stand-off Markup Authoring Needs Simplification. Journal of the Text Encoding Initiative vol. 10 (2019). https://journals.openedition.org/jtei/1838. doi:https://doi.org/10.4000/jtei.1838.

Spadini, Elena and Turska, Magdalena. XML-TEI Stand-off Markup: One Step Beyond. Digital Philology: *A Journal of Medieval Cultures* vol. 8 (2019) pp. 225-239. doi:https://doi.org/10.1353/dph.2019.0025.

Suzgun, Mirac, Stuart M. Shieber, and Dan Jurafsky. 2023. "String2string: A Modern Python Library for String-to-String Algorithms." https://arxiv.org/abs/2304.14395.

Wu, Shuo, Xiaodan Yin, and Yue Zhang. (2023). "Improving Sentence-Level Alignment with Contrastive Learning in Multilingual Contexts." In *Proceedings of the 2023 Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics*.

Navigating the Digital Transition: Lessons from a Hybrid Critical Edition Project

Elisa Bastianello[1], Reto Baumgartner[2]
[1] Bibliotheca Hertziana – Max Planck Institute for Art History, Italy/Germany – elisa.bastianello@biblhertz.it
[2] University of Zurich, Switzerland - reto.baumgartner@zi.uzh.ch

ABSTRACT (ENGLISH)

This article explores the development and publication of a digital critical edition of the complete works of Heinrich Wölfflin, resulting from a collaborative project between the University of Zurich and the Bibliotheca Hertziana – Max Planck Institute for Art History. The project, started as a traditional print edition, aimed to create a digital edition not only of the original volumes but also of the newly published critical volumes, enriched with annotations and facsimile images. The article details the hybrid approach of publishing printed volumes first, followed by their digital versions, and the challenges faced in converting traditional transcription methods into structured digital formats. It highlights the use of legacy tools like Visual Basic for Applications and more modern XSLT transformations to streamline the process. The publication phase emphasizes strategies for ensuring long-term maintenance of the digital edition, including the adoption of open-source platforms and institutional support. The article concludes by reflecting on the collaborative efforts and diverse expertise that contributed to the project's success, offering insights for other teams embarking on similar digital publishing endeavors.
Keywords: Digital Critical Edition; Text Annotation; TEI XML; Open-Source Platforms; Digital Humanities

ABSTRACT (ITALIANO)

Navigare la transizione digitale: lezioni da un progetto di edizione critica ibrida

Questo articolo esplora lo sviluppo e la pubblicazione dell'edizione critica digitale delle opere di Heinrich Wölfflin, risultato di un progetto collaborativo tra l'Università di Zurigo e la Bibliotheca Hertziana – Istituto Max Planck per la storia dell'arte. Il progetto, nato come una edizione critica tradizionale a stampa, mirava a creare un'edizione digitale non tanto dei volumi originali ma di nuovi volumi critici stampati, arricchiti con annotazioni e immagini facsimili. L'articolo descrive l'approccio ibrido nella pubblicazione prima dei volumi stampati, seguiti dalle loro versioni digitali, e le sfide affrontate nel convertire i metodi di trascrizione tradizionali in formati digitali strutturati. Sottolinea l'uso di strumenti del passato come Visual Basic for Applications e più moderni come le trasformazioni XSLT per ottimizzare il processo. La fase di pubblicazione enfatizza le strategie che sono state messe in atto per garantire la manutenzione a lungo termine dell'edizione digitale, inclusa l'adozione di piattaforme open-source e il supporto istituzionale. L'articolo si conclude riflettendo sugli sforzi collaborativi e le diverse competenze che hanno contribuito al successo del progetto, offrendo spunti per altri team che intraprendono iniziative di pubblicazione digitale simili.
Parole chiave: edizione critica digitale; annotazione testuale; TEI XML; piattaforme open source; Digital Humanities

1. INTRODUCTION

In November 2024, following years of meticulous work, the digital critical edition of Heinrich Wölfflin's collected works was formally launched online.[1] This moment presents an opportune occasion to reflect on the journey that culminated in the digital publication of the first two volumes in this series. The project for the complete works of the Swiss art historian represents a collaboration between the University of Zurich and the Bibliotheca Hertziana - Max Planck Institute for Art History in Rome, under the supervision of Tristan Weddigen, Oskar Bätschmann, and Joris van Gastel.

The edition project, initially funded by the Swiss National Science Foundation (SNSF),[2] with additional support by the Bibliotheca Hertziana,[3] evolved beyond its original conception as a traditional printed edition when the first volumes had already reached the typesetting stage. Consequently, the digital edition was developed not merely as a parallel version of the new printed books but rather as a synoptic view of both (and potentially multiple) editions of the original volumes. Beyond the new commentaries and critical

[1] The edition is available at *https://hwgw.humanitiesconnect.pub*.
[2] The SNSF funding began as a research project (grant n. *160081*) from 2015 to 2017 and continued as an editorial project from 2017 to 2024 (grants nn. *157979* and *198243*).
[3] Grant id *BH-P-19-32*, since 2017.

apparatus, the digital edition has been enhanced with an annotation layer that integrates bibliography and named entity linking, alongside IIIF facsimile images of the *editio princeps*. This comprehensive approach ensures that full text search capabilities and index apparatuses are now seamlessly integrated across all volumes.
This article presents the team's experience throughout all stages of creating the digital edition, including challenges and missteps encountered along the way. Our aim is to provide other groups with insights into our process, enabling them to adopt and adapt the aspects that might prove useful for their own digital humanities projects.

2. PREPARATION OF THE EDITION

The first note goes to the fact that this is a hybrid critical edition, where, at least so far, the volumes are first published in traditional printed form and later prepared for the online edition. The edition was grafted onto a long path of traditional transcription with OCR[4] of Heinrich Wölfflin's printed sources that had been manually corrected in the form of MS Word documents. The idea of working with native TEI XML[5] was discarded at the beginning of the project because it was deemed too complex by the team. This choice took into account both the curators' greater familiarity with the editor and the manual workflow used within the publishing house.
The editors, in collaboration with the typesetters, established a list of editorial rules to mark layout and philological information in the text files. These included using double underlining or highlighters to denote different character styles (such as spaced text – Sperrsatz in German), images, or text requiring manual editing during layout. As for the philological notes, including page breaks, location and captions of images, author's pencil notes, and factual errors in the text were indicated by curly brackets. Critical notes were inserted as endnotes, while the author's notes were converted into regular footnotes.
To overcome the limitation of the text editor, which does not allow automatic notes attached to the text of other notes, the critical annotations related to the footnotes of the original text had been supplemented with the use of curly brackets. In many cases, the co-presence of philological information and critical notes led to the nesting of curly brackets. Additionally, while the fonts and sizes of individual elements (e.g., body of the text, footnotes, and citations) had been defined in the editorial rules, they had not been uniformly and correctly applied with character and paragraph styles by the editors, possibly due to the copy-and-paste function that retains the source text style. For these reasons, the files, visually perfect on screen, contained issues such as endnotes occasionally using the headings paragraph style, or titles being just normal text in bold.
The print output, though requiring time-consuming and tedious manual work, was not compromised by the documents' totally incongruent file structure. While not ideal for the printed edition, the challenges were overcome thanks to the fact that it was entirely hand-set layout and it did not create problems in the generation of the typesetting. For instance, notes to the text are in their own InDesign story,[6] disconnected from reference numbers in the text body, that are simple Roman and Arabic numbers formatted as superscript. Even decoding the nested curly brackets was left to the experienced eye of the typesetters.

3. FROM CHAOS TO STRUCTURE

When these same files were used to begin preparing an XML/xHTML version, the total lack of structure in the text become a huge obstacle, making us fear the repetition of the manual typesetting work even for the creation of the base text of the digital edition. At the time, the final version of the InDesign document that was then sent to print was not available, but even when the files arrived, they were not easier to convert than the original unstructured word processor files.
With this in mind, we split the work with one side converting the XML content of the DOCX files into semi-structured TEI, and the other cleaning up the DOCX documents to improve the success chances. The first results were achieved by combining the experience in XML file development and manipulation of the Central IT of the University of Zurich in one hand with long practice in manipulating and normalizing text files delivered by authors in MS Word format for conversion to structured HTML and import in InDesign on the Digital Publications unit of Bibliotheca Hertziana. This collaboration resulted in an initial manual

[4] Optical Character Recognition. On the problems of working with poor quality results (Cordell, 2017).
[5] (*Text Encoding Initiative*, n.d.).
[6] In Adobe InDesign, a story is a group of text frames with flowing text, i.e. main text, appendices, index of names.

prototype of a portion of the text, which was essential for identifying a clean structure template that could be easily transformed between formats.
The built-in style names (i.e. Body Text, Footnote Text, Heading 1) in a template for the word processor (MS Word) are localized between different system and languages. For this reason, it was necessary to create a set of character and paragraph styles fully independent of the program's default styles to ensure consistency among the international research team.
The next step was to adapt the text documents with critical comments to the template. The manual procedure involved identifying local modification of the font (e.g., italicized or underlined text) and applying the correct character style to them. This made the text marked stable when it was necessary to apply the correct paragraph style, cleaning up the text from unwanted styles and returning it to its base character state. Since this was a process repeated several times both on the same file and in different files rather than repeating the sequence of changes by hand, a series of macros were created in Visual Basic for Applications (VBA), then combined together so that they were all executed in the correct order. VBA proved to be the most efficient way to deal with the complex search and replace, based mainly on how the content had been locally formatted.[7]
Once the styles cleanup was finished, it was necessary to distinguish the different uses of curly brackets, converting them, for example, into pseudo tags and font styles. To quickly identify the most common instances, such as page numbers, a regular expression search was used, again within MS Word, to automatically capture and transform as many instances as possible. Since the final plan was to convert the content into TEI XML, when possible, those curled brackets were replaced with "pseudo tags" that could easily be captured and substituted in any subsequent XSLT transformation. The macro for the page numbers, for instance, identifies parenthesis pairs that contain only Arabic or Roman numerals and replaces them with the `<pb>…</pb>` tag while keeping the number originally inside.
The most complex work involved the recognition and extraction of critical notes within the footnotes. Using the TEI note template as a reference, where author's notes are embedded as `<note>…</note>` tags in the superscript reference position, while critical comments are all relegated to a stand-off section of the file, we decided to move the footnotes inline, applying a different color (blue) and enclosing them between a pair of symbols that could not appear in the text (in our case the lower and upper half of the mathematical function symbol). Moving the original footnotes as a part of the main text makes it possible to use the standard endnotes for the critical comment. At this point it was possible to extract the text of the comments in curly brackets (again with a VBA macro), and move their contents to standard automatic endnotes. In order to retain the nested philological annotations, that were planned in curled brackets in the final rendition, it was necessary to temporary capture their content in pseudo tags. This could be performed almost automatically in the most common cases (single words such as {sic!} or references to page numbers) with the support of regular expressions in macros. Unfortunately, in other cases, where the brackets were not properly balanced or there were less common semantic notes or special annotations, hand verification was necessary.
The normalization of the work documents used by the editors of each volume was applied even to work in progress documents, in order to reduce the necessity of inserting curled brackets, although in some cases the editors preferred to wait until the critical work was finished according to previous editorial norms rather than find the notes integrated into the text and insert the comments directly as footnotes. The macros have been organized in a version of the template, which remains separate from that used by the authors because of MS Word's security limitations against documents with macros.[8]
The final text was then converted to TEI XML using the TEI XSLT Stylesheets[9]. Subsequent transformations catch the styles defined by the project-specific template and introduce tags specific to our document types. A series of substitutions based on regular expressions replace the pseudo tags and symbols with the correct XML markup. While the combination of transformations and substitutions brings the danger of breaking the document structure, we found this useful to detect issues in the markup of the MS Word documents which we could fix there and re-try the transformation.
To ease development and to ensure consistency, these steps are orchestrated by scripts specific to the parts of the books.[10]

[7] Templates, with and without activated macros, and macros are available as dataset (Bastianello & Baumgartner, 2025).
[8] See note n.7.
[9] (*TEIC/Stylesheets*, 2013/2025).
[10] The scripts and stylesheets are available as dataset (Baumgartner & Bastianello, 2025).

In addition, numerous transformations were applied to create the bibliography and link it correctly to the bibliographic abbreviations from the formatted text.

4. PUBLICATION OF THE TEXT

The second phase is concerned with bringing the text online. Our personal experience with digital editions is that when the publishing interface is created specifically for an edition, there is a high risk that at the end of the funds for the edition the maintenance of the infrastructure is delegated to the goodwill of the team members still working with the institution hosting the edition. This has led to the loss of many editions created since the late 1990s, only in some cases recovered through specific recovery projects in recent years.[11]

Because our digital publishing project depends on funds that, in part, expired on December 2024, several strategies have been developed to ensure long-term maintenance. One is the use of an open-source platform, TEI Publisher, whose maintenance and updating is ensured by an extended community of users and developers independent from the specific edition. The other is to associate the management of the digital edition, once development and coding are completed, with the digital publications manager position at Bibliotheca Hertziana.

At the beginning of development in 2019, the choice fell on TEI Publisher (Turska et al., 2016), at the time in version 6, because the other open-source promising alternative, EVT 2 (Turco et al., 2014, 2019) had not yet been fully developed as a web service. On the one hand, this choice proved successful, because with the inclusion of annotation tools in version 7 (August 2021)[12], linking named entities became intuitive and simple. In fact, it was only a matter of determining the authority file source of the identifiers (in our case the GND[13]) and how to store the information not present in the form of local stored data. Annotations also became a valuable control tool, since the researcher who was in charge of entity tagging could also mark on the fly typos or other items to be rechecked later. At the same time, TEI Publisher's way of interpreting the TEI XML by manipulating the display using ODD[14] is not completely agnostic and requires the insertion of a part of special tags, for example to indicate parallel editions (original edition and new critical printed edition) as milestones. Also, the web components used in the service boxes, for example in the one containing critical comments or images of the facsimile and illustrations, had not yet reached full maturity. This meant a major upgrade of the entire edition, albeit limited to the volume used as a test, during the upgrades to versions 7, 8 and 9. Future version 10, coming soon, is expected to simplify the platform upgrade without requiring updates to the content of the edition, but will require editing of the code prior to the upgrade.

An additional advantage of using TEI Publisher is that it is an application that operates on top of eXist-db, with a powerful indexing system based on Apache Lucene. This makes it possible to configure an integrated structured search of both the individual volume and across the entire critical edition and potentially across several editions. HWGW was created as a stand-alone application and not as a part of a basic instance of TEI Publisher but of a specific application, for which it was possible to change entirely even the appearance of the site that welcomes users.[15] The online edition is fully responsive, to ensure readers can check the content on their desktop and mobile devices.

5. IMAGES AND EXTERNAL SITES

Images in the Digital Critical Edition are available through a IIIF[16] viewer integrated into the TEI Publisher platform. Up to version 8, it was only possible to include individual images using the IIIF image API,[17] while starting with version 9, it is possible to use a manifest JSON file and the presentation API,[18] even if only v.2. This allowed us to integrate additional images such as book covers and blank pages with comments by the author. Furthermore the option to add a local manifest allowed the integration of

[11] A case in point is the "Biblioteca delle fonti storico-artistiche" formerly available at the Scuola Normale Superiore in Pisa, created by Paola Barocchi and restored several times. The relics of the project are available through the Way Back Machine *https://web.archive.org/web/20190913121736/http://fonti-sa.sns.it/fsaInfo.php*.

[12] (*Annotation Editor Released with New TEI Publisher 7.1.0*, n.d.).

[13] Gemeinsame Normdatei (*GND - Homepage*, n.d.).

[14] One Document Does it all (Rahtz & Burnard, 2013).

[15] The instance is available in GitHub (*Biblhertz/Hwgw*, 2025/2025).

[16] International Image Interoperability Framework. (*IIIF Home*, 2025; Kelli & Di Cresce, 2019).

[17] (*Image API 3.0*, 2020).

[18] (*Presentation API 3.0*, 2020).

metadata missing from the official version hosted on the Archive.org servers. The decision to link directly to the source images seemed the most appropriate, as it maintained the direct connection with the Getty Research Institute, which owns the author's personal volumes complete with his manuscript annotations. Unfortunately, in October 2024, a DDoS attack[19] on the archive.org servers forced the institution to shut down most of its public services, and in particular the IIIF server, which had not been restored by the publication of the online edition. Fortunately, the Getty Research Institute allowed the team to publish the images directly from the institutional IIIF server, which is on the same cluster as the critical edition. This experience has made us even more aware of the fragility of using linked resources, even though common and shared protocols such as the IIIF, within a critical edition.

6. FROM THE PROTOTYPE TO THE ONLINE EDITION

Although the first alpha version of the volume edition was developed from the MS Word file sent to the publisher of the print edition, it was necessary to incorporate into the text all the corrections that the editors had added in the layout during proofreading. In particular, moving critical comments into footnotes had revealed that some interlinear annotations, which appeared clear and obvious as long as they were within an in-text citation, became ineffective in footnotes and needed rework and rethinking.
Unfortunately, due to the manual typesetting described above, it was not possible to directly convert the final texts to XML as we did for the MS Word file. It was therefore necessary to export the individual InDesign stories as RTF files and convert them to xHTML. For the volumes where we already had the TEI XML data, we had to compare the text content with the xHTML data. In the future, for volumes that will not be typeset for printing by the publisher, the InDesign typesetting will be generated directly from TEI XML. In this way we can avoid dealing with different versions.
We have developed a comprehensive transformation pipeline that bridges the gap between our TEI XML source files and InDesign layouts[20]. This pipeline uses XSLT transformations to map TEI elements to corresponding InDesign styles through a structured mapping configuration. The process first converts TEI XML to a specialized intermediate format compatible with InDesign's XML import functionality, preserving both structural hierarchy and stylistic information. When importing into InDesign, each element is automatically assigned its corresponding paragraph or character style based on our predefined mapping schema. This bidirectional workflow ensures that any edits made to the source TEI are accurately reflected in the InDesign layout during subsequent updates, while maintaining layout integrity. This approach not only streamlines PDF generation for print outputs but also ensures perfect consistency between the digital TEI XML edition and any print derivatives. By implementing this semi-automated approach, we will significantly simplify maintaining consistency between the TEI XML source and the InDesign layout, eliminating the need for manual synchronization between digital and print versions.
The output is not the only part that will be connected. We have developed a direct transformation between neural OCR transcription in PAGE XML[21] format and TEI XML format for editions of print texts not yet transcribed and for manuscripts.[22] The ideal path was to have editors annotate directly in TEI format, either on TEI Publisher or on an XML editor, but this route was eventually abandoned. Developing a comfortable annotation interface for curators required time and resources that were not available, while annotating directly in XML on code editors required additional technical skills from curators. It was preferred to transform the generated TEI back into MS Word (with inline footnotes) and proceed as usual for the ongoing and future volumes (see Fig. 1).
To ensure easier access to the data, both creation information for the custom TEI Publisher instance [23] and the TEI XML-formatted texts[24] have been published on GitHub. In particular, for long term preservation, the edition data are synchronized on a research data repository that publishes the latest released version of the code, complete with project metadata.[25]

[19] Distributed Denial-of-Service (Brewster, 2024)
[20] This workflow is part of the editorial general process at Bibliotheca Hertziana.
[21] Page Analysis and Ground truth Elements (*PRImA*, n.d.).
[22] The code is published as open-source in (*Biblhertz/Trans2tei*, 2021/2024) and was detailed on (Bastianello & Baumgartner, 2023). The workflow uses Pylaia engine with a model trained to add the information and a P2PaLA model for layout tagging, in the Transkribus platform (*Transkribus - Unlocking the Past with AI*, n.d.)
[23] The custom app information is available at (*Biblhertz/Hwgw*, 2025/2025).
[24] The TEI XML data are available at (*Biblhertz/Hwgw-Data*, 2025/2025).
[25] (Bastianello et al., 2025)

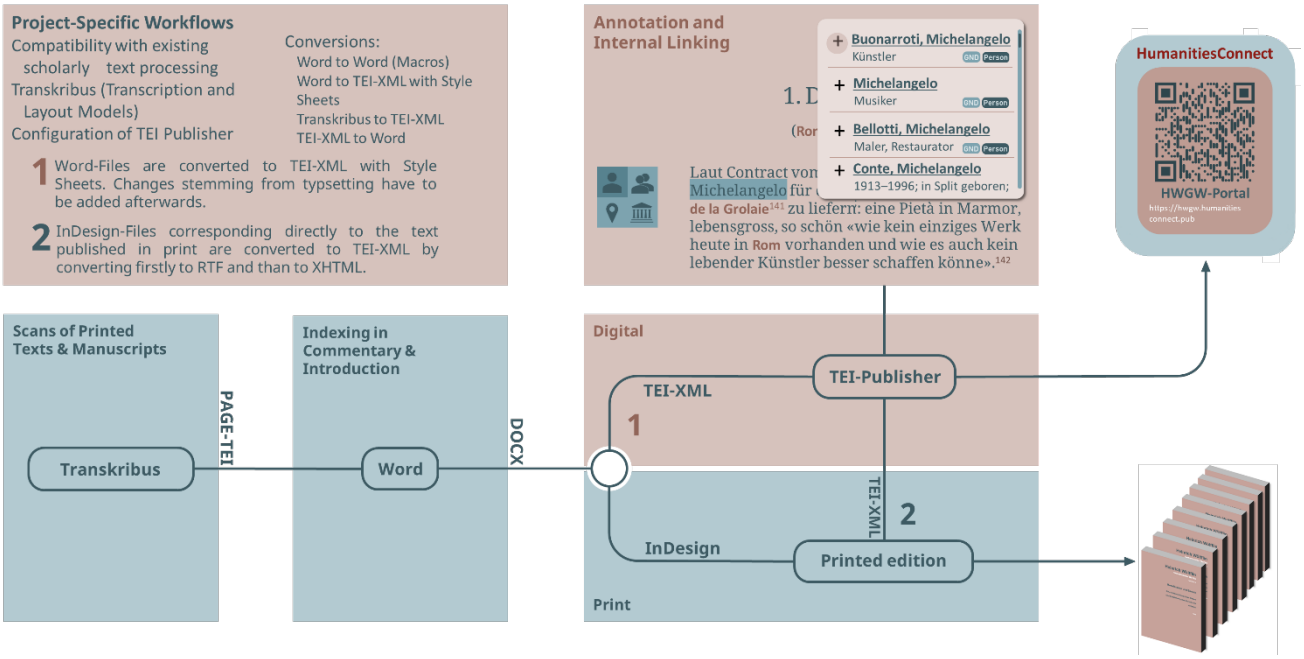


Figure 1 - Schema of the critical edition workflow.

7. CONCLUSIONS

Publishing a digital critical edition from a traditional workflow is not a linear path; it often requires going back and looking at issues from different perspectives. The collaboration of team members with experience and skills in different fields has been one of the key advantages of our working group, so we have decided to make all the development public, in the hope that it will also be of help to other teams, especially those who cannot rely on sufficient funds and such broad expertise.

ACKNOWLEDGEMENTS

This critical edition was funded by SNSF as a research project (grant n. *160081*) from 2015 to 2017 and continued as an editorial project from 2017 to 2024 (grants nn. *157979* and *198243*). Support to the project has been granted by the Bibliotheca Hertziana – Max Planck Institute for Art History since 2017 with grant id *BH-P-19-32*.

REFERENCES

Annotation editor released with new TEI Publisher 7.1.0. (n.d.). Retrieved January 26, 2025, from *https://www.e-editiones.org/posts/annotation-editor-released-with-new-tei-publisher-7-1-0/*.

Bastianello, E., & Baumgartner, R. (2023). L'applicazione del riconoscimento testi neurale per la realizzazione di ristampe digitali. In E. Carbé, G. Lo Piccolo, A. Valenti, & F. Stella (Eds.), *La memoria digitale: Forme del testo e organizzazione della conoscenza. Atti del XII Convegno Annuale AIUCD* (pp. 15–23). Associazione per l'Informatica Umanistica e la Cultura Digitale.

Bastianello, E., & Baumgartner, R. (2025). *Heinrich Wölfflins –Gesammelte Werke (HWGW) Digital Edition: Word clean-up conversion scripts* (Version 1.0) [Dataset]. Edmond. *https://doi.org/10.17617/3.4A6MNT*.

Bastianello, E., Baumgartner, R., Meier, S., & Weddigen, T. (2025). *Heinrich Wölfflins –Gesammelte Werke (HWGW) Digital Edition Dataset* (Version 1.0). Edmond. *https://doi.org/10.17617/3.QHJW4D*.

Baumgartner, R., & Bastianello, E. (2025). *Heinrich Wölfflin – Gesammelte Werke (HWGW) Digital Edition: XSLT transformations* (Version 1.0) [Dataset]. Edmond. *https://doi.org/10.17617/3.15WQ42*.

Biblhertz/hwgw. (2025). [HTML]. Bibliotheca Hertziana Digital Humanities Lab. *https://github.com/biblhertz/hwgw* (Original work published 2025).

Biblhertz/hwgw-data. (2025). [XQuery]. Bibliotheca Hertziana Digital Humanities Lab. *https://github.com/biblhertz/hwgw-data* (Original work published 2025).

Biblhertz/trans2tei. (2024). [XSLT]. Bibliotheca Hertziana Digital Humanities Lab. *https://github.com/biblhertz/trans2tei* (Original work published 2021).

Brewster, K. (2024, October 18). *Internet Archive Services Update: 2024-10-17 | Internet Archive Blogs*. *https://blog.archive.org/2024/10/18/internet-archive-services-update-2024-10-17/*.

Cordell, R. (2017). "Q i-jtb the Raven": Taking dirty OCR seriously. *Book History*, *20*(1), 188–225. *https://doi.org/10.1353/bh.2017.0006*.

GND - Homepage. (n.d.). Retrieved January 26, 2025, from *https://gnd.network/Webs/gnd/EN/Home/home_node.html*.

Image API 3.0. (2020). *https://iiif.io/api/image/3.0/*.

international image interoperability framework—Home. (2025, January 7). *https://iiif.io/*.

Kelli, B., & Di Cresce, R. (2019). Impact of international image interoperability framework (IIIF) on digital repositories. In K. J. Varnum (Ed.), *New top technologies every librarian needs to know* (pp. 181–196). ALA Neal-Schuman.

Presentation API 3.0. (2020). *https://iiif.io/api/presentation/3.0/*.

PRImA. (n.d.). Retrieved January 26, 2025, from *https://www.primaresearch.org/tools/PAGELibraries*.

Rahtz, S., & Burnard, L. (2013). Reviewing the TEI ODD system. *Proceedings of the 2013 ACM Symposium on Document Engineering*, 193–196. *https://doi.org/10.1145/2494266.2494321*.

TEIC/Stylesheets. (2025). [XSLT]. Text Encoding Initiative Consortium. *https://github.com/TEIC/Stylesheets* (Original work published 2013).

Text Encoding Initiative. (n.d.). Retrieved January 26, 2025, from *https://tei-c.org/*.

Transkribus—Unlocking the past with AI. (n.d.). Retrieved February 15, 2025, from *https://www.transkribus.org/*.

Turco, R. R. D., Buomprisco, G., Pietro, C. D., Kenny, J., Masotti, R., & Pugliese, J. (2014). Edition Visualization Technology: A Simple Tool to Visualize TEI-based Digital Editions. *Journal of the Text Encoding Initiative*, *Issue 8*, Article Issue 8. *https://doi.org/10.4000/jtei.1077*.

Turco, R. R. D., Pietro, C. D., & Martignano, C. (2019). Progettazione e implementazione di nuove funzionalità per EVT 2: Lo stato attuale dello sviluppo. *Umanistica Digitale*, *7*, Article 7. *https://doi.org/10.6092/issn.2532-8816/9322*.

Turska, M., Cummings, J., & Rahtz, S. (2016). Challenging the Myth of Presentation in Digital Editions. *Journal of the Text Encoding Initiative*, *Issue 9*, Article Issue 9*. https://doi.org/10.4000/jtei.1453*.

Human–LLM Synergy in Higher Education Publishing: Two ChatGPT Use Cases within Editorial Pipelines

Gianluca Pavani[1],
[1] Università degli Studi di Roma Tor Vergata, Italia - gianlpav94@gmail.com
Edra S.p.A., Italia - g.pavani@lswr.it

ABSTRACT (ITALIANO)

Il presente contributo intende analizzare l'utilizzo di Large Language Models (LLM), con particolare riferimento a ChatGPT, nell'ambito dell'editoria scientifica e accademica. Sebbene il ruolo dell'intelligenza artificiale generativa nella scrittura creativa e nella narrativa sia stato ampiamente esplorato, la sua integrazione nei processi di scrittura scientifica ha ricevuto, ad oggi, una attenzione significativamente più limitata. Gran parte del dibattito attuale si concentra su implicazioni di natura etica e sul possibile indebolimento del rigore scientifico e autoriale. Questo studio adotta un approccio qualitativo per esaminare in che modo i LLM possano essere concretamente integrati nei flussi editoriali, presentando due casi applicativi sviluppati nel contesto dell'editoria universitaria. L'obiettivo è individuare metodologie operative e buone pratiche emergenti che favoriscano un utilizzo responsabile ed efficiente dell'intelligenza artificiale nella produzione di contenuti accademici.
Parole chiave: Large Language Models (LLMs); ChatGPT; scrittura scientifica; editoria universitaria

ABSTRACT (ENGLISH)

Human–LLM Synergy in Higher Education Publishing: Two ChatGPT Use Cases within Editorial Pipelines. This paper investigates the use of Large Language Models (LLMs), with a focus on ChatGPT, in the field of scientific and academic publishing. While the role of generative AI in creative and fictional writing has been extensively explored, its integration into scientific writing processes has received comparatively limited attention. Much of the existing discourse remains centered on ethical implications and the possible weakening of scientific rigor and authorial credibility. In contrast, this study takes a qualitative approach to examine how LLMs can be practically embedded in editorial workflows, presenting two real-world applications within higher education publishing. The objective is to identify concrete methodologies and emerging best practices that support the responsible and efficient use of AI in the production of academic content.
Keywords: Large Language Models (LLMs); ChatGPT; scientific writing; academic publishing

1. INTRODUCTION

The spread of generative artificial intelligence (AI) systems, particularly OpenAI's well-known ChatGPT, has sharply brought issues of potential uses, regulations, and risks into the educational landscape. The wide range of these crucial topics encompasses all stakeholders: authors, publishers, academic and school settings, professors, researchers, and students. This paper aims to investigate potential pipelines for implementing generative tools. The two case studies presented were conducted in collaboration with *Edra*, an Italian medical-scientific publishing house, and the discussed results will be considered as possible good practices for generative AI tools in the editorial production chain of educational content. Finally, the paper is also intended as a study on the AI's ability in scientific writing, a much more complex field than creative and fiction writing. Generative systems, often seen as unreliable facilitators prone to hallucinations, needs to be examined from a multi-level and interdisciplinary perspective, with an awareness that the effort of stakeholders should be directed towards proper implementation and regulation, rather than ineffectual limitations.

2. WRITING DRAFTS FROM NOTES: PROMPT-BASED APPROACH TO CHATGPT

The case study under examination pertains to an editorial project for the medical-scientific catalog of the Italian publishing house Edra. This project aims to publish two volumes of university textbooks on Biotechnology. The editorial initiative is part of a larger project comprising six volumes on Biomedicine, scheduled for publication by 2026. The first two volumes, "Principles of Biotechnology" and "Biotechnology Drugs," are anticipated to be released by early 2025 and will target an audience of approx. 1300-1500 students, with about two-thirds corresponding to the first volume, which is intended for the Bachelor's degree cohort.

The case study involved the application of ChatGPT, in versions 4 and 4o, for building the draft of the first two volumes of Biotechnology. The first manual is structured into eleven chapters and second one into twelve, totaling approximately three hundred and twenty printed pages each. OpenAI's LLM was utilized to generate drafts for twenty-one chapters. The editorial project involves a scientific board, composed of the headline authors, and coordinated scientifically by three external editors. Both the editors and the authors serve as Associate Professors or Full Professors at Biotechnology Faculties in the Universities of Milan (Università di Milano Statale, Università di Milano-Bicocca, Università Vita-Salute San Raffaele).
Drafting process using ChatGPT began with topics outlined in PPTX files, which comprised over five hundred slides in total. These PowerPoint files were created by the headline authors and their assistants as teaching support in courses, subsequently collected and selected by the scientific editors according to the editorial project. Consequently, the provided material, as a source of interpreted literature, is considered scientifically validated from the outset. The use of ChatGPT in versions GPT-4 and GPT-4o can be described as an extension of the content from all over five hundred slides, a process referred to by the system as "Transforming Slide Concepts". ChatGPT was employed throughout the entire process of drafting, expanding the textual content present in the PPTX files. The available slide material typically consists of unstructured texts, bullet points, images, charts, and quotes, which were transformed into extended and argumentative text. Furthermore, ChatGPT was tasked with providing a logical sequence and coherent indexing of the generated paragraphs, while maintaining existing structures from the slides (e.g., bullet and numbered lists) and extending their contents. GPT-4 and 4o were also employed to create captions for some images that lacked them.
The entire draft production process (e.g., Figure 1) was based on two primary criteria: the presence of external forms of scientific validation and the adoption of appropriate prompting strategies.

1. Forms of Scientific Validation: given the use of generative AI tools in drafting scientific texts for educational purposes, primary importance was placed on the presence of sufficient validation forms at both ends of the process. The first form of validation is present from the inception of the case study. The prompts exclusively included texts, data, graphs, images, and quotes from the provided PPTX files. As previously mentioned, the slides supplied by the authors were an integral part of the university teaching material, considered as sources of interpreted literature and thus validated upstream. These slides were deemed consistent with the editorial project during the material collection by the coordinating editors. The coordinating editors also handle the initial review of the scientific content. This phase is preceded by a structural check of the drafts by the catalog editor and followed by submission to the authors for final revisions, including any necessary cuts and additions. The outcome will be the second draft, formatted in Word, which will undergo another review by the coordinating editors for final approval.

2. Prompting Strategies: the efficiency of utilizing LLMs (Large Language Models) such as ChatGPT is inextricably linked to the proper construction of prompts, particularly in the context of models that perform reasoning tasks (Mialon et al., 2023). During the process of expanding unstructured scientific text into extended and argumentative prose, having prompting strategies is considered *conditio sine qua non* for obtaining satisfactory outputs. The strategies adopted can be categorized as follows:

- Extent Measurement: inclusion of specifics regarding the desired length for text extension provided in the prompt. ("*Extend the text provided in the prompt to no less than (n. of tokens)*," with the highest efficiency achieved within a range of two hundred to five hundred tokens. In this phase, prompts were mainly zero-shot.
- Linguistic Types of the Text to be Generated: prompts included instructions on the linguistic register and terminology to be used. ("*Use formal language appropriate for university textbooks*", "*Utilize terminology belonging to the technical jargon of the subject – Biotechnology*"). For this task, one-shot prompts have been largely used.
- Structure of the Text to be Generated: essential for obtaining the desired textual form. ("*Transform into extended and argumentative text, divide into titled paragraphs*", "*Insert bullet and numbered lists where necessary and maintain those already present*"). This phase required mostly few-shot prompts.
- Variable Prompts for Targeted Corrections: inserted as needed to correct parts of the generation, significantly facilitating the editing phase. ("*Avoid redundancies in the conclusions*").

The extension process, started at the end of April 2024, produced twenty-one AI-generated drafts of the first two volumes of Biotechnology by the end of August 2024. To date, all drafts generated with ChatGPT have been considered scientifically and editorially coherent from the first review by the editors.
From the application of ChatGPT for the generation of drafts in the medical-scientific editorial context, the following observations were made:

- Constant attention was directed towards some of the major issues related to the generation of scientific texts, namely the potential occurrence of hallucinations and biases in the outputs. Neither during the control phase in editing nor from the scientific evaluation by the editors did any hallucinations or biased generations emerge.
- ChatGPT proved effective in the extent and quality of text expansion, in the creation of ad hoc captions for images, including complex ones. Additionally, it provided correct indexing and titling of paragraphs, helping to build the structure of subsections, paragraphs, and subparagraphs (the overarching chapter structure was known from the start).
- ChatGPT capabilities allowed for minimizing the editing phase, which primarily focused on correcting redundancies and occasional typos.
- By providing scientifically validated material in the prompt, ChatGPT transformed the slide content into coherent text with logical sequencing and a remarkable level of quality and originality. Furthermore, through AI, it was possible to contextualize and integrate quotes, data, documentary content, and images that were otherwise detached from the rest of the content in the slides.

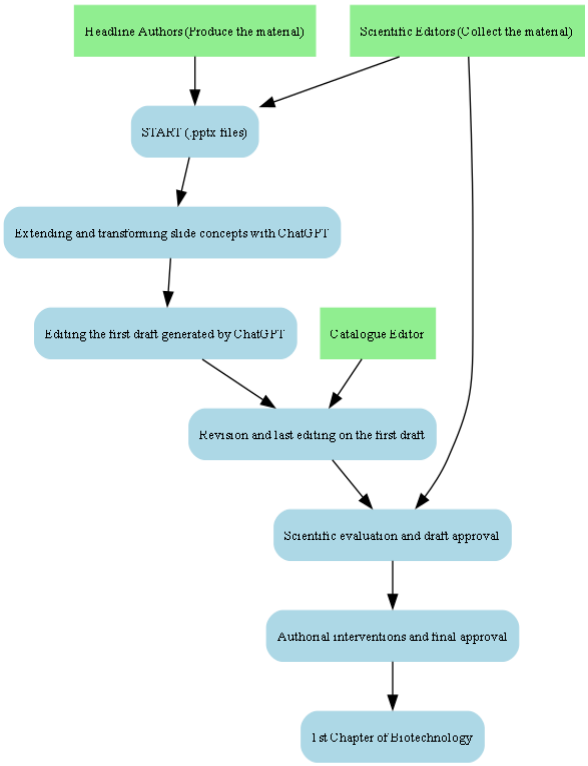


Figure 1. Workflow and validation steps

3. DOMAIN-GROUNDED, PROMPT-BASED CUSTOMIZATION AND EMBEDDED DOCUMENT CONTEXT VIA CHATGPT API

The second case study presented here illustrates an alternative implementation – currently ongoing – of a Large Language Model within editorial workflows. In contrast to the first project, which focused on extending and reshaping unstructured source material, this initiative aims to distill content from extensive textual sources for the purpose of generating structured drafts. Final output will consist of a concise revision study guide book designed for second and third-year medical students enrolled in the Medical Physiology course. The project involves collaboration among a subject-matter expert responsible for curating content and defining the indexing structure, as well as editors and catalog editors from the

publishing house. Final publication is scheduled for late 2025, pending further refinement of the editorial project.
Production process relied on both open-access materials and proprietary textbooks and study guides from the Edra catalog. Given the need to compress and restructure large bodies of content, and in light of the token limitations of direct prompt-based interaction with the chatbot interface, the workflow adopted a more technically structured approach using the ChatGPT API, Python scripting, and document storage functionalities. The technical workflow and editorial integration followed the key stages detailed as follows:

1. Model and Temperature Control: the language models (GPT-4 and GPT-4o) were accessed via API and configured with Python to ensure consistent, accurate, and context-aligned summarization. After some testing, the temperature parameter was adjusted within a range of 0.2 and 0.5. Lower values (0.2–0.3) were used for dense, technical chapters ensuring terminological precision and minimal creative deviation. Slightly higher values (0.4–0.5) were tested when summarization required more abstraction and paraphrasing.
This regulation allowed the model to maintain both semantic fidelity and stylistic consistency with the domain-specific content.

2. File Upload and Contextual Anchoring: source documents, such as university-level textbooks, were uploaded as TXT files to the ChatGPT file storage via API. Each document was then referenced in the prompt using its file ID in Python, providing the model with grounding in domain-specific terminology and disciplinary structure. These files acted not only as the objects to be summarized, but also as semantic and stylistic context, anchoring the generation in the established discourse of medical physiology.

3. Chunking Strategy: files were segmented to manage long documents and align with token window limitations. Each chunk was sized at approximately 1500-2000 tokens, equivalent to 1000–1600 words depending on the source format. This granularity allowed sufficient context retention while keeping the input within manageable token limits for API calls, showing consistency and completeness in summarization, with balance between context and focus.

4. Target Output Length and Compression Ratio: each input chunk was summarized into approximately 200–400 tokens, which equates to a compression ratio of 5:1 to 8:1, depending on the source density and repetition. This compression level ensured that all relevant concepts were preserved in a concise, revision-oriented form, avoiding excessive abstraction or loss of terminological accuracy.

5. Persona, Prompting and Content Constraints: the prompt consistently assigned the model the role of a senior ghostwriter with expertise in scientific and medical writing, specialized in Medical Physiology. Beyond directing the model toward coherent disciplinary tone with the persona, each prompt included specific instructions to maintain alignment with the source material. These prompts enforced:

- Terminological precision, avoiding simplification of medical vocabulary.
- Tone appropriate for academic writing, neutral and informative rather than creative or colloquial.
- Structural mirroring of the source material, when applicable (e.g., maintaining topic subdivisions and preserving sequential logic).

The prompt reinforced that the summaries should serve as revision tools, assuming that the reader had prior exposure to the full content via textbooks and lectures.

6. Human Oversight and Indexing: the output underwent a manual editorial phase by expert and editors, ensuring fidelity to scientific and medical standards. This phase included refining terminology and phrasing, eliminating redundancies occurring across adjacent chunks (use of ChatGPT was included for this task) and validating the factual accuracy of complex physiological descriptions. Subsequently, all summaries were indexed by the editorial team of the publishing house, according to the structure developed by the collaboration between expert and university-affiliated library partner.

Between July and November 2024, the process resulted in the completion and editorial approval of draft versions for twelve chapters of the Medical Physiology study guide book. This timeline primarily reflects the iterative nature of editorial validation and content refinement. Since all processing is handled via the OpenAI API, the workflow itself remains technically lightweight and the infrastructure requirements minimal, making this approach scalable and replicable even in small editorial settings with limited resources.

4. DISCUSSION

It is important to preface the following discussion of the two case studies with the observation that it is primarily grounded in a coherent interpretation of the results obtained, given the previously mentioned scarcity of literature on comparable use cases or studies.

Firstly, from a general observation, ChatGPT demonstrates remarkable effectiveness in writing activities within the editorial workflow of a university textbook. Moreover, given the generation of texts pertaining to a well-defined domain, the study also examined ChatGPT's behavior in the complex task of scientific writing. The application of ChatGPT —and LLMs in general— in the field of scientific writing is subject to significant controversy. The main ambiguities revolve around scientific reliability, respect for authorship, and copyright. While it is true that these issues are widely discussed in the literature, many of the existing studies have stemmed from considerations on cases different from the ones presented here, such as the growing use of generative AI as a ghostwriter in scientific articles, papers, and literary reviews. This phenomenon, particularly in the medical-scientific field, is perceived as a threat to the integrity of academic ethics (Khalifa, 2023); the reason lies in the partly compromised principle of authors' agency, as well as in the risk of plagiarism, which is notably difficult to detect when GPT is tasked with generating scientific content.

However, these risks, though undoubtedly significant, seem to be particularly relevant in two circumstances. The first involves generation from scratch, or more precisely, generation based solely on the statistical-probabilistic criteria applied to GPT's memory. Although improvements have been made in this regard with the latest versions, ChatGPT is not designed to provide precise bibliographical references and struggles to distinguish Open Source literature, moreover in previous versions it provided wrong PMID numbers (Alkaissi & McFarlane, 2023). Additionally, there is the risk of potential hallucinations, such as fictitious references (Babl & Babl, 2023). The second circumstance involves the input of external material into the prompt to summarize existing literature on a specific topic (Salvagno et al., 2023).

When implementing ChatGPT within the first editorial project, all the major known issues were taken into account, and the factor that enabled justified skepticism to be overcome was the inherent characteristics of the project itself. The principle of authorship, or authors' agency if preferred, is respected insofar as the validity of the draft produced by ChatGPT—scientifically, conceptually, and terminologically—has been verified. It is the authors' additions and modifications that ultimately shape the final version of the text. The issue of bias and hallucinations is strongly mitigated by the use of sources that are considered validated at the outset (albeit in an unstructured form) and, above all, through the adoption of various steps to validate scientific rigor by qualified personnel. It is worth emphasizing that all drafts produced were approved by the scientific board, as no hallucinations or significant errors emerged in the generation of scientific content.

While this first case demonstrates the potential of LLMs for expanding loosely structured content through flexible, interface-based prompting, it could also revealed certain limitations in terms of replicability and scalability on domain control, consistency, and contextual fidelity, particularly when working with larger textual contents. To explore how these challenges might be addressed through a more structured and domain-grounded setup, the second case study shifts focus toward the summarization of complex source materials via API orchestration. This approach prioritizes semantic alignment and editorial control, offering a technically more sophisticated, albeit narrower, implementation of LLM-assisted content production.

In the second case of use, rather than on a fine-tuning, the system relied on a domain-specific adaptation pipeline based on the ChatGPT API and documents as contextual anchors within the generation process; this technique can be described as a lightweight domain adaptation strategy, on top of a general-purpose LLM via API-level configuration. While not altering the underlying model architecture or parameters, the system effectively simulated domain-specific competence by combining embedded content references with a document-linked prompt orchestration structure. In a way, a dynamic and semi-structured form of knowledge grounding – clearly distinct from Retrieval-Augmented Generation (RAG) – yet could be functionally comparable in terms of enhancing relevance and coherence within a predefined domain.

The result so far seems to be a controllable and reproducible output pipeline tailored to the editorial aim of producing drafts of high-utility texts for study guide books.

Both use cases can lead to broader observations about the content generated by ChatGPT: different types of writing—on one hand generalist, creative, and fictional; on the other, scientific, in-depth and rigorous—inevitably do not suffer from the same critical issues. The use of LLMs in creative writing is often regarded as a potential threat to the quality and variety of content, and the greatest risks are particularly associated with its use during school-age years, when writing skills are still being developed (Shidiq, 2023). Indeed,

while generative AI can be credited with a certain degree of originality (which is undoubtedly improving in the latest versions of the models), widespread use of these tools would sooner or later lead to redundant content, both in terminological and syntactic form and in substance. As mentioned, this application of AI has been primarily studied in other contexts, and the majority of the literature on the topic focuses on ethical concerns (Otmar et al., 2024). Only a few studies have attempted to explore alternative forms of collaboration aimed at positively impacting the performance of both parties —humans and LLMs— involved (Chackrabarty, Laban & Wu, 2025). Research concerning the use of LLMs in scientific writing—particularly in the medical field—has almost exclusively addressed the drafting of journal articles or literature reviews (Ahn, 2024), rather than editorial practices related to the publication of academic textbooks.

However, based on the findings of this editorial project, their use in the context of scientific writing seems to lead to a different paradigm. Scientific knowledge, particularly that aimed at academic teaching, is inherently more fixed, despite its constant evolution. In a way, scientific writing allows for greater flexibility in form, provided that the essential requirement of rigor is upheld. In this sense, the use of LLMs could be seen as a tool to facilitate the transmission of scientific knowledge, as they are highly effective at reinterpreting its form. The case presented here constitutes an example, as ChatGPT was not used to produce significant scientific insights but rather to structure non-cohesive and non-argumentative texts, leveraging its Natural Language Processing (NLP) capabilities. Through AI, it was possible to transform unstructured scientific sources from interpreted literature, such as slides used in university lectures, into a different textual form in significantly less time than was previously required.

In other words, it could be argued that LLMs may represent highly valuable tools in the future for reworking and disseminating scientific knowledge (Markowitz, 2024), or contribute to making scientific communication more accessible, as has been particularly noted in the use of small models subjected to fine-tuning (Jiang, Shi & Luo, 2024). From an editorial perspective, it could translate into greater productivity in publishing and more rapidly updatable catalogs. Such benefits represent an undeniable advantage in presenting scientific knowledge that is aligned with the state of the art, which is fundamental in rapidly evolving fields, such as the medical-scientific one, as well as in more promptly addressing the needs of students and, more generally, the academic environment.

Despite the divergent technical setups and editorial aims of the two case studies, they share two fundamental commonalities. In both cases, the textual material served a dual function; it operated simultaneously as the target of generation (i.e., the content to be transformed or summarized) and as a source of contextual and domain-specific grounding. Moreover, both projects were characterized by a structured collaboration between human academic and editorial figures and the language model, with the latter operating not in isolation but as a component within a carefully supervised pipeline.

These findings suggest that, rather than striving for full automation, the human–LLM synergy can offer a more promising and sustainable path forward for integrating generative models into academic publishing workflows.

5. CONCLUSIONS

The results emerging from both case studies support a broader reflection on the evolving role of generative language models within editorial and publishing ecosystems. The main challenges associated with these systems have often been treated as limitations to be highlighted rather than as opportunities for exploring potential solutions. Rather than replacing human expertise, the integration of LLMs into these pipelines points toward a symbiotic model of collaboration, in which the machine enhances human capabilities without displacing them. This human–LLM synergy enables a division of labor that is both effective and sustainable: the model contributes scale and consistency, while human editors and experts ensure semantic precision, disciplinary relevance, and alignment with pedagogical intent. Such an approach reframes generative AI not as a tool of full automation, but as a cooperative agent embedded in domain-specific processes, aligning with emerging paradigms of human-in-the-loop knowledge production, where credibility and scientific rigor remain central. If we consider the balance between efficiency, cost-effectiveness, time constraints, and reliability as crucial for the application of AI systems in higher education publishing contexts, collaborative and symbiotic frameworks may enhance the domain relevance and the trust in the outputs—key factors for a broader acceptance, and consequently applicability, of LLMs in this field.

REFERENCES

Ahn, S. (2024). The transformative impact of large language models on medical writing and publishing: current applications, challenges and future directions, Korean J Physiol Pharmacol 2024;28(5):393-401.
https://doi.org/10.4196/kjpp.2024.28.5.393

Alkaissi, U., & McFarlane, S. I. (2023). Artificial Hallucinations in ChatGPT: Implications in Scientific Writing, *Cureus* 15(2): e35179.
DOI 10.7759/cureus.35179

Babl, F. E., & Babl, M. P. (2023). Generative artificial intelligence: Can ChatGPT write a quality abstract?. Emergency Medicine Australasia, 35: 809-811.
https://doi.org/10.1111/1742-6723.14233

Chakrabarty, T., Laban, P., & Wu, C. S. (2025). Can AI writing be salvaged? Mitigating Idiosyncrasies and Improving Human-AI Alignment in the Writing Process through Edits, in Yokohama' 25: ACM (Association of Computing Machinery) CHIconference on Human Factors in Computing Systems, April 26–May 1, 2025, Yokohama, Japan. ACM, New York, NY, USA.
https://doi.org/10.48550/arXiv.2409.14509

Jiang, G., Shi, X., & Luo, Q. (2024). LLM-Collaboration on Automatic Science Journalism for the General Audience.
https://doi.org/10.48550/arXiv.2407.09756

Khalifa, A. A. (2023). Does the use of AI tools in academic publishing represent a new form of ghost authorship? *Croatian Medical Journal*, 64:292-4.
https://doi.org/10.3325/cmj.2023.64.292

Markowitz, M. M. (2024). From Complexity to Clarity: How AI Enhances Perceptions of Scientists and the Public's Understanding of Science. *PNAS Nexus*.
https://doi.org/10.48550/arXiv.2405.00706

Mialon, G., et al. (2023). Augmented Language Models: a Survey, in Machine Learning Research, June 2023.
https://doi.org/10.48550/arXiv.2302.07842

Otmar, R., et al. (2024). Ethics and the use of generative AI in professional editing.
https://doi.org/10.1007/s43681-024-00521-7

Salvagno, M., Taccone, F. S., & Gerli, A. G. (2023). Can artificial intelligence help for scientific writing?, *Critical Care* (2023) 27:75.
https://doi.org/10.1186/s13054-023-04380-2

Shidiq, M. (2023). The Use of Artificial Intelligence-Based Chat-Gpt and its Challenges for the World of Education; from the Viewpoint of the Development of Creative Writing Skills. *Proceedings of International Conference on Education, Society and Humanity*, Vol. 01 No. 01 (2023).
https://ejournal.unuja.ac.id/index.php/icesh

Taming the Hydra: A Model for Textual Dynamics and Constellations of Goethe's *Venetian Epigrams*

Daniele Fusi[1], Matteo Zupancic[2], Franz Fischer[3], Claus Zittel[4]
[1]Stuttgart University, Germany – daniele.fusi@ts.uni-stuttgart.de
[2]Università di Pisa, Italy – matteo.zupancic@unipi.it
[3]Università Ca' Foscari Venezia, Italy, franz.fischer@unive.it
[4]Università Ca' Foscari Venezia, Italy, claus.zittel@unive.it

ABSTRACT (ENGLISH)

This paper presents a digital model and software created in the context of the *VEdition* project, to provide a critical digital edition of Goethe's *Venetian Epigrams*. The paper proposes an innovative textological approach, focusing on a generic and reusable model of autographs to represent the dynamic nature of the creative process. While preserving the "objective" reproduction of documents separate from subjective scholarly interpretations, the model focuses on a single structured, computable and compact graph-based data structure, allowing to generate multiple text versions, annotated at any granularity level, for both textual and visual content. A full-fledged web UI (and an alternative complementary DSL) facilitates the creation of content, allowing scholars to focus on the reconstruction of the creative process at a higher abstraction level, while providing virtually unlimited export formats for integration with TEI-based production flows.

Keywords: digital scholarly editing; textology; epigrams; modelling

ABSTRACT (ITALIANO)

Domando l'idra: Modellare dinamiche e costellazioni testuali negli Epigrammi Veneziani di Goethe

Questo contributo presenta modello digitale e software creati nell'ambito del progetto *VEdition* per un'edizione critica digitale degli *Epigrammi veneziani* di Goethe. L'articolo propone un approccio testologico innovativo, incentrato su un modello generico e riusabile per rappresentare le genesi degli autografi, restituendo la natura dinamica del processo creativo. Preservando la riproduzione "oggettiva" dei documenti, unita alle interpretazioni soggettive degli studiosi, il modello si concentra su un un'unica struttura dati computabile, strutturata e basata su un grafo, che consente di generare versioni multiple del testo, annotate a qualsiasi livello di granularità, in ambito sia testuale che visivo. Un'interfaccia web (e un DSL complementare alternativo per un inserimento più rapido) facilita la creazione dei contenuti, consentendo agli studiosi di concentrarsi sulla ricostruzione del processo creativo a un livello di astrazione più elevato, e fornendo formati di esportazione virtualmente illimitati per l'integrazione con flussi basati su TEI.

Parole chiave: edizione scientifica digitale; testologia; epigrammi; modellizzazione

1. THE *VENETIAN EPIGRAMS*[1]

Despite an upsurge of scholarly interest in recent years, the *Venetian Epigrams* remain an anomaly within the extensive corpus of Goethe scholarship. To date, there is only one comprehensive monograph on the subject (Oswald, 2014), while more recent specialised research has begun to free the VE from the limiting association with the famous *Roman Elegies*, unearthing specific aspects of the work but leaving behind a still incomplete picture of it. Philological research has not progressed further than Jochen Golz's attempts to establish the possible criteria for a critical edition of the VE (Golz, 1998), which remain unfulfilled, and his diplomatic edition of two extensive documents of the work, edited with Rosalinde Gothe (Golz & Gothe, 1999). Consequently, the representation of Goethe's epigrams remains provisional and fragmented in the absence of an authoritative critical edition. The reasons for the scholarly reluctance to engage with the VE are manifold, but can be traced primarily to its complicated genesis, coupled with the burden of its scandalous content, which has weighed heavily on its subsequent reception. Above all, the composition of the VE is marked by a considerable but discontinuous manuscript tradition (Golz, 1998 and 1999), which shows a lack of any developmental coherence (Zittel, 2023)[2]. This documentary history must then be

[1] This paper was mutually agreed by the authors, with Matteo Zupancic writing sections 1 and 2 and Daniele Fusi writing sections 3, 4 and 5.

[2] Following some early drafts in Goethe's personal edition of Martial's poems, the poet began collecting a series of epigrams during his second stay in Venice (1790). Evidence of this phase can be found within a small quarto notebook (H 54, GSA 27/60), which includes at least 42 incomplete poems, and in a small selection of complete epigrams sent as

interwoven with a network of self-censorship and external censorship (Wilson, 2015), which constitutes a major obstacle to any philological reconstruction[3].

2. PREVIOUS PHILOLOGY AND A POLYP-LIKE ONTOLOGY

Besides the protean transformations from one document to another, we should also consider the issues posed by the *Ausgabe letzter Hand* (1827-1830), which was presumably approved by Goethe and published after his death by his closest circle of collaborators. The weight of Goethe's approval had farreaching consequences for later philology. The versions of the texts contained in this edition were long regarded as Goethe's last will, and thus as the final version of his works that he intended. For this reason, the *Weimarer Ausgabe*, (1887 onwards) set itself the goal of either preserving the versions of the *Ausgabe letzter Hand* or at least establishing them as the final goal of the philological work on the *Nachlass*[4]. However, the editors of the later *Akademie-Ausgabe*, pioneers in the application of textology to the reconstruction of the compositional history of Goethe's oeuvre, had noted that, in addition to Goethe's own changes, some variants had indeed been added posthumously by his collaborators without the poet's consent (Grumach, 1952). This congenital defect of the *Ausgabe letzter Hand* has led to a chain of philological unreliability and defies the possibility of adopting a genetic approach that leads to it as a final goal[5]. In the light of those problems, a new edition capable of going back to the manuscripts is needed in order to recover a complex and open-ended compositional history. Its provisionality is inscribed in the microgenetic and macrogenetic dimensions of the text (Golz, 1998) and seems to be linked not only to contingent reasons but also to a progressive, 'epigenetic' poetology that Goethe began to develop in the first months of 1790 (Zupancic, 2023-24) and that can be traced up to the second part of *Faust* (Bohnenkamp, 2024). Working on an extension of his morphological theory from plants to animals, the poet showed a growing interest in the domain of *infusoria* and invertebrates. These small creatures provided an uncharted territory for studying the laws of self-organisation and transformation of living organisms. Almost simultaneously, the poet began to compose the VE collection, which, as we have seen, is characterised by the metamorphosis of both its poems and its overall organisation. The influence of invertebrates, in particular molluscs and the freshwater polyp (*Hydra vulgaris*), on the genesis of the work can be traced through Goethe's original sources, letters, scientific annotations and a close reading of some of the earlier epigrams dedicated to the young acrobat Bettine (Zupancic, 2023-24). Consequently, Goethe's subtle comparison between the combinatory potential of the self-structuring ontology of the freshwater polyp and the similarly malleable form of the epigram must be addressed by the *VEdition* itself. In fact, the edition has to deal with a massive number of authorial alterations and their various combinations, while attempting to provide an interactive diplomatic rendition of challenging handwritten

---

a letter to his friend Knebel (H 60). Furthermore, the GSA also contains individual pencil sketches of a further seven epigrams, with unclear datation. The creation of a quarto notebook (H 56, GSA 25/W 61), written in ink in Goethe's own hand, followed. This notebook encompasses 106 complete epigrams and exhibits indications of projectuality. This number is to be supplemented with a short notebook (H 61) from the later Silesian Journey (July 1970), in which 19 new epigrams came to light. Following this preparatory work, the most significant and comprehensive document of the VE emerged: namely, a second quarto notebook (H 55, GSA 25/W62) containing the fair copy of 138 epigrams in Latin cursive, all written by Goethe in his own hand in ink after October 1790. This includes 6 epigrams from the Silesian notebook. H 55 also demonstrates the pivotal role played by Goethe's private circle of friends and colleagues in the stylisation and later reception of the VE as a whole. With the aim of publishing a selection of epigrams in Schiller's Musenalmanach (1796), the poems contained in H 55 underwent an extensive process of revision by Schiller's hand, documented by a series of pencil annotations on the document itself, and were subsequently published accordingly. Finally, a gift copy of the work (H 59, GSA 25/W 63), written by an unknown hand and addressed by Goethe to Duchess Anna Amalia, is of considerable philological importance, as it testifies to an extremely polished version of the VE, unrelated to the changes made for publication in the Musenalmanach.

[3] The process was initiated by Goethe himself from his first raw drafts in Venice (H 54), where the poet can be seen reconsidering and replacing potentially controversial expressions with milder ones. His opinion is documented by a letter to Schiller (26.10.1794), in which some of the epigrams are explicitly defined as "irreproducible". In particular, their eroticism (Wilson, 2012a, 2012b and 2015; Immer, 2013; Rohde & Valk, 2013; Zapperi, 2016) has had a significant impact on the reception of the work. In fact, beyond self-censorship, the VE encountered the even harsher resistance of external censorship at the hands of Schiller, and the subsequent interpolations by the editors of the *Weimarer Ausgabe*.

[4] The shortcomings of such an approach are manifold, but its main issue stems primarily from the point of view of the author himself, who looked back on his work from the perspective of his maturity and was ready to change it accordingly.

[5] The *Münchner Studienausgabe* provides compelling evidence of this, as it chose to reproduce the VE according to the number and arrangement of the previous, denser manuscript: H 55.

documents that form the backbone of the VE. Following the textological approach of the editors of the *Akademie-Ausgabe*, the present model aims at representing a complex series of alterations upon alterations on the handwritten document by generating a multinodal graph of operations, which stems from the base-layer and branches in connected sequences of alterations[6].

3. MODELING AUTOGRAPHS

The challenges posed by this scenario have both theorical and practical implications. On the theorical side, we need a model capable of representing the intrinsically dynamic nature of the creative process as reflected by our documents. This must be done both at the textual and at the visual level, while preserving the separation between the "objective" reproduction of the document and its subjective interpretation by the scholars in charge of reconstructing the level and order of changes. On the practical side, the essential idea is to provide a quick and user-friendly way of creating highly structured content with this model. In this context, the model and tools created for this project aim to provide a reusable content-creation paradigm, working as a complement rather than as a replacement for TEI-based representations. The highly structured and compact content thus created is typically exported into wider, more traditional data flows mostly based on specific TEI flavors. Once the model has been exported into some TEI flavor, this output could not be different from one created manually or with the aid of other tools and could thus join TEI-based flows using e.g. popular tools like EVT or production flows for digital editions like DiScEPT. This is another reason for the higher abstraction adopted by the model, which can easily adapt to different XML flavors, including those required for the integration of the project output with the larger constellation of Goethe's works. Once TEI becomes a software output, it is easier to provide many different flavors of it. In XML-based flows, we usually focus on a set of parallel TEI documents, carefully crafting the best encoding practices to produce an annotated text with metadata reflecting its transformations during the creative process. This produces a set of documents, each representing a specific version of the text, all connected via shared identifiers to the virtual root of all these offsprings. Often, these documents provide ultra-diplomatic representations of the text which face challenges connected to encoding strategies and overlap constraints. Just like standoff annotations for multiple structures, such practices may quickly increase the overall complexity of the documents and consequently the process of their creation and maintenance. In a context like *VEdition*, where human resources are limited, and additionally an authoritative edition of the text is missing, this represents a serious issue; and that's right this practical scenario, coupled with the desire to fully engage the possibilities offered by digital resources[7], which prompted the creation of the model presented here[8].

4. THE SNAPSHOT MODEL

The general idea here is focusing on the source of all the documents, adopting a generative strategy parallel to that underlying their representation. Rather than manually creating many versions of what we consider the same text, each with its set of annotations linking it to the source, we focus on their common source, defining it in operational terms. In *VEdition*, the source of epigrams is one or more autograph sheets from Goethe's handbooks. This material support, whether it is a portion, one, or more sheets referring to a single composition, represents our *carrier*. Of course, a carrier's content is far from being a linear text; it rather is a sort of *snapshot* of the creative process which in the author's mind would have led to the intended text. The carrier is thus the material support of the snapshot, which virtually contains many versions of a text, in a compact and often chaotic form, where annotations on a base text represent changes to it via editing operations like deletions, insertions, replacements, etc. While we can describe the surface of a carrier with its visuals in a relatively objective way, the main issue posed by it is the subjective reconstruction of the creative process as reflected by the snapshot it conveys. With all these annotations on top of a text, it is like having all the ingredients of a recipe on top of a table, while missing the recipe itself. The annotations, whatever their visual form, hint at changes to a text; but we can't be sure about how to select and order them to generate one or more versions of it. Defining this lost recipe based on its ingredients is right the task of the scholar, and of course it relies on his subjective judgement, while

[6] This is particularly useful for documents such as H 54, whose mobile material situation have prevented its complete transcription and is revealing new aspects of the work. The same goes for the allographic interpolations scattered throughout H 55 or parallel sources, such as some handwritten suggestions by August Wilhelm Schlegel.

[7] Cf. the considerations by Elena Pierazzo about the Proust prototype at http://epierazzo.blogspot.com/2012 (cons. 2024). A privileged reference for the design of this model was the TEI draft encoding model for Genetic Editions and Genetic Editing, here shortened with TCW19.

[8] For more information see https://vedph.github.io/gve-doc.

constrained by data provided by the snapshot[9]. To properly model a snapshot we thus want to represent it in a computable, generative way, focusing on the process which leads to text versions; in it, we want to preserve the distinction among our ingredients (the base text with its annotations), recipes (selection and ordering of editing operations), and outputs (the text versions); finally, we want to represent both the textual and the visual form of a snapshot. The graphical representation of each annotation (its visuals) can be encoded as both a possible presentation form and as a complement to its interpretation, thus allowing scholars to inspect all data and even provide their own alternative.

At the core of this model there is a general-purpose data structure, the *chain*[10], representing multiple linear combinations of a set of entities. These entities are the nodes of a graph, a structure which has already been proposed in many flavors to deal with complex text encoding issues with overlapping structures or digital authorial philology[11]. In our case, entities are characters, and their linear combinations build sequences representing the versions of a text. Just like in a snapshot text is laid on top of existing text, which whatever its destiny continues to occupy the space it was originally allotted, chain nodes once added are preserved. Each operation just links them in different ways, optionally introducing new nodes. Operations are thus the surface endpoints which change the state of the chain structure; and as they refer to human interpretation, they are not limited to just additions or deletions, but include many higher-level, user-friendly types, including replacements, movements or swaps. A simple, yet totally abstract example can be represented by the mock facsimile of Figure 1: here we just wrote the characters ARZDC, and then added some changes in form of annotations.

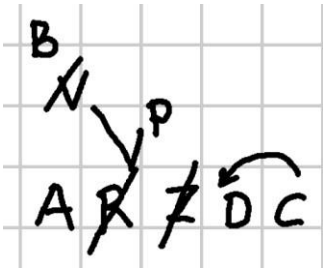


Figure 1. A mock snapshot

The visuals on this snapshot hint to operations like deletions (stroke on Z), replacements (stroke on R, replaced on one branch with V and then B, and on another one with P), and movements (C before D). Of course, it is up to the scholar's judgement to select and order them in some way and to define which outputs are to be considered as staged versions along the path leading from one text to another through all the intermediate steps defined by each operation's outcome. If we interpret this snapshot starting from ARZDC (v0 in Figure 2) going through deletion of Z, replacements R=V, V=B, R=P, and finally movement of C before D, we get 5 such "versions", among which zero or more can be defined as true, "staged" versions of the original text: ARDC (v1), AVDC (v2), ABDC (v3), APDC (v4), ABCD (v5).

[9] These are among the capital concepts also in TCW19: focus on process ("the genetic approach [...] aims not only to identify 'what is on the page', but also to reconstruct the process necessary to produce <it>"), and distinction between fact and interpretation, as the record ("Befund") is distinct from its interpretation ("Deutung"). Even if from a stricter point of view everything could be envisaged as interpretation, TCW19 points to a well-defined distinction between it and an "objective" plane, thus differentiating between "what's there" (document/fact) and "how does it relate" (text/interpretation).

[10] Technically this is a *tagged multigraph linked list*: *linked list*, because each node has at most one child in each version; *multigraph*, because it allows multiple edges between nodes; and *tagged*, because each version has a unique tag.

[11] The use of graphs for representing complex textual structures can be traced back to Colwell and Tune 1964, as pointed out by Elisa Nury during an exchange with Paolo Monella, to whom we owe this suggestion (more bibliography is given below). Nowadays, graphs for texts are mostly used for visualizing alignments from collations (like in TRAViz) and representing variants in the context of the more general problem of overlapping structures (e.g., Schmidt and Fiormonte). There, to represent multiple versions of a text, a graph is used where nodes are the void points across which links are drawn. It is the links which carry a text segment, with the identifiers of all the versions presenting it. This fits a scenario focused on a given linear text with multiple, parallel segments; in *VEdition* instead, where there is no authoritative final text and collation only happens above the level of the single snapshot, we rely on unordered sets of nodes and many subsets of links, one for each version, thus decoupling them from a specific linear sequence.

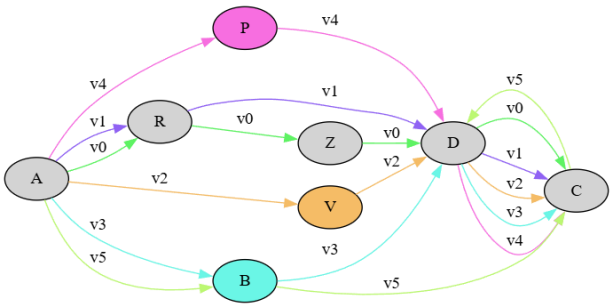


Figure 2. Chain model representing many versions of a text at the same time

Each version has its set of tagged links, defining a specific linear combination of the same set of nodes[12]. So, the generative model is very compact: you start from a base text and just add all the operations you reconstruct from the snapshot. Each operation starts with a specific step (thus also allowing complex branching in the reconstructed process) and generates a new one. Besides generating text, an operation has a model which allows it to inject an open set of metadata into either specific nodes or into the whole generated version. This allows us to output text with highly granular annotations, which can then be the starting point for exporting XML documents. The same model also provides a well-defined way for optionally representing on a separate layer also the visual part of the snapshot. Again, this stems directly from the nature of the snapshot itself: this being a set of operations on top of a base text, as suggested by visuals, we can just add diplomatic metadata to each operation, just like we do for textual metadata. The core here is represented by SVG code, reproducing a surrogate visual representation of any details found on the snapshot; each SVG element can even get additional metadata for further injections. Also, we can add to this 2D SVG representation the third dimension of time via animations, defined in a declarative way by leveraging a more abstract layer, based on a GSAP-oriented implementation. This allows us to interpolate between each step defined by operations, transforming a set of stills into a sort of a movie.

5. ARCHITECTURE AND INTEGRATION

This highly abstract and structured model requires a way for easily creating content based on it. The model has in fact been designed on par with an open-source software editor providing many ways for entering data in that form, ranging from full-fledged UI to DSL-based entry methods[13]. On the user side, software is fully containerized for easy distribution. On the developer side, it has been designed for integration into third-party environments, by making it highly modular and using different technologies according to their layer. For instance, at the lowest level most of the complex logic for visualizing a snapshot from its data is implemented by a pure custom web component, in turn adopted by an Angular-based library for editing, in turn wrapped in a Cadmus-based UI infrastructure[14], with the corresponding backend components and an underlying, JSON-based data store. This way, a highly compact and structured graph-based model, with separate layers for "objective" and subjective interpretations, text and their visuals, and even animations can be easily created on a web UI, and export any number of deeply annotated documents, whatever their format. This makes this solution potentially reusable for creating content capable of taking the way of existing and popular TEI-based production flows, fostering a higher integration among them.

[12] Each node in this graph represents a single character in the chain, but this does not have any implications for the granularity of markup which can be generated from the model. The chain itself is a templated class, meaning that its textual content can be made a character as well as a string (or even more complex objects); the UI provided for creating snapshots allows to freely select ranges of characters to act on them, just like on a word processor; and the export process leverages a dynamic segmentation method, designed to maximize the extent of each segment carrying the same types and values of annotations. So, while exporting each segment of text will be built by merging all the subsequent characters targeted by the same set of annotations, which can be freely selected among those provided by the model.

[13] For a real-world example, see https://vedph.github.io/gve-doc/model/limerick.html.

[14] See e.g. https://vedph.github.io/cadmus-doc. The choice of Cadmus here is mainly due to its open-ended nature and its export subsystem, coupled with the requirement of editing a full hierarchy of higher-level entities on top of the snapshot model, representing epigrams and collections in the context of a full-fledged critical edition.

ACKNOWLEDGEMENTS

The *VEdition* is a project financed and hosted by the Istituto Italiano di Studi Germanici, in cooperation with the Stuttgart Research Center for Text Studies, the Venice Centre for Digital and Public Humanities, the BemboLab at Ca' Foscari University of Venice and the Klassik Stiftung Weimar.

REFERENCES

Andrews, Tara & Diehr, Franziska & Efer, Thomas & Kuczera, Andreas & Zundert, Joris van (ed.). 2022. Graph Technologies in the Humanities - Proceedings 2020 CEUR Workshop Proceedings. https://ceur-ws.org/Vol-3110/

Bohnenkamp, Anne. 2024. "'Gestaltenlehre ist Verwandlungslehre': Zur 'genetischen Methode' als einem interdisziplinären Schlüsselkonzept Goethes." Publications of the English Goethe Society 93:34-51. https://doi.org/10.1080/09593683.2024.2308931

Burnard, Lou. 2013. "Resolving the Durand Conundrum." Journal of the Text Encoding Initiative 6. https://doi.org/10.4000/jtei.842.

Buzzoni, M. & Cucurnia, D. & Fenu, C. & Rosselli Del Turco, R. & Tancredi, G. 2024. Progetto di edizione genetica digitale del Canzon iere manoscritto di U. Saba (1919-20). In Me.Te. Digitali. Mediterraneo in rete tra testi e contesti, edited by Antonio Di Silvestro and Daria Spampinato, 215-220. https://hdl.handle.net/10278/5078681.

Colwell, E.C. & Tune, E.W. 1964. "Variant Readings: Classification & Use." Journal of Biblical Literature 83 (3): 253-261. https://doi.org/10.2307/3264283.

Cummings, James 2018. "A world of difference: Myths and misconceptions about the {TEI}." Digital Scholarship in the Humanities. https://doi.org/10.1093/llc/fqy071.

Genovesi, Chiara & Isolani, Alida & Lo Rito, Claudia & Marotta, Daniele & Matteoli, Marco & Tozzinni, Cinzia 2009. "Topic Maps and MVD for the Representation of Interpretative Variants." In Digital Humanities 2009. Conference Abstract, University of Maryland, College Park, June 22-25, 2009. The Maryland Institute for Technology in the Humanities (MITH). https://dhabstracts.library.virginia.edu/works/948.

Golz, Jochen 1998. "Alle Ordnung ist vorläufig. Über den Zusammenhang von Textgenese und Entstehungskontext in Goethes *Venezianischen Epigrammen*." Editio 12: 69-78. https://doi.org/10.1515/9783484604230.69

Golz, Jochen & Gothe, Rosalinde (ed.). 1999. "Venezianische Epigramme: eigenhändige Niederschriften, Transkription und Kommentar."

Grumach, Ernst. 1952. "Aufgaben und Probleme der modernen Goetheedition." In Dokumente zur Geschichte der neugermanistischen Edition, edited by Rüdiger Nutt-Kofoth, 2005, S. 154-161. https://doi.org/10.1515/9783110926927.154

Immer, Nikolas. 2013. "Die Götter Italiens. Goethes mythoerotische Elegien." In Goethes Liebeslyrik. Semantiken der Leidenschaft um 1800, edited by Carsten Rohde and Thorsten Valk, 107-124. https://www.degruyterbrill.com/document/doi/10.1515/9783110312041.107/html

Pierazzo, Elena & Rehbein, Malte & Galley, Amanda. "An Encoding Model for Genetic Editions.", last accessed September 13, 2024 at https://tei-c.org/Vault/TC/tcw19.html.

Oswald, Stephan. 2014. "Früchte einer großen Stadt – Goethes *Venezianische Epigramme*."

Rohde, Carsten & Valk, Thorsten (ed.). 2013. "Goethes Liebeslyrik. Semantiken der Leidenschaft um 1800."

Sahle, Patrick. 2016. "What Is a Scholarly Digital Edition?" In Digital Scholarly Editing: Theories & Practices, edited by Matthew James Driscoll & Elena Pierazzo, 19-40. Open Book Publishers. https://doi.org/10.11647/OBP.0095.02.

Schmidt, Desmond & Colomb, Robert. 2009. "A data structure for representing multi-version texts online." International Journal of Human-Computer Studies 6 (67): 497-514. https://doi.org/10.1016/J.IJHCS.2009.02.001.

Schmidt, Desmond. "Merging Multi-Version Texts: a Generic Solution to the Overlap Problem." Presented at Balisage: The Markup Conference 2009, Montréal, Canada, August 11 - 14, 2009. In Proceedings of Balisage: The Markup Conference 2009. Balisage Series on Markup Technologies, vol. 3 (2009). https://doi.org/10.4242/BalisageVol3.Schmidt01.

Schmidt, Desmond & Fiormonte, Domenico. 2010. "Multi-Version Documents: A Digitisation Solution For Textual Cultural Heritage Artefacts." Intelligenza Artificiale 1 (4): 56-61.

Schmidt, Desmond & Eggert, P. 2019. "The Charles Harpur Critical Archive." International Journal of Digital Humanities. https://doi.org/10.1007/s42803-019-00021-9.

Turska, Magdalena & Cummings, James & Rahtz, Sebastian. 2016. "Challenging the myth of presentation in digital editions." Journal of the Text Encoding Initiative 9. https://doi.org/10.4000/jtei.1453.

Wilson, W. Daniel. 2012a. "Masturbation, Prostitution, Sodomy: The Imagination and Non-Reproductive Sexuality in Goethe." Colloquia Germanica 45 (3/4), Themenheft: Triangular Readings, edited by Martin Kagel and Alexander Sager: 221-237.

Wilson, W. Daniel. 2012b. "Goethe Männer Knaben. Ansichten zur 'Homosexualität'."

Wilson, W. Daniel. 2015. "Goethes Erotica und die Weimarer 'Zensoren'."

Zapperi, Roberto. 2016. "Eros a Venezia. Gli Epigrammi Veneziani di Goethe." In L'Italia di Goethe, edited by Marino Freschi, 79-103.

Zittel, Claus. 2023. "«…denn was ich berühre, / Wird mir unter der Hand gleich ein behendes Gedicht». Textuelle Metamorphosen und poetisches Kalkül in Goethes Venezianischen Epigrammen." In Verwandlung der Worte. Textuelle Metamorphosen in Goethes Schriften: Fassungen, Ausgaben, Übersetzungen, edited by Gabriella Catalano and Giovanni Sampaolo, 167-191.

Zupancic, Matteo. 2023-2024. "L'enérgeia dell'organismo letterario. Goethe, Venezia e i prodromi della Gattungstheorie morfologica." Studi Germanici – I quaderni dell'AIG 6: 211-226. https://www.studigermanici.it/wp-content/uploads/2025/03/Q_AIG6_Zupancic.pdf

Computational Methods

The Influence of AI Tools on University Students' Writing Style: A Stylometric Analysis of Narrative Texts

Dimitris Bilianos1, Katerina Florou2,
1 National and Kapodistrian University of Athens, GREECE – DBILIANOS@ILL.UOA.GR
2 National and Kapodistrian University of Athens, GREECE – kathyflorou@ill.uoa.gr

ABSTRACT (ENGLISH)

This study investigates the impact of ChatGPT on the writing styles of university students learning Italian as a foreign language. Using a subset of the UniC corpus, which includes narrative texts from 2023 (produced with ChatGPT assistance) and 2024 (produced without AI tools under supervised conditions), we employed stylometric techniques such as Principal Component Analysis (PCA) and K-means clustering to analyze stylistic patterns. Results revealed that while stylistic groupings were evident, they did not align distinctly with ChatGPT usage or year of authorship. Texts from both years appeared in overlapping clusters, suggesting that individual writing habits and pedagogical influences may obscure the stylistic impact of AI assistance. These findings underscore the complexity of distinguishing AI-generated content from human-authored texts and highlight the nuanced interplay between AI tools and student writing practices. Limitations, including a small dataset and variability in ChatGPT usage, are discussed, along with suggestions for future research to further explore AI's role in educational contexts.
Keywords: Stylometric Analysis, AI-Generated Texts, Writing Style in Higher Education

ABSTRACT (ITALIANO)

L'impatto degli strumenti di intelligenza artificiale sullo stile di scrittura degli studenti universitari: Un'analisi stilometrica di testi narrativi

Questo studio analizza l'impatto di ChatGPT sullo stile di scrittura di studenti universitari che apprendono l'italiano come lingua straniera. Utilizzando un sottoinsieme del corpus UniC, che comprende testi narrativi del 2023 (redatti con l'assistenza di ChatGPT) e del 2024 (prodotti senza strumenti di intelligenza artificiale in condizioni supervisionate), sono state applicate tecniche stilometriche come l'Analisi delle Componenti Principali (PCA) e il clustering K-means per esaminare i modelli stilistici. I risultati mostrano che, sebbene siano emerse alcune aggregazioni stilistiche, queste non corrispondono in modo netto all'uso di ChatGPT o all'anno di composizione. Testi di entrambi gli anni si trovano in cluster sovrapposti, suggerendo che abitudini di scrittura individuali e influenze pedagogiche possano offuscare l'impatto stilistico degli strumenti di intelligenza artificiale. Questi risultati evidenziano la complessità di distinguere i contenuti generati dall'AI da quelli scritti interamente dagli studenti, sottolineando l'interazione sfumata tra strumenti di AI e pratiche di scrittura accademica. Sono discusse le limitazioni dello studio, inclusa la dimensione ridotta del campione e la variabilità nell'uso di ChatGPT, insieme a suggerimenti per ricerche future volte a indagare più a fondo il ruolo dell'intelligenza artificiale nei contesti educativi.
Parole chiave: Analisi Stilometrica, Testi Generati dall'Intelligenza Artificiale, Stile di Scrittura nell'Istruzione Universitaria

1. INTRODUCTION

Recent advancements in artificial intelligence (AI) have sparked increasing interest in their potential to transform educational environments. Among these, OpenAI platforms, particularly ChatGPT, have garnered significant attention for their role in enhancing distance and online learning (Bozic & Poola, 2023; Firat, 2023). The growing incorporation of AI tools in education underscores their potential to reshape teaching methodologies and learning experiences, offering new opportunities for both educators and students. A specialized body of research is emerging that focuses on ChatGPT's application as a natural language processing (NLP) tool in higher education, shedding light on its capacity to support academic research and its broader implications for pedagogical practices (Fuchs, 2023).

As students increasingly integrate ChatGPT into their language learning routines, they employ it for a wide range of tasks. These include generating original text, clarifying definitions, verifying spelling, and exploring complex syntactic structures. This shift suggests a significant departure from traditional language resources, with ChatGPT serving as a modern alternative to conventional tools such as dictionaries, grammar guides, and automated translation services like Google Translate (Aydin, 2024; Baskara & Mukarto, 2023). Moreover, ChatGPT's interactive nature offers learners an engaging platform to

practice and refine their speaking and comprehension skills, fostering a low-stress environment where students can freely experiment with vocabulary and grammar without fear of judgment from peers (Klimova et al., 2024; Vo & Nguyen, 2024).
However, despite its numerous advantages, ChatGPT is not without limitations. Ethical concerns have emerged, particularly regarding its potential for misuse in academic settings. One major issue is the risk that students may rely on ChatGPT to complete assignments by copying and pasting generated responses, circumventing academic integrity and teacher oversight (Nugroho et al., 2023). This limitation is central to our investigation, as it raises questions about how the tool may influence students' writing practices, particularly when used for both primary production and as a supplementary aid. As AI tools like ChatGPT continue to be integrated into academic practices, it also becomes increasingly important to develop reliable methods for distinguishing between AI-generated and human-written content (see [Author 1] among others).
In this study, we aim to explore how ChatGPT affects the style of students' writing, focusing on the potential shifts in language use, structure, and originality when it is employed as either a primary writing tool or an assistive resource.

2. CORPUS DEVELOPMENT AND DATA PRESENTATION

Each sub-corpus represents a small collection of texts developed over the course of an academic year as part of a larger corpus, UniC (University Corpus). UniC was created primarily for studies on Greek learners of Italian as a foreign language and their interlanguage development (Author 2). The project aimed to document the written language of university students and future teachers in Italian as a foreign language. The topics and sizes of the sub-corpora vary, but the texts consistently adopt a narrative genre each year. For this study, we selected data from the last two academic years.

Sub Corpus	N. Words	N. Learners	Use of online tools (for corrective purposes)
UniC 2021	42.751	43	Google translate
Unic 2022	27.069	23	Google translate
Unic 2023	31.215	31	Chat GPT
Unic 2024	4.214	7	NONE

Table 1: Uni Corpus

For the purposes of our study, we analyzed a set of short narratives written by students in two distinct years: 2023 and 2024. The purpose was to explore stylistic patterns and determine whether texts written with the aid of large language models, like ChatGPT, differ stylistically from those written without such assistance.
The dataset consists of narratives written by students across two different years. In 2023, a total of 31 narratives were produced, while in 2024, only 7 narratives were collected. A key distinction between these two years lies in the conditions under which the narratives were composed. In 2023, students were encouraged to use ChatGPT to assist in their writing, resulting in texts that potentially blend human and machine-generated stylistic features. By contrast, in 2024, students were explicitly instructed not to use any assistance from large language models (LLMs). These narratives were written at the university under supervised conditions, ensuring they reflect purely human-written styles.
All texts were saved in plain .txt format to maintain consistency across both years. Text files from 2023 were labeled sequentially (e.g., 001.txt, 002.txt), while those from 2024 included the year in their filenames (e.g., 2024-001.txt). This labeling system facilitated easy differentiation between the two datasets.

While students in 2023 were encouraged to use ChatGPT as an assistive tool, the extent and nature of their interaction with the AI (e.g., for generating entire drafts, for specific suggestions, for editing) were

not systematically recorded. Similarly, the specific prompts used by students were not collected, which introduces a degree of variability in the AI-assisted texts.

3. PURPOSE OF THE ANALYSIS

For this analysis, a set of word n-grams (specifically, unigrams and bigrams) were extracted from the texts and weighed using Term Frequency-Inverse Document Frequency (TF-IDF). This method assigns a higher weight to words that appear frequently in a specific text but infrequently across the entire corpus, thus highlighting terms that are characteristic of individual writing styles. We considered the top 1000 most salient of these TF-IDF weighed n-grams as features for our analysis. These features were chosen as word usage patterns, including combinations of adjacent words, which are commonly used in stylometric studies to capture variations in writing style.

To investigate through clustering whether stylistic differences in narrative writing can be attributed to the use of LLMs such as ChatGPT, we made use of the following techniques:

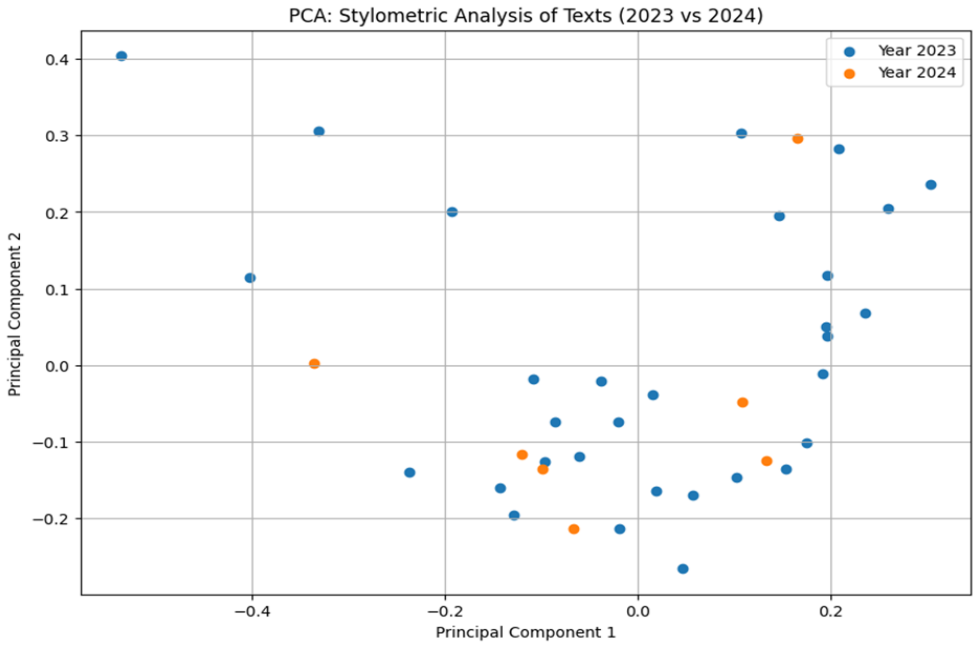
- Principal Component Analysis (PCA) for dimensionality reduction and visualization of stylistic variance.
- K-means clustering to identify groupings of texts based on stylistic similarities or differences.

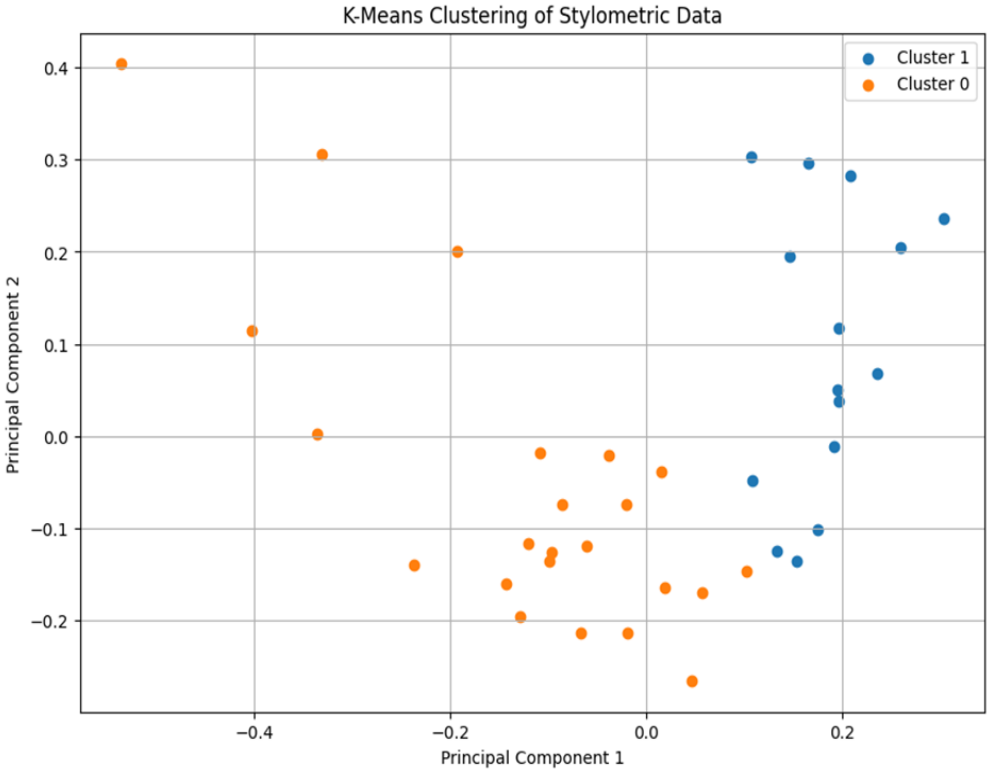
While Principal Component Analysis (PCA) and K-means clustering are machine learning techniques, they were applied here to analyze patterns within a set of extracted linguistic features relevant to writing style. Our approach aimed to identify potential groupings based on the overall distribution of these features across the texts.

PCA is a statistical technique used to reduce the dimensionality of datasets while preserving as much variance as possible. It transforms the original data into a set of new, uncorrelated variables called principal components. The first principal component captures the greatest variance in the data, while the second principal component captures the next highest variance, and so on. By projecting high-dimensional data onto these components, PCA simplifies the visualization and interpretation of complex datasets (Jolliffe, 2002). In this analysis, PCA enables us to map the stylistic features of the texts onto a two-dimensional space, making it easier to visually inspect potential groupings or patterns based on their stylistic variance.

K-means clustering is an unsupervised machine learning algorithm used to partition data into distinct clusters. The algorithm works by initializing a set number of cluster centroids and then iteratively assigning data points to the nearest centroid based on their distance, recalculating the centroids until the optimal configuration is reached. The goal is to minimize the variance within each cluster while maximizing the variance between clusters. In this study, K-means clustering allows us to group the narratives based on their stylistic features, helping to identify potential differences in style between the 2023 and 2024 texts.

4. RESULTS





Based on the clustering results, it appears that the stylistic differences between the 2023 and 2024 texts do not strongly correlate with the use of ChatGPT. While the K-means clustering algorithm divides the texts into two clusters, these clusters do not strictly correspond to the year of authorship or the use of AI assistance. Both 2023 and 2024 texts are present in both Cluster 0 and Cluster 1, indicating that stylistic similarities and differences are distributed across years rather than being exclusively tied to whether ChatGPT was used.

Cluster 1 contains a higher proportion of 2023 texts, suggesting that some texts written with ChatGPT share certain stylistic features that were captured by the clustering. However, Cluster 0 also includes a substantial number of 2023 texts, which highlights variability in how students used or edited AI-assisted content. Interestingly, some 2024 texts also appear in Cluster 1, implying that purely human-written texts can sometimes resemble AI-generated styles. This could be due to individual writing habits, common educational influences, or even shared narrative structures that students employed across both years.

The overlap between the clusters suggests that the stylistic impact of ChatGPT is not uniform or easily distinguishable from human writing through clustering alone. Instead, the clustering may reflect broader stylistic differences, such as variations in sentence complexity, use of specific vocabulary, or narrative structure, which are present in both AI-assisted and non-AI-assisted texts.

In summary, the clustering results demonstrate that while there are detectable stylistic groupings, these do not clearly separate the texts based on ChatGPT usage. This indicates that other factors, such as individual writing style or pedagogical influences, may play a significant role in shaping the stylistic characteristics of the narratives.

5. CONCLUSION

The analysis of the narratives from 2023 and 2024 aimed to explore whether the use of ChatGPT resulted in identifiable stylistic differences in the students' writing. Through stylometric techniques, including Principal Component Analysis (PCA) for dimensionality reduction and K-means clustering for grouping, we examined patterns across the texts. While there were observable stylistic groupings, these clusters did not distinctly separate texts based on the use of ChatGPT or the year of authorship. Both AI-assisted (2023) and non-AI-assisted (2024) texts were present in both clusters, suggesting that the impact of ChatGPT on writing style is neither uniform nor easily distinguishable from human-authored content. The findings highlight the complexity of attributing stylistic features to the use of AI, as individual writing habits, narrative structures, and pedagogical influences may overshadow the direct effects of AI assistance.

This study underscores the importance of analyzing the nuanced relationship between AI-assisted and human-written texts. While some stylistic trends were observed in 2023 texts, likely influenced by ChatGPT, the overlap between years and clusters suggests that human intervention, such as editing AI outputs, can blur distinctions.

Even though these results contribute to ongoing discussions about the role of AI in education and writing, particularly regarding its impact on students' stylistic development, it is important to note that our use of PCA and K-means clustering served to explore patterns in the distribution of extracted linguistic features, offering a broad overview of potential stylistic differences rather than a deep dive into specific stylistic markers.

6. LIMITATIONS

Despite the insights gained, this study has several limitations. First, the dataset is relatively small, with only 31 texts from 2023 and 7 from 2024. This limited sample size may reduce the robustness of the clustering analysis and the generalizability of the findings. A larger and more diverse dataset might reveal clearer patterns or more nuanced stylistic differences. Second, the analysis did not consider deeper semantic or syntactic structures. Incorporating advanced linguistic features or additional computational techniques could provide more comprehensive insights.

Another limitation lies in the variability of how students may have used ChatGPT in 2023. Some may have relied heavily on AI-generated outputs, while others might have used it sparingly or extensively edited its suggestions. This heterogeneity complicates efforts to draw clear stylistic distinctions. Furthermore, while 2024 texts were written without AI assistance under supervised conditions, individual writing habits and external influences may still have contributed to stylistic overlap with 2023 texts.

A significant limitation of this study is the lack of detailed content analysis of the narratives. It is possible that the clustering patterns observed are, to some extent, driven by differences in the topics or themes explored by students in 2023 and 2024, rather than solely by variations in their writing style due to AI assistance. Future research should consider incorporating content analysis methodologies to control for this potential confounding factor.

These limitations suggest avenues for future research, including larger datasets, expanded stylistic features, and more nuanced clustering methods, to deepen our understanding of the intersection between AI and human authorship. More specifically, future research could benefit from incorporating dedicated stylometric tools to extract a wider range of linguistic features before applying machine learning techniques. This could provide a more granular and potentially more accurate stylistic analysis. However, preliminary data suggests that while stylometric methods can identify patterns, the boundaries between AI and human stylistics are not always clear-cut.

ACKNOWLEDGEMENTS

We would like to express our gratitude to the students of the Department of Italian Language and Literature for granting us permission to use their written works. We are also grateful to the personnel and the director of the Laboratory of Linguistic Analysis and Computational Processing of Romance Languages, Prof. Gianna Giannoulopoulou. We would like to thank the anonymous reviewers for their comments.

REFERENCES

Aydin, S. (2024). Using ChatGPT in Foreign Language Research: An Overview. DOI: https://doi.org/10.29329/irelt.2024.1045.2

Baskara, R., Mukarto, M. (2023). Exploring the Implications of ChatGPT for Language Learning in Higher Education. *Indonesian Journal of English Language Teaching and Applied Linguistics*, 7(2), 343- 358

Božić V, Poola I. (2023). ChatGPT and education. *Preprint*.. https://www.researchgate.net/profile/Velibor-Bozic-2/publication/369926506_Chat_GPT_and_education/links/64350844ad9b6d17dc4d3a79/Chat-GPT-and-education.pdf

Firat M. (2023). How ChatGPT can transform autodidactic experiences and open education?. https://osf.io/preprints/osf/9ge8m

Florou, K. (2024). The Effect of English as a Second Foreign Language On Learning Italian as a Third Foreign Language: A learner corpus-based research in written speech. *International Journal of Language and Literary Studies*, 6(3), 407-419.

Fuchs K. (2023). Exploring the opportunities and challenges of NLP models in higher education: is ChatGPT a blessing or a curse?. In *Frontiers in Education* 2023 May 17 (Vol. 8, p. 1166682). Frontiers Media SA.

Jolliffe, I. T. (2002). Principal Component Analysis. Springer.

Klimova B, Pikhart M and Al-Obaydi LH (2024). Exploring the potential of ChatGPT for foreign language education at the university level. Front. Psychol. 15:1269319. doi:10.3389/fpsyg.2024.1269319

Mikros, G. K., Koursaris, A., Bilianos, D., & Markopoulos, G. (2023). AI-Writing Detection Using an Ensemble of Transformers and Stylometric Features. In *IberLEF*@ SEPLN.

Nugroho, A., Putro, N. H. P. S., & Syamsi, K. (2023). The potential of ChatGPT for language learning: Unpacking its benefits and limitations. *Register Journal*, 16(2), 224-247.

Usare i Large Language Model per l'analisi del testo narrativo: strategie di prompt engineering per il riconoscimento del discorso indiretto libero nella narrativa italiana 1830-1930

Aurora Argenzio, Fabio Ciotti, Anna Chiara Corradino
Università di Roma -Tor Vergata, Italia
argenzioaurora@gmail.com; fabio.ciotti@uniroma2.it; annachiara.corradino1@gmail.com

ABSTRACT (ITALIANO)

Questo studio presenta i risultati di un esperimento di prompt engineering condotto nell'ambito del progetto "Leggere il romanzo italiano a distanza (1830-1930) (RIND)". Il progetto intende utilizzare metodi computazionali per rivalutare la periodizzazione tradizionale della letteratura italiana, concentrandosi in particolare sul fenomeno del discorso indiretto libero (DIL) come marcatore di cambiamento stilistico e narrativo del modernismo italiano. RIND ha predisposto un corpus di 1.000 testi letterari, di cui 500 romanzi italiani originali e 500 traduzioni. Per il presente esperimento è stato selezionato un sottoinsieme di 100 testi, bilanciando anno di pubblicazione e lunghezza dei testi. La metodologia ha combinato l'annotazione manuale di circa 3.000 frasi con test sistematici su modelli GPT-4 e Claude 3.5 Sonnet. L'esperimento ha utilizzato strategie di prompting, confrontando approcci zero-shot e *few-shot*, (con e senza spiegazioni preliminari del fenomeno linguistico, sia in italiano che in inglese). I risultati mostrano che i prompt few-shot ottengono le migliori prestazioni nell'identificazione del DIL. Abbiamo anche rilevato che la lingua dei prompt nell'analisi dei testi originali italiani non produce oscillazioni significative nei risultati sull'identificazione di DIL. La lunghezza ottimale del testo per l'analisi si è stabilizzata tra i 1.800 e i 5.000 caratteri. La tokenizzazione per *sentence* si è rivelata l'approccio più efficace per rilevare i marcatori linguistici e sintattici del DIL. Nel 50% dei casi di errore, le interpretazioni alternative proposte dai modelli erano comunque contestualmente valide, mostrando la complessità e le capacità dell'analisi letteraria automatizzata. Lo studio intende fornire un'ipotesi iniziale di framework metodologico per l'analisi di tecniche narrative attraverso modelli linguistici. I risultati suggeriscono l'importanza di elaborare una tassonomia chiara, ma flessibile, dei prompt utilizzabili per l'analisi dei testi.
Parole chiave: prompt engineering; discorso indiretto libero; analisi narrativa computazionale; Large Language Models; Romanzo Italiano

ABSTRACT (ENGLISH)

Using Large Language Models for Narrative Text Analysis: Prompt Engineering Strategies for Identifying Free Indirect Speech in Italian Fiction (1830–1930)

This abstract presents the findings of a prompt engineering experiment conducted as part of the "Reading the Italian Novel at a Distance (1830–1930) (RIND)" project. The project applies computational methods to reassess traditional periodisations of Italian literature, focusing on free indirect speech (FIS) as a marker of stylistic and narrative shifts in Italian Modernism. RIND includes a corpus of 1,000 texts—500 original Italian novels and 500 translations. For this specific experiment, a subset of 100 texts was selected, balanced by publication year and length. The methodology combined manual annotation of 3,000 sentences with systematic tests on GPT-4 and Claude 3.5 Sonnet using zero-shot, one-shot, and few-shot prompting strategies, with and without prior explanations of the phenomenon in Italian and English. Few-shot prompts performed best in identifying FIS, and the language of the prompts did not significantly affect results. Text lengths between 1,800 and 5,000 characters proved optimal for analysis, with sentence tokenization yielding the most accurate detection of FIS's linguistic and syntactic markers. In 50% of errors, the models' alternative interpretations were still contextually valid, highlighting the complexity and potential of automated literary analysis. This study emphasizes the need for a clear yet adaptable taxonomy of prompts to enhance text analysis.
Keywords: prompt engineering; Free Indirect Speech; LLMs; Italian Novel; computational narratology.

1. INTRODUZIONE

L'introduzione dei modelli linguistici di grandi dimensioni (Large Language Model, o LLM) generativi, oltre a costituire un sostanziale avanzamento nel campo dell'Intelligenza Artificiale, rappresenta un'occasione di profonda innovazione per gli studi letterari computazionali. In linea generale i modelli generativi, a

differenza dei sistemi di reti neurali discriminativi, producono nuovi dati sulla base della distribuzione di probabilità dei dati di input. Nei modelli linguistici questa modalità operativa si traduce nella capacità di produrre sequenze linguistiche sintatticamente e semanticamente coerenti con una sequenza testuale di input, detta *prompt*. Con opportune tecniche di training, gli LLM si sono dimostrati in grado di rispondere in modo appropriato a domande o istruzioni, rivelando una imprevista capacità di adattarsi a task complessi non esplicitamente previsti nella fase di addestramento, soprattutto se la formulazione dei prompt viene progettata in modo adeguato (Ciotti, 2023). Di conseguenza negli studi sulle proprietà degli LLM ha assunto rilevanza la definizione di tecniche di progettazione dei prompt, o *prompt engineering* (Mollick, 2023; Wei et al., 2023).
Questo lavoro intende presentare i risultati di una sperimentazione nell'uso degli LLM per condurre sofisticate analisi di testi narrativi, spesso eccedenti le possibilità offerte delle tecniche di machine learning tradizionali (Byszuk et al., 2020), concentrandosi in particolare sulle strategie di *prompt engineering* e *prompt implementation*. La nostra sperimentazione si colloca nell'ambito del progetto di ricerca "Leggere il romanzo italiano a distanza (1830-1930)" (RIND) finanziato dal programma PRIN2022 del MUR. L'obiettivo generale del progetto è quello di utilizzare metodi computazionali e quantitativi per rivalutare e validare sulla base di evidenze empiriche la tradizionale periodizzazione della letteratura italiana dal 1800 ai primi anni del 1900. L'arco temporale scelto abbraccia l'emergere del romanzo in Italia, il dominio del romanzo storico e realista e l'ascesa del Modernismo. A questo fine sono state individuate due tipi di features testuali che la storiografia e la critica letteraria hanno considerato fondamentali per tracciare i cambiamenti diacronici nelle strategie narrative e dunque nei periodi letterari: la prima è la presenza caratterizzante del discorso/pensiero riportato (DR) e del discorso indiretto libero (DIL), una classe di fenomeni morfosintattici che gli studi indicano come strategie discorsive volte a far emergere la coscienza e il pensiero dei personaggi riducendo la mediazione del narratore; la seconda riguarda l'ambientazione narrativa e i ruoli assunti dagli attori umani all'interno dei testi, con l'obiettivo di studiare il riflesso dell'evoluzione sociale nelle transizioni tra i periodi letterari. In questa sede ci concentriamo sulle analisi condotte sul primo insieme di fenomeni, illustrando i risultati preliminari emersi da numerosi test di prompt engineering condotti sistematicamente sui campioni testuali scelti – sulla base di preciso bilanciamento - nell'ambito del corpus di riferimento del progetto. Gli LLM adottati nei test sono i modelli più capaci e performanti attualmente disponibili (sebbene gli esperimenti siano stati condotti anche su altri modelli): Open AI GPT-4o e Claude 3.5 Sonnet.

2. STATO DELL'ARTE

La periodizzazione storico-critica delle correnti letterarie è da sempre oggetto di accese controversie. René Wellek scriveva che "the concept of period is certainly one of the main instruments of historical knowledge" (Wellek, 1956: 268), tuttavia, la cultura opera come un sistema dinamico, rendendo problematica l'imposizione di rigidi confini di periodizzazione temporale. Le sfide sono ulteriormente complicate dai limiti epistemologici del tradizionale close-reading, che per lungo tempo ha foraggiato i tentativi di periodizzazione degli studiosi di letteratura. Negli ultimi decenni, l'emergere di metodologie quantitative ha arricchito gli studi letterari, portando all'adozione di approcci come il distant-reading, introdotto da Franco Moretti (2013; Ciotti, 2022a), che sfidano le consuete periodizzazioni letterarie, come dimostrato da Moretti (2005), Jockers (2013: cap. 6), Piper (2018: cap. 4), Underwood (2019) per la narrativa inglese, Jannidis e Lauer (2014) per quella tedesca e Ciotti (2022b) in un esperimento sulla letteratura italiana. Nell'ambito specifico della storia letteraria italiana, critici e storici di spicco (Castellana, 2010; Donnarumma, 2012; Luperini, 2018; Tortora, 2018; Cangiano, 2018) hanno tradizionalmente, sebbene variamente, datato l'emergere del Modernismo italiano all'inizio del XX secolo. Tra le molte caratteristiche che segnano questa transizione, il discorso indiretto libero (DIL) spicca come particolarmente significativo; Il DIL è uno dei tipi di Discorso Riportato (DR) il cui scopo è quello di fondere la voce del narratore con quella del personaggio: esso è un tratto distintivo della letteratura europea di inizio Novecento, che riflette una crescente attenzione alla profondità psicologica e all'interiorità dei personaggi narrati. Tuttavia, la presenza di DIL che precedono la canonica periodizzazione letteraria mette in discussione l'idea di una netta divisione tra Verismo e Modernismo, suggerendo un'evoluzione stilistica più graduale che può essere computata proprio attraverso l'analisi della sua presenza in testi che precedono l'inizio del Novecento. Basandosi su intuizioni teoriche (Calaresu, 2000; Sullet-Nylander et al., 2014) sul DIL e sulla distinzione dei diversi DR, la nostra sperimentazione ha inteso esplorare le capacità dei modelli linguistici di individuare le occorrenze di DIL in un corpus di romanzi e racconti italiani. Dopo una fase di test preliminari condotti su GPT-4o e su Claude 3.5 Sonnet, su porzioni di testo variabili tra i

1.800 e i 5.000 caratteri, abbiamo deciso di valutare in modo sistematico varie strategie di prompt engineering (Bommasani et al., 2021; Liu et al., 2021) elaborate nella ricerca più recente sugli LLMs (Bach et al. 2022).
Il concetto di prompting è generalmente definito come il processo di istruzione di una IA generativa affinché esegua un compito ben preciso. Per "prompt engineering" (o prompt design) intendiamo qui "the activities of tailoring a prompt to a specific task and utilizing the possibilities of the respective generative AI model" (Böhmker et al., 2023: 559). Data l'assenza di modelli fissi per la progettazione dei prompt, la facilità di formularli attraverso il linguaggio naturale e soprattutto la velocità di cambiamento dei modelli utilizzati, la progettazione dei prompt è un processo soggetto a un notevole grado di arbitrarietà (Zamfirescu-Pereira et al., 2023). Recenti studi hanno fatto emergere l'importanza delle componenti singole dei prompt e il loro impatto sui risultati prodotti (Liu et al., 2021) e mostrato l'inefficacia di prompt molto lunghi e mancanti di concatenazione logico-consquenziale (Wu et al., 2022). Pertanto, la classificazione tassonomica dei prompt (come quelle di Böhmker et al., 2023; Kundisch et al., 2021; Nickerson et al., 2013) risulta un punto di partenza necessario e utile per ottenere risultati soddisfacenti da analisi narratologiche che comprendono porzioni di testo lunghe, intricate e spesso protette da copyright. I diversi approcci quantitativi e qualitativi di costruzione dei prompt, con particolare riguardo ai prompt zero, one- e few- shot (si veda *ultra*), hanno mostrato applicazioni adattabili a diverse tipologie di file, testo e contesto forniti (cf. anche Dang et al. 2022; Kojima et al. 2022).

3. IL CORPUS E LA METODOLOGIA

Il nostro esperimento, nella sua visione più generale, mira a indagare quantitativamente la presenza del DIL e di altre tracce di interiorità nei testi letterari italiani pubblicati tra il 1830 e il 1930. A tale fine è stato costituito un corpus cercando di conseguire la massima rappresentatività possibile, stante le intrinseche difficoltà nella definizione di campione rappresentativo per la tradizione letteraria (Bode, 2018; Moretti, 2017). Il corpus RIND è costituito da 1.000 romanzi, equamente suddivisi tra 500 italiani e 500 traduzioni in italiano. Il corpus è stato bilanciato in modo da garantire una rappresentazione completa della produzione letteraria italiana degli anni presi in analisi, considerando fattori quali il genere dell'autore, la provenienza geografica e l'appartenenza al canone. Per questo esperimento specifico, abbiamo escluso le traduzioni per concentrarci esclusivamente sulla prosa italiana originale. Ogni testo è stato preparato in formato solo testo con codifica UTF-8 per garantire uniformità e compatibilità con gli strumenti di calcolo. Dal corpus iniziale, abbiamo selezionato un sottoinsieme bilanciato (per anno e per genere) di testi per un'analisi dettagliata del DIL: 40 testi e per ogni testo abbiamo selezionato porzioni di circa 5000 caratteri a testo, estratti randomicamente da un programma Python, per poter sottoporre i diversi chunk all'analisi di GPT e di Claude attraverso differenti tipologie di prompt con l'obiettivo di identificare il DIL nei testi forniti.
Come fenomeno linguistico, i DIL si estendono oltre i confini della frase e spesso mancano di marcatori linguistici espliciti che ne segnalino inizio e fine, ponendo una sfida per l'analisi quantitativa classica del fenomeno. Gli approcci NLP tradizionali non riescono infatti a identificare sistematicamente i DIL nei testi letterari (Brunner et al., 2019; Tu et al., 2019; Taivalkoski-Shilov, 2019). Per affrontare queste difficoltà, abbiamo utilizzato i LLM in modo da rilevare e classificare i tipi di discorso riportato - diretto, indiretto e indiretto libero - sulla base dei verbi del discorso, dei marcatori pragmatici e degli indizi contestuali. In particolare, abbiamo testato i modelli Open AI GPT-4o e Claude 3.5 Sonnet per mappare quantitativamente l'evoluzione del DIL nei vari periodi letterari. Inizialmente abbiamo definito l'oggetto dell'analisi, identificato i gruppi di utenti target e specificato lo scopo della tassonomia. Abbiamo costruito diversi tipi di prompt basandoci sui criteri esposti qui nel paragrafo 4 e seguendo parzialmente la classificazione tassonomica e la metodologia di Böhmker et al. (2023). I test sono stati eseguiti sia su campioni di testo tokenizzati per frase e archiviati in formato tabellare (CSV), sia su porzioni di testo dei file in formato "solo testo" dei romanzi.
La fase iniziale del nostro esperimento si è concentrata sulla valutazione dell'accuratezza e delle capacità degli LLM nell'identificare il DIL all'interno del nostro sotto-corpus di riferimento. Per avere un raffronto verificabile, abbiamo taggato manualmente, evidenziando la presenza dei diversi DR, i 40 testi citati. Questo set di dati annotati ha fornito un'importante base di confronto per la valutazione delle prestazioni dei modelli e posto le basi per la seconda fase dell'esperimento, in cui si prevede di mettere a punto almeno due modelli più piccoli e di libero accesso e di confrontare i risultati con quelli dei grandi modelli testati finora. Data la variabilità della struttura e dello stile narrativo nel nostro corpus, abbiamo tentato diverse strategie di tokenizzazione per determinare l'unità più efficace per rilevare i DIL: a livello di frase,

di paragrafo e di porzioni testo significative (intorno ai 5000 caratteri), utilizzando un programma di approssimazione Python per definire porzioni di testo comparabili. Dopo un test comparativo, la tokenizzazione per *sentence* (sebbene la definizione stessa di "sentence" sia complessa) è emersa come l'approccio più efficiente per catturare i marcatori linguistici e sintattici dei DIL che operano a livello di concatenazioni di frasi. Utilizzando il set di dati taggati manualmente come "verità di base", abbiamo poi testato e confrontato ChatGPT (in particolare lo strumento di analisi dei dati di OpenAI) e Claude 3.5 Sonnet. Il processo di test prevedeva l'inserimento di testo in questi modelli e la loro interrogazione per classificare le istanze di discorso diretto, indiretto e DIL in base a criteri linguistici predefiniti. Successivamente, i risultati dei modelli sono stati sistematicamente confrontati con i dati taggati manualmente per valutarne l'accuratezza e l'affidabilità nel riconoscimento dei DIL. Abbiamo costruito diversi tipi di prompt e utilizzando un approccio *empirical-to-conceptual* basandoci sui seguenti livelli di analisi e seguendo la classificazione tassonomica di Böhmker et al. 2023 per poter costruire una nuova tassonomia che intendiamo proporre per l'analisi del romanzo italiano tra Ottocento e Novecento:

a) Analisi comparativa dei prompt zero-shot (tab.1) e few-shot (tab.2): i.e. Richiesta di analisi di porzione di testo intorno ai 5000 caratteri senza e con esempi tratti dalla letteratura critica di base sul tema.
b) Analisi di precisione testuale e della progressione human-in-the-loop (tab. 3.) con richiesta di spiegazione preliminare del fenomeno linguistico DIL.
c) Analisi dei diversi esempi da fornire da GPT e Claude singolarmente, con il prompt di riferimento few-shot e human-in-the-loop.

Tab. 1. Zero-Shot Prompts

Tipo	Italiano	English
Base	Leggi e analizza il seguente testo e identifica eventuali passaggi in discorso indiretto libero [allega .csv/.txt]	Read and analyze the following text and identify any instances of free indirect speech: [attach .csv/.txt]
Dettagliato	Leggi il testo nel file CSV. Per ogni frase, valuta se contiene un discorso indiretto libero. Creare un nuovo file CSV che includa una colonna aggiuntiva etichettata "Discorso indiretto libero" con una voce "sì" o "no" per ogni frase, indicando la presenza o l'assenza di discorso indiretto libero. rinominare il file con il nome del file [allega .csv]	Read the text in the CSV file. For each sentence, evaluate if it contains free indirect speech. Create a new CSV file that includes an additional column labeled "Free Indirect Speech" with a "yes" or "no" entry for each sentence, indicating the presence or absence of free indirect speech. rename the file with the file name:[attach .csv]

Tab. 2. Few-Shot Prompts

Tipo	Italiano	English
Multi-esempio	Ecco tre esempi di DIL da diversi autori: [esempi]. Usando questi come riferimento, leggi e analizza il seguente testo: [attach -.csv/.txt]	Here are three examples of FIS from different authors: [examples]. Using these as reference, read and analyze the following text: [attach .csv/.txt]
Comparativo GPT	Leggi e riscrivi tutto il seguente testo indicando con "si" e "no" se è presente il discorso indiretto libero e sistemalo in una tabella [allega .txt]; Riconduci l'analisi di cui sopra alla luce di questi esempi: [esempi] [allega .txt]	Read and rewrite the whole of the following text, indicating with "yes" and "no" whether free indirect speech is present and arrange it in a table [attach .txt]; Reconduct the above analysis in the light of these examples: [examples].
Comparativo Claude	Individua il Discorso indiretto libero nel seguente testo, questi sono alcuni esempi: [esempi] [allega csv].	Identify Free Indirect Speech in the following text, these are some examples: [examples], attach csv.

Tab. 3. Prompt con Progressione Human-in-the-Loop

Fase	Italiano	English
Iniziale	Ciao! Spiegami cos'è il discorso indiretto libero	Hi, tell me what's Free Indirect Speech
Verifica	Individualo in questo csv e riscrivi un altro csv con il testo con una colonna si no in corrispondenza delle frasi con DIL	Analyse this file and produce a csv file with an additional column "Free Indirect Speech" with "yes/no" label
Raffinamento	Rivedi la tua analisi considerando anche [esempi]:	Review your analysis considering also [examples]:
Conclusione	Fornisci una nuova classificazione finale dei passaggi, indicando il grado di certezza per ciascuna identificazione:	Provide a final classification of passages, indicating the degree of certainty for each identification:

Questo approccio sistematico, che combina tokenizzazione a livello di frase, taggatura manuale e test comparativi dei prompt, ci ha permesso di costruire un framework metodologico solido, ma flessibile, per l'identificazione del DIL. La nostra esperienza dimostra come un'attenta progettazione delle fasi di analisi, dalla scelta dell'unità testuale fino alla costruzione dei prompt, sia fondamentale per ottenere risultati affidabili nello studio computazionale di fenomeni narrativi complessi.

4. RISULTATI DELL'ESPERIMENTO E DISCUSSIONE

L'analisi quantitativa della progressione metodologica, dalla baseline zero-shot fino all'implementazione few-shot, ha mostrato un incremento significativo nelle performance predittive sia di GPT sia di Claude, ma soprattutto di quest'ultimo. I prompt zero-shot hanno mostrato una tendenza alla sovra-classificazione, con tassi di non corrispondenza rispetto al tagging manuale del 33,7% per Claude S.3.5 e del 35,0% per GPT 4.o. I prompt few-shot hanno mostrato un miglioramento significativo nelle metriche di valutazione, con una riduzione della non corrispondenza: Claude è passato dal 33,7% (zero-shot) al 24,6% (few-shot), mentre GPT è passato dal 35,0% (zero-shot) al 34,3% (few-shot). In termini di corrispondenza con il tagging manuale, Claude S.3.5 ha raggiunto il 75,4% nella configurazione few-shot, mostrando il miglioramento più significativo tra i modelli testati. I grafici (Figura 1) mostrano chiaramente i cambiamenti dall'analisi da zero-shot a few-shot rispetto alla nostra baseline (il nostro tagging). Su un totale di 975 frasi analizzate con Claude in modalità zero-shot, 640 (66,3%) corrispondevano al tagging manuale, mentre con l'approccio few-shot, su 686 frasi analizzate, 517 (75,4%) mostravano corrispondenza. Per GPT 4.o, su 1108 frasi analizzate in modalità zero-shot, 720 (65,0%) corrispondevano al tagging manuale, mentre con l'approccio few-shot, su 1263 frasi, 830 (65,7%) mostravano corrispondenza.

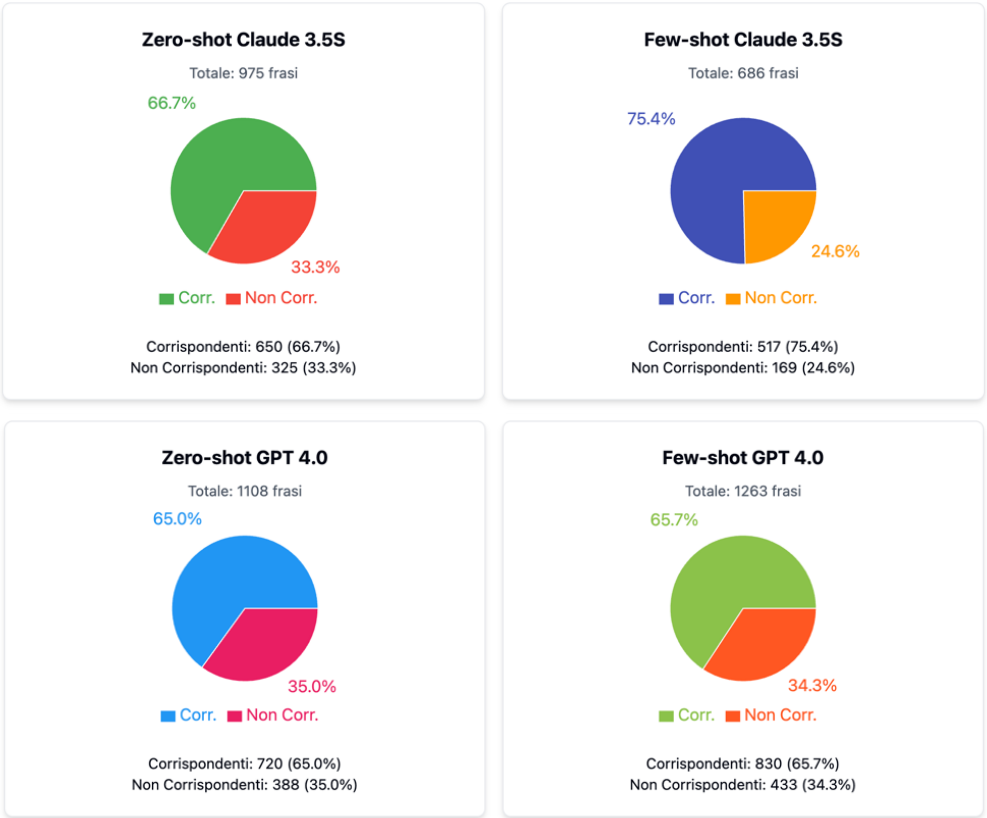


Figura 1 Risultati progressione Zero>Few-shot con GPT e Claude

In una seconda analisi dei dati (analisi in questo caso non automatizzata) abbiamo osservato che l'approccio few-shot, arricchito da una formalizzazione teorica preliminare del fenomeno linguistico target tramite esempi di varia natura ha permesso agli LLM di sviluppare rappresentazioni più granulari del DIL, ottimizzando la discriminazione tra le diverse tipologie di discorso riportato.

L'analisi contrastiva dei verbi di comando nelle due lingue target ha rivelato una correlazione positiva tra la coerenza linguistica prompt-testo e l'accuratezza predittiva (i.e. è preferibile utilizzare prompt in italiano su testi in italiano, sebbene la variazione non sia così sostanziale nei risultati, ma test aggiuntivi si richiedono per fornire questa affermazione di solidità scientifica). I risultati non sono qui riportati, ma sono stati eseguiti anche test con prompt strutturalmente più complessi, dimostrando che la variazione verbale abbia una ovvia ricaduta sui risultati forniti da GPT e Claude. In particolare si riportano qui due problematiche interessanti: nel caso di Claude l'utilizzo dei verbi "rewrite" e "riscrivi" pone problemi dal punto di vista di copyright (l'output di alcuni testi viene bloccato e si è rimandati alla copyright policy page) aggirabili parzialmente fornendo testi più brevi (intorno ai 1800 caratteri); mentre nel caso di GPT l'utilizzo del verbo "analyze" in coppia con i file CSV porta GPT a creare un programma di tokenizzazione Python per l'analisi del DIL secondo alcune regole arbitrarie e sistematiche che tuttavia eludono l'elemento contestuale; si è pertanto deciso, nel caso di GPT di fornire chunk di testo .txt per aggirare il problema e di far tokenizzare il testo per sentence direttamente a GPT (per questo si troverà nella tabella finale una variazione significativa nel numero di frasi analizzate tra Claude e GPT).
Il processo human-in-the-loop è stato fondamentale per l'ottimizzazione dei prompt. Attraverso più di 100 cicli di feedback, sono stati identificati e corretti pattern di errore ricorrenti, conducendo alla definizione di prompt sempre più robusti e *context-aware*. Tuttavia, l'utilizzo di questa metodologia ha anche dimostrato che il processo stesso sia utile per la formazione e creazione di prompt concretamente più performanti, sebbene non lo sia a livello di analisi vera e propria. Dalle nostre analisi, il coinvolgimento dialogico porta, infatti, sia GPT sia Claude a un alto livello di sovrainterpretazione del testo.
La finestra ottimale per l'analisi testuale si è attestata nell'intervallo 1.800-5.000 caratteri (per motivi di gestibilità del documento da parte delle IA e per motivi di copyright, come detto), con una degradazione non lineare dell'accuratezza oltre tale threshold. Nel 50% dei casi di classificazione scorretta, le interpretazioni alternative generate dagli LLM presentavano una validità ermeneutica significativa secondo la nostra annotazione manuale: è stato rilevato inoltre che nel 32% di casi l'annotazione positiva di DIL operata dagli LLM fosse giustificabile contestualmente e in parte superasse le sviste del tagging manuale, fornendo importanti spunti di riflessione sui labili confini del fenomeno preso in analisi e sulla necessità di un più ampio spettro di annotatori manuali. Le sfide più complesse riguardano l'identificazione dei marcatori DIL *low-salience* e la gestione dell'ambiguità interpretativa, aspetti che richiedono non solo un ulteriore raffinamento delle strategie di prompting, ma anche l'implementazione di tecniche di tokenizzazione semanticamente informate.

5. CONCLUSIONI

I risultati dell'esperimento mostrano la capacità degli LLM di applicare strategie di classificazione su fenomeni intrinsecamente complessi e sfumati, soggetti a variabilità interpretativa anche da parte di esperti umani, aprendo spazi di sperimentazione senza precedenti per lo studio su vasta scala di testi complessi come il romanzo, come mostrano anche le ricerche di (Piper & Bagga, 2024) e di (Underwood, 2023) - ma anche il testo filosofico, si veda ad esempio (Mosca, 2024). Allo stesso tempo esso evidenzia come queste capacità analitiche siano soggette a una forte variabilità e dipendenza dal contesto (tecniche di prompt, tipo e dimensione dei frammenti di testo, basi teorico-critiche labili etc.); a questo si aggiunge alla ben nota mancanza di perspicuità del comportamento in fase di inferenza delle reti neurali di grandi dimensioni pone sostanziali problemi di validazione, affidabilità e accettabilità epistemica. In questa ottica, se ovviamente è necessario innalzare la soglia di attenzione nella valutazione critica dei risultati da parte dei ricercatori, è anche opportuno esplorare l'efficacia di modelli linguistici di medie o piccole dimensioni, soggetti ad opportune di tecniche di fine tuning verticali su dati di training di alta qualità. Nell'ambito del progetto RIND abbiamo iniziato a sondare queste possibilità, nei limiti concessi da una scala progettuale limitata come quella del programma PRIN. Considerata tuttavia la complessità e i costi per lo sviluppo e l'utilizzo di sistemi di AI generativa, consideriamo vieppiù opportuno promuovere sperimentazioni su frammenti piccoli, ma significativi dell'analisi, come nel caso delle analisi qui proposte sul prompt engineering.

BIBLIOGRAFIA

Bach, S. H., Sanh, V., Yong, Z. X., Webson, A., Raffel, C., Nayak, N. V., Sharma, A., Kim, T., Bari, M. S., Févry, T., Alyafeai, Z., Dey, M., Santilli, A., Sun, Z., Ben-David, S., Xu, C., Chhablani, G., Wang, H., Fries, J. A., ... Rush, A. M. (2022). Promptsource: An integrated development environment and repository for natural language prompts. arXiv, abs/2202.01279.

Bode, K. (2018). *A world of fiction: Digital collections and the future of literary history*. University of Michigan Press.

Böhmker, M., Greve, M., Kegel, F., Kolbe, L. M., & Beyer, P. E. (2024). Can (A)I have a word with you? A taxonomy on the design dimensions of AI prompts. Proceedings of the Annual Hawaii International Conference on System Sciences.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., & Liang, P. (2021). On the opportunities and risks of foundation models.

Brunner, A., Tu, N. D. T., Weimer, L., & Jannidis, F. (2019). Deep Learning for Free Indirect Representation. In Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019): Short Papers (pp. 241-245). Erlangen, Germany. https://ids-pub.bsz-bw.de/frontdoor/index/index/docId/9315

Byszuk, J., Woźniak, M., Kestemont, M., Leśniak, A., Łukasik, W., Šeļa, A., & Eder, M. (2020). Detecting Direct Speech in Multilingual Collection of 19th-century Novels. In R. Sprugnoli & M. Passarotti (A c. Di), *Proceedings of LT4HALA 2020—1st Workshop on Language Technologies for Historical and Ancient Languages* (pp. 100–104). European Language Resources Association (ELRA). https://aclanthology.org/2020.lt4hala-1.15/

Calaresu, E. (2000). Il discorso riportato. Una prospettiva testuale. Il Fiorino.

Cangiano, M. (2018). La nascita del modernismo italiano: Filosofie della crisi, storia e letteratura. 1903-1922. Quodlibet Studio.

Castellana, R. (2010). Realismo modernista. Un'idea del romanzo italiano (1915-1925). Italianistica, 39(1), 23-45.

Ciotti, F. (2022a). Una nuova svolta negli studi letterari: La convergenza tra computazione, cognizione ed evoluzione. In La narrazione come incontro (pp. 19-36). Firenze University Press. https://doi.org/10.36253/979-12-215-0045-5.04

Ciotti, F. (2022b). Computational approaches to literary periodization: An experiment in Italian narrative of 19th and 20th century. In Digital Humanities 2022. Conference Abstracts (pp. 181-183). DH2022 Local Organizing Committee.

Ciotti, F. (2023). Minerva e il pappagallo. IA generativa e modelli linguistici nel laboratorio dell'umanista digitale. *TESTO & SENSO*, *26*, 289–315. https://doi.org/10.58015/2036-2293/671

Dang, H., Mecke, L., Lehmann, F., Goller, S., & Buschek, D. (2022). How to prompt? Opportunities and challenges of zero- and few-shot learning for human-AI interaction in creative applications of generative models. http://arxiv.org/abs/2209.01390

Donnarumma, R. (2012). Tracciato del modernismo italiano. In R. Luperini & M. Tortora (Eds.), Sul modernismo italiano (pp. 13-38). Liguori Editore.

Jannidis, F., & Lauer, G. (2014). Burrows's Delta and Its Use in German Literary History. In M. Erlin & L. Tatlock (Eds.), Distant readings. Topologies of German culture in the long nineteenth century (pp. 29-54). Camden House.

Jockers, M. L. (2013). Macroanalysis: Digital methods and literary history. University of Illinois Press.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. http://arxiv.org/abs/2205.11916

Kundisch, D., Muntermann, J., Oberländer, A. M., Rau, D., Röglinger, M., Schoormann, T., & Szopinski, D. (2022). An update for taxonomy designers. *Business & Information Systems Engineering, 64*(4), 421-439.

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2021). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. http://arxiv.org/abs/2107.13586

Luperini, R. (2018). Il modernismo italiano esiste. In M. Tortora (Ed.), Modernismo e modernità. Liguori Editore.

Mollick, E. (2023, settembre 16). Working with AI: Two paths to prompting. *One Useful Thing*. https://www.oneusefulthing.org/p/working-with-ai-two-paths-to-prompting

Moretti, F. (2005). Graphs, maps, trees: Abstract models for a literary history. Verso.

Moretti, F. (2013). Distant reading. Verso.

Moretti, F. (A c. Di). (2017). *Canon/Archive: Studies in Quantitative Formalism*. N+1.

Mosca, F. (2024). Wittgensteinian Network and Wittgensteinian Oracle. Two humanistic-digital tools in and beyond lexicological enquiry. *Umanistica Digitale*, *8*(18), 145–173. https://doi.org/10.6092/issn.2532-8816/20549

Nickerson, R. C., Varshney, U., & Muntermann, J. (2013). A method for taxonomy development and its application in information systems. *European Journal of Information Systems, 22*(3), 336-359.

Piper, A. (2018). Enumerations: Data and literary study. The University of Chicago Press.

Piper, A., & Bagga, S. (2024). Using Large Language Models for Understanding Narrative Discourse. In Y. K. Lal, E. Clark, M. Iyyer, S. Chaturvedi, A. Brei, F. Brahman, & K. R. Chandu (A c. Di), *Proceedings of the The 6th Workshop on Narrative Understanding* (pp. 37–46). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.wnu-1.4

Sullet-Nylander, F., Roitman, M., Muñoz, J.-M. L., Marnette, S., & Rosier, L. (Eds.). (2014). Discours rapporté, genre(s) et médias. Stockholm University Press.

Taivalkoski-Shilov, K. (2019). Free Indirect Discourse: An Insurmountable Challenge for Literary MT Systems? Translation Spaces, 8(1), 9-29. https://doi.org/10.1075/ts.00002.tai

Tortora, M. (2018). Il modernismo italiano. Carocci.

Tu, N. D. T., Krug, M., & Brunner, A. (2019). Automatic Recognition of Direct Speech Without Quotation Marks: A Rule-Based Approach. In Proceedings of the Digital Humanities Conference (DHD). https://www.bibsonomy.org/bibtex/13c802571e843affc5bba0f30a334938e/krug

Underwood, T. (2019). Distant horizons: Digital evidence and literary change. The University of Chicago Press.

Underwood, T. (2023, marzo 19). Using GPT-4 to measure the passage of time in fiction. *The Stone and the Shell*. https://tedunderwood.com/2023/03/19/using-GPT-4-to-measure-the-passage-of-time-in-fiction/

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2023). *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models*. https://doi.org/10.48550/ARXIV.2201.11903

Wellek, R. (1956). Periods and movements in literary history. In R. Wellek & A. Warren, Theory of literature (pp. 263-283). Harcourt, Brace & World.

Wu, T., Terry, M., & Cai, C. J. (2022). AI chains: Transparent and controllable human-AI interaction by chaining large language model prompts. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1-22.

Zamfirescu-Pereira, J. D., Wong, R. Y., Hartmann, B., & Yang, Q. (2023). Why Johnny can't prompt: How non-AI experts try (and fail) to design LLM prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1-21.

Historical GIS e metodologie digitali per una storia della copertura boschiva

Vincenzo Colaprice[1]
[1] Università degli Studi di Torino, Italia – v.colaprice1@gmail.com

ABSTRACT (ITALIANO)
La recente approvazione del Nature Restoration Law ha risollevato il dibattito sulle strategie da adottare per il ripristino degli ecosistemi. Tra gli strumenti di ricerca applicata prospettati si aprono spazi anche per gli Historical GIS, intesi come geodataset di fonti storiche utili alla ricostruzione delle trasformazioni dei paesaggi anche per orientare l'intervento dei decisori pubblici. Ai fini di valutare questa prospettiva, questo contributo si concentra sul caso studio della Liguria, regione italiana con il più alto indice di boscosità, dove il progressivo spopolamento e l'abbandono delle pratiche locali hanno determinato un'espansione boschiva significativa, spesso non priva di criticità. Ricorrendo all'analisi storico-regressiva e a metodologie comparative, è stato sviluppato un sistema informativo geografico in grado di rappresentare e analizzare l'evoluzione della copertura boschiva nell'area appenninica compresa tra la Val Trebbia e la Val d'Aveto tra XIX e XXI secolo. Parte del progetto PRIN 2022 "Bridging geography and history of woodlands", il lavoro offre spunti interdisciplinari per riflettere sulle sfide metodologiche rappresentate dalla gestione dei dati geo-storici e per rispondere alle sollecitazioni provenienti dal dibattito pubblico, proponendo un modello esportabile in contesti analoghi.
Parole chiave: Historical GIS; Liguria; boschi; cartografia; Nature Restoration Law

ABSTRACT (ENGLISH)
Historical GIS and digital methodologies for a history of forest cover. The recent approval of the Nature Restoration Law has reignited debates on strategies for ecosystem restoration. Among the proposed tools for applied research, Historical GIS offers potential as a geodataset of historical sources that can reconstruct landscape transformations and guide public policy interventions. To evaluate this perspective, this study focuses on Liguria, the Italian region with the highest forest cover index, where progressive depopulation and the abandonment of traditional practices have led to significant forest expansion, often accompanied by critical issues. Using a historical-regressive analysis and comparative methodologies, a Geographic Information System has been developed to represent and analyze the evolution of forest cover in the Apennine area between Trebbia and Aveto Valleys from the 19th to the 21st century. As part of the PRIN 2022 project "Bridging Geography and History of Woodlands", this work offers interdisciplinary insights into the methodological challenges of managing geo-historical data. It also responds to public debate by proposing a scalable model applicable to similar contexts.
Keywords: Historical GIS; Liguria; woodlands; cartography; Nature Restoration Law

1. INTRODUZIONE
Negli ultimi anni, il dibattito pubblico sorto attorno alla crisi climatica ha reso centrale la riflessione sulla gestione delle risorse ambientali, spingendo le istituzioni ad individuare provvedimenti destinati a sostenere nuove forme di gestione sostenibile degli ecosistemi a rischio. L'ultimo, in ordine di tempo, è stato adottato nel giugno 2024 dal Parlamento e dal Consiglio dell'Unione Europea, ovvero il Nature Restoration Law (NRL)[1]. Il provvedimento si presenta come un regolamento che vincola gli Stati Membri «a riportare la natura e gli ecosistemi a un buono stato di conservazione». Gli interventi previsti mirano a ripristinare entro il 2030 almeno il 20% degli ecosistemi degradati in aree terrestri e marine e la restante parte entro il 2050.
L'attuazione di questo regolamento avrà nei prossimi anni un'influenza diretta sulle politiche definite da ciascuno Stato e dalle istituzioni locali. Il carattere di questo "ripristino della natura" è da intendersi come «un miglioramento [degli habitat] da attuare secondo specifici criteri valutati come efficaci in termini di sostenibilità» (Gabellieri, 2024: 73). In merito alla definizione di questi criteri il NRL indica agli Stati Membri dell'Unione Europea la necessità di predisporre «piani nazionali di ripristino dettagliati sulla base delle migliori evidenze scientifiche disponibili» (art. 65). Non a caso, il regolamento prevede l'esecuzione di attività di mappatura utili a individuare le aree in cui intervenire, accompagnate da un'analisi diacronica

1 Testo integrale del Nature Restoration Law disponibile al sito: https://eur-lex.europa.eu/legal-content/IT/TXT/?uri=CELEX:32024R1991. Ultimo accesso: 21 gennaio 2025.

che possa orientare i piani nazionali «al fine di evitare conseguenze indesiderate» e favorire il controllo e la previsione di catastrofi naturali (art. 65 e 68).
Di fatto, seppur evitando di esplicitare un collegamento diretto con il raggiungimento degli obiettivi previsti dal regolamento, il NRL conferisce una certa rilevanza agli studi scientifici in relazione ad alcuni ambiti d'intervento, sollecitando l'adozione di alcune metodologie di ricerca e l'implementazione di tecnologie digitali, con particolare attenzione ai sistemi informativi geografici quali strumento di condivisione di dati con la programmazione ambientale. Inoltre, se l'apertura alla dimensione storica degli ecosistemi appare ancora ridotta (Gabellieri, 2024), questa è richiamata in alcuni passaggi del NRL in cui si riconosce che «registri documentati sulla distribuzione e sulla superficie storica […] dovrebbero servire da base per la determinazione delle superfici di riferimento favorevoli per tipi di habitat» (art. 65).
La necessità di mappare le aree di intervento e sviluppare un'analisi su scala diacronica apre la strada all'impiego degli Historical GIS, i quali consentono di rispondere a queste esigenze attraverso il dialogo interdisciplinare e puntando a influenzare i processi decisionali (Grava et al., 2020). L'utilizzo degli Historical GIS può rispondere alle sollecitazioni avanzate dal NRL (Gabellieri, 2024: 75) ed è da tempo attestato tra gli strumenti utili a ricostruire le dinamiche evolutive del paesaggio. A tal proposito, questa prospettiva è stata sostenuta nell'attività di ricerca relativa alle relazioni che intercorrono tra le comunità locali delle aree interne e il patrimonio culturale materiale e immateriale, evidenziando le potenzialità insite nelle attività di mappatura digitale per la definizione di strategie politiche *place-based* (Colaprice & Barbuti, 2023).
Le metodologie di analisi che contemplano l'utilizzo degli Historical GIS consentono di adottare una prospettiva integrata che supera la presunta dicotomia tra ecosistemi culturali e naturali e privilegia paradigmi di più recente formazione, come il concetto di *bio-cultural heritage*, inteso come «a link between people, history and biodiversity» (Rotheram, 2015: 3425). Questa proposta unitaria consente di andare oltre le interpretazioni del concetto di "ripristino della natura" volte a intendere un vago ritorno ad un passato incontaminato in cui non vi sia traccia della presenza umana. Nonostante questa narrazione abbia ispirato alcune politiche protezionistiche (Tigrino, 2015: 219) e si rifletta in parte nel NRL, l'esigenza di giungere a proposte efficaci di miglioramento degli ecosistemi spinge nella direzione di un approccio che metta al centro i processi storici che hanno determinato lo stato di attivazione e abbandono delle risorse ambientali. Tali fenomeni sono da ricondurre alle attività antropiche e all'impatto sociale ed ecologico esercitato dallo spopolamento e dalla deruralizzazione, processi che in Italia si sono manifestati in particolar modo nel contesto appenninico (Tigrino, 2015: 219), dando luogo ad una progressiva erosione delle risorse ambientali e alla contestuale perdita di saperi, pratiche e prodotti (Cevasco, 2007: 264; Rotheram et al., 2018: 269).
Dunque, appare evidente come la necessità di indagare questi processi storici assuma maggiore rilevanza. Lo sviluppo di un sistema informativo geografico capace di restituire su scala diacronica l'evoluzione paesaggistica e ambientale risponde direttamente alle sollecitazioni avanzate dal NRL e favorisce l'adozione di metodologie di ricerca interdisciplinari che mettono in relazione Digital Humanities, storia, geografia e archeologia. Il progetto PRIN 2022 "Bridging geography and history of woodlands" si inserisce in questo approccio interdisciplinare nello studio degli ecosistemi, mirando ad indagare i processi storici e socio-ecologici che hanno prodotto il patrimonio boschivo italiano e guardando alle ricadute pratiche in termini di gestione sostenibile[2]. Le attività di ricerca si basano sull'approfondimento dei due casi studio individuati, ovvero l'area appenninica compresa tra la Val Trebbia e la Val d'Aveto in Liguria e l'area dolomitica della Val di Fiemme in Trentino.
Il presente contributo intende illustrare le metodologie adottate nel corso del progetto per la costruzione di un Historical GIS volto a restituire l'evoluzione storica della copertura boschiva in Italia tra età moderna e contemporanea con riferimento al caso di studio ligure. Particolare attenzione è data al trattamento digitale della documentazione disponibile nel contesto locale, tenendo conto dei ben noti problemi relativi al reperimento di fonti omogenee per tutto il territorio nazionale in epoca preunitaria.

2. METODOLOGIA

La Liguria è la regione d'Italia con l'indice di boscosità più elevato. Quello che potrebbe apparire a prima vista come un dato positivo si scontra con diverse criticità che espongono il territorio al rischio di calamità. Nel corso dell'ultimo secolo, il progressivo abbandono di pratiche di gestione del bosco e lo spopolamento delle comunità locali ha determinato un avanzamento consistente della copertura boschiva. Questa ipotesi

2 Sito ufficiale del progetto PRIN: https://birnamscape.wixsite.com/birnam. Ultimo accesso: 21 gennaio 2025.

di ricerca ha orientato il lavoro sul caso studio ligure che insiste su un'area di circa 220 km^2, compresa nei comuni di Rovegno, Fontanigorda, Rezzoaglio e Santo Stefano d'Aveto, appartenenti alla provincia di Genova. Inoltre, all'interno di questi quattro territori, sono state individuate tre aree boschive che saranno oggetto di approfondimenti ulteriori nel corso del progetto, evidenziate dai confini tratteggiati nella figura 1.

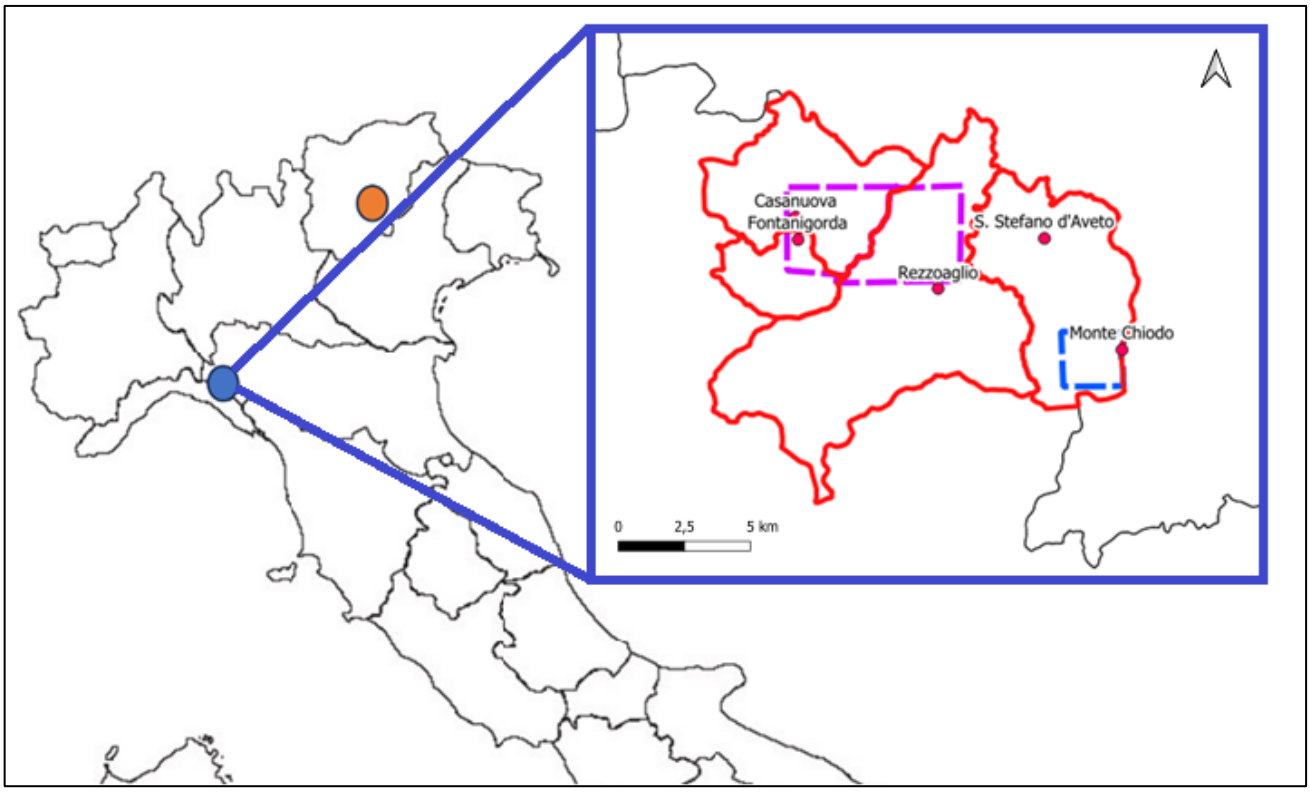


Figura 1. Localizzazione del caso di studio ligure.

La costruzione di un Historical GIS è stata essenziale per verificare la validità dell'ipotesi di partenza, già esplorata in casi di studio regionali a scala locale (Cevasco et al., 2005; Gabellieri & Primi, 2013; Priarone, 2023). A tal proposito, è stata adottata una metodologia di analisi basata sul metodo storico regressivo, volta ad individuare la persistenza di elementi di continuità e discontinuità nei processi storici (Moreno et al., 2005). Questo approccio rende essenziale l'incrocio di fonti di natura diversa, richiedendo la costituzione di una serie documentaria che contenga fonti testuali e cartografiche. Pertanto, è stato necessario costruire un repertorio cartografico che permettesse di rilevare la presenza della copertura boschiva negli ultimi due secoli. La definizione di questa serie cartografica si è avvalsa dell'utilizzo di alcune carte storiche della prima metà dell'Ottocento già studiate per l'area presa in esame (Cevasco, 2002) e di carte topografiche e digitali relative al Novecento e agli anni più recenti. Di seguito l'elenco delle fonti cartografiche utilizzate, eterogenee per caratteristiche e scala:

- Corine Land Cover (1990-2018), scala 1:100.000;
- Carta Tecnica Regionale (1986), scala 1:25.000;
- Carta d'Italia IGM (1936-1937), scala 1:25.000;
- Carta forestale del Regno d'Italia (1936), scala 1:100.000;
- Carta topografica degli Stati in Terraferma di S.M. il Re di Sardegna (1853), scala 1:50.000;
- Carta topografica degli Stati di Terra Ferma di S.S.R.M. Carlo Alberto Re di Sardegna (1831), scala 1:50.000;
- Carte preparatorie per la carta topografica degli Stati di Terra Ferma (1827-1829), scala 1:20.000.

Definito il repertorio cartografico, è stato possibile avviare in ambiente GIS un'analisi regressiva su scala diacronica adottando la metodologia del filtraggio cartografico, che prevede «il confronto di una sequenza di documenti cartografici rilevati e redatti in date diverse che rappresentano uno stesso sito o area» (Cevasco, 2007: 74). Come dato di partenza è stata utilizzata la rilevazione della copertura boschiva effettuata nel 2018 dall'agenzia europea Copernicus per il rilascio dell'ultima versione del dataset vettoriale Corine Land Cover (CLC), inclusivo dei dati risalenti fino al 1990. In questo modo, si è ottenuto un primo saggio dell'analisi regressiva, osservando minimi mutamenti nel trentennio preso in esame.

Successivamente, il GIS è stato arricchito con l'importazione della "Carta forestale del Regno d'Italia", realizzata nel 1936 dalla Milizia Forestale e oggetto di vettorializzazione in tempi recenti (Ferretti et al., 2018). La comparazione tra le due carte ha permesso di rilevare uno scarto consistente in termini di copertura boschiva tra gli anni Trenta e gli anni Novanta del Novecento, stimolando il ricorso alla cartografia storica per approfondire questi primi esiti della ricerca. Sono state incluse nel GIS carte tecniche regionali (1986), IGM (1936-1937) e carte risalenti al periodo preunitario prodotte dal Corpo di Stato Maggiore Sardo tra il 1816 e il 1868. Tale azione è stata preceduta dalla scansione a 300 DPI delle fonti cartografiche individuate e seguita dal processo di georeferenziazione dei file raster. Attraverso la decifrazione della simbologia propria di ciascuna carta è stato possibile individuare l'estensione della copertura boschiva, rappresentandola visivamente attraverso un processo di vettorializzazione poligonale. Le carte prodotte dal Corpo di Stato Maggiore Sardo hanno richiesto un lavoro di vettorializzazione differente, stante l'utilizzo di una simbologia letterale che indicava la localizzazione delle aree boschive. In questo caso, non è stato possibile visualizzare l'estensione della copertura boschiva, ma rappresentarne il posizionamento sulla carta e la tipologia attraverso una vettorializzazione di tipo puntuale.

3. RISULTATI OTTENUTI

Al netto dei limiti e della parzialità delle fonti cartografiche utilizzate, già oggetto di discussione in altre sedi scientifiche, il risultato ottenuto dall'attività di filtraggio cartografico ha permesso di confermare l'ipotesi di partenza verificando un progressivo avanzamento della copertura boschiva all'interno di un arco cronologico di circa 200 anni. Inoltre, appare evidente come la crescita del bosco sia aumentata in maniera consistente nella seconda metà del Novecento, quando l'abbandono di determinate pratiche di gestione del bosco e lo spopolamento hanno impattato maggiormente sull'area oggetto di studio. Questo fenomeno è evidenziato dalle percentuali riportate nella figura 2, le quali rappresentano il rapporto tra la superficie boschiva e quella complessiva dei quattro comuni, testimoniando il passaggio dal 54,28% della Carta forestale del 1936 all'86,06% rilevato nel 2018.

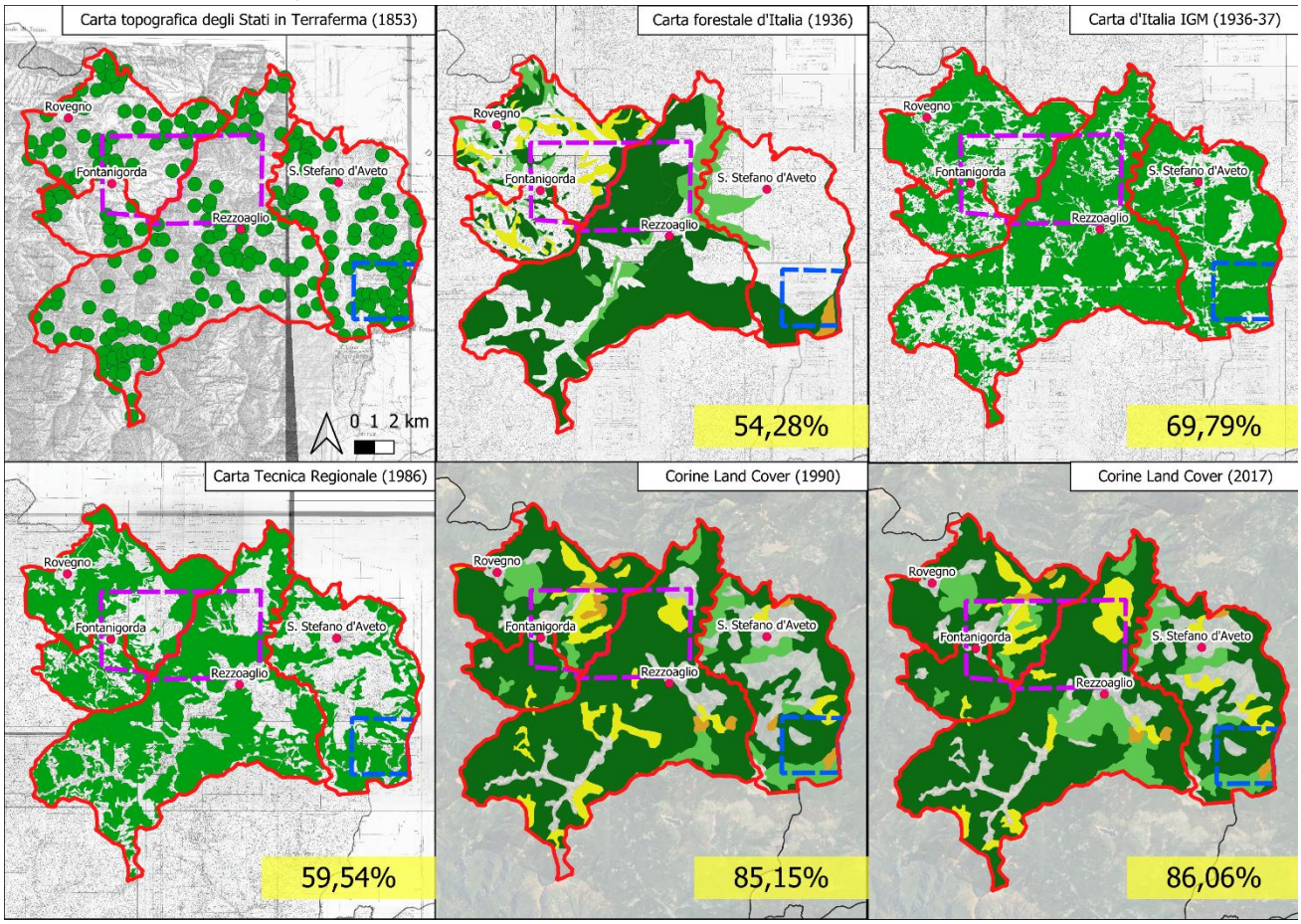


Figura 2. Esito del filtraggio cartografico: le carte riproducono l'estensione del bosco nel 1853, 1936, 1936-1937, 1986, 1990 e 2017. La percentuale evidenziata in ciascuna carta indica l'estensione della superficie boschiva.

A partire dal corpus di fonti cartografiche utilizzato, la creazione di un Historical GIS consente di svolgere un'attività di analisi a scala locale (per quanto permesso dalla fonte), rilevando dati quantitativi e qualitativi che si possono ricavare dallo studio delle carte. Una delle attività esperite ha riguardato il censimento dei toponimi indicati dalla cartografia, permettendo di individuare 1825 record che sono stati classificati per tipologia e grado di variazione su base diacronica, seguendo la metodologia proposta da Dai Prà, Gabellieri e Scanu (2022). La rilevazione di questi dati consente di fare emergere risvolti interdisciplinari, nonché la possibilità di agganciare altre fonti documentali ricche di informazioni essenziali per lo studio della storia del patrimonio boschivo. È il caso delle "Consegne dei Boschi e Selve", redatte per l'alta Val d'Aveto nel 1822, che consentono di cogliere aspetti qualitativi attraverso le descrizioni di boschi comunali e privati in esse contenuti (Bertolotto e Cevasco, 2000). Attraverso l'analisi delle descrizioni e l'incrocio con i toponimi a cui si riferiscono, è possibile dare una profondità qualitativa al lavoro di ricostruzione storica, individuando la presenza o la scomparsa di determinate specie vegetali. Si tratta di un lavoro essenziale, tuttora in stato di esecuzione, che supera alcuni limiti che caratterizzano le fonti cartografiche relative all'Appennino ligure, prive di quella cartografia catastale ottocentesca presente in altri territori. Un lavoro simile è stato realizzato per l'altro caso studio di cui si occupa il progetto, ovvero la Val di Fiemme, come dimostrato dalla recente pubblicazione di Gabellieri e Sarzotti (2024), dove la trasposizione in ambiente GIS della cartografia e delle fonti documentali prodotte dalla comunità locale nel XVIII secolo ha consentito di localizzare con precisione le aree boschive esistenti tra Settecento e Ottocento e rilevare tendenze e mutamenti relativi alla presenza di determinate specie vegetali oggi in gran parte scomparse, ampliando significativamente la diacronia permessa dalle sole fonti catastali.

4. RIFLESSIONI CONCLUSIVE

Lo sviluppo di un Historical GIS non rappresenta unicamente uno strumento a disposizione di storici e geografi. Attraverso il suo utilizzo si può provare a rispondere a sollecitazioni derivanti dal dibattito in corso sulla conservazione ambientale. Se in questo caso è stato possibile individuare le linee di tendenza in termini di evoluzione della copertura boschiva nell'arco di circa due secoli, la possibilità di rilevare dati qualitativi permette di determinare la storicità di alcune pratiche di gestione del bosco, così come la presenza o scomparsa di specie vegetali ad esse connesse. In questo modo si può provare ad intercettare quella richiesta di realizzazione di analisi diacroniche come previsto dal Nature Restoration Law.
Sul piano metodologico, si possono avanzare alcune considerazioni relative all'utilizzo della cartografia storica in ambiente GIS. Sebbene siano stati compiuti importanti passi avanti nell'automazione della georeferenziazione di raster e vettori attraverso il ricorso all'Intelligenza Artificiale, i software GIS non consentono ancora lo sviluppo di un'analisi integrata che preveda l'estrazione e la vettorializzazione automatizzata delle informazioni presenti in vari esempi di cartografia peritale sette-ottocentesca quale quella utilizzata. Il lavoro manuale risulta essere preminente, garantendo un elevato grado di controllo sul processo di digitalizzazione al costo di tempi di esecuzione lunghi. Le metodologie illustrate in questo contributo potrebbero giovarsi dell'utilizzo di plugin o di software GIS capaci di includere funzionalità legate all'IA, a partire dalla possibilità di riconoscere attraverso la tecnologia OCR i toponimi presenti all'interno delle carte. Un'innovazione che richiederebbe al contempo la capacità del software di distinguere tra i nomi legati agli elementi geografici e le informazioni testuali non pertinenti, nonché riconoscere stili e caratteri topografici nelle carte risalenti all'Ancien Régime.
Infine, l'utilizzo di fonti testuali all'interno di ambienti GIS pone il problema metodologico di tradurre grandi quantità di informazioni topografiche in dataset geospaziali utilizzabili. Fonti documentali come le "Consegne" a cui si è fatto cenno, pur offrendo un ricco apparato descrittivo, spesso contengono dati che non sono facilmente localizzabili con precisione. Queste sfide metodologiche indicano la necessità di sviluppare soluzioni innovative, non solo per ottimizzare i processi di digitalizzazione e analisi, ma anche per garantire una maggiore interoperabilità e una rappresentazione più accurata dei dati geo-storici.

RINGRAZIAMENTI

Il lavoro di ricerca esposto in queste pagine è finanziato dall'Unione Europea con fondi Next Generation EU, nell'ambito del bando PRIN 2022, progetto "Bridging geography and history of woodlands" (2022EKECST) – CUP E53D23010170006.

BIBLIOGRAFIA

Bertolotto, S., & Cevasco, R., (2000, gennaio). Fonti osservazionali e fonti testuali. Le «consegne dei boschi» e il sistema dell'«alnocoltura» nell'appennino ligure orientale (1822). Quaderni Storici: Nuova Serie, 103 (1), 87-108. https://www.jstor.org/stable/43779225.

Cevasco, R. (2002). La copertura vegetale dell'Alta Val Trebbia nelle ricognizioni topografiche del Corpo di Stato Maggiore Sardo (1816-1852): approccio storico all'ecologia dei siti. Archeologia Postmedievale, 6, 195-214. DOI: 10.1400/210860.

Cevasco, R., Marullo, E., & Stagno, A. M. (2005). L'analisi della cartografia storica per lo studio delle variazioni della copertura vegetale nel SIC RoccaGrande-M.te Pu (Liguria orientale). Atti della Nona Conferenza Nazionale ASITA (Federazione delle Associazioni Scientifiche per le Informazioni Territoriali e Ambientali), Catania 15-18 novembre 2005. 683-688. https://hdl.handle.net/11567/1026974.

Cevasco, R. 2007. Memoria verde. Nuovi spazi per la geografia. Reggio Emilia: Edizioni Diabasis. ISBN: 978-88-8103-487-1.

Colaprice, V., & Barbuti, N. (2023). Una nuova mappatura digitale per i borghi delle aree interne. Umanistica Digitale, 15, 25-46. DOI: 10.6092/issn.2532-8816/15612.

Dai Prà, E., Gabellieri, N., & Scanu, N. (2022). Dalla mappa al geodatabase: un modello di raccolta, digitalizzazione e analisi sincronica e diacronica in ambiente GIS del patrimonio toponomastico del territorio trentino da fonti cartografiche storiche (XIX-XXI secolo). Il capitale culturale, 25, 603-633. DOI: 10.13138/2039-2362/2815.

Gabellieri, N. (2024, dicembre). Nature Restoration Law e programmazione ambientale: quali prospettive per la ricerca geografico-storica? Rivista geografica italiana, CXXXI (4), 69-78. DOI: 10.3280/rgioa4-2024oa18973.

Gabellieri, N., & Primi, A. (2017). Uso del suolo e rischio idrogeologico: historical GIS e analisi geostorica della Val Bisagno (GE) dal XIX secolo ad oggi. Atti XXI Conferenza Nazionale Asita, Salerno 21-23 novembre 2017. 571-579. https://hdl.handle.net/11567/882863.

Gabellieri, N., & Sarzotti, E. (2024, October). Forest planning, rural practices, and woodland cover in an 18th-century Alpine Valley (Val di Fiemme, Italy): A geohistorical and GIS-based approach to the history of environmental resources. AIMS Geosciences, 10 (4), 767–791. DOI: 10.3934/geosci.2024038.

Grava, M., Berti, C., Gabellieri, N., & Gallia, A., (2020). Historical GIS. Strumenti digitali per la geografia storica in Italia. Trieste: Edizioni Università di Trieste. ISBN: 978-88-5511-158-4.

Moreno D. T., Guido, M. A., & Montanari, C. (2005). L'approccio storico-archeologico alla copertura vegetale: il contributo dell'archeologia ambientale e dell'ecologia storica. In G. Caneva (Ed.), La biologia vegetale per i beni culturali. Firenze: Nardini, 463-498. https://hdl.handle.net/11567/252572.

Priarone, E. (2023). L'analisi diacronica dell'uso del suolo nella cartografia: l'alta val di Vara (SP) tra 1936 e 2019. Bollettino dell'Associazione Italiana di Cartografia, 178, 114-127. https://www.openstarts.units.it/handle/10077/36748.

Rotheram, I. D. (2015). Bio-cultural heritage and biodiversity: emerging paradigms in conservation and planning. Biodivers Conserv, 24, 3405–3429. DOI: 10.1007/s10531-015-1006-5.

Rotheram, I. D., Çolak, A. H., & Kırca, S. (2018). European Woodland History and Management: some concluding thoughts. In A. H. Çolak, S. Kırca & I. D. Rotheram (Eds.), Ancient Woodlands and Trees: A Guide for Landscape Planners and Forest Managers. Vienna: IUFRO World Series Volume 37, 268-271. DOI: 10.53478/TUBA.2018.018.

Tigrino, V. (2015). Sovranità, comunità, possesso e lavoro nell'Appennino imperiale. Intorno ad una mappatura settecentesca della Val d'Aveto. In R. Leggero (Ed.), Montagne, comunità e lavoro tra XIV e XVIII secolo. Mendrisio Academy Press, 217-256. https://www.share.usi.ch/arc/montagne_comunita.pdf.

Experiments on the Use of LLMs for the Translation of the Babylonian Talmud

Mafalda Papini[1], Davide Albanesi[1], David Dattilo[2], Emiliano Giovannetti[1], Simone Marchi[1]
[1] Cnr-Istituto di Linguistica Computazionale "A. Zampolli", Italy - name.surname@ilc.cnr.it
[2] Progetto Traduzione Talmud Babilonese S.c.a r.l., Italy - david.dattilo@talmud.it

ABSTRACT (ENGLISH)
In this paper, we present an experiment on the use of Large Language Models (LLMs) in the translation of the Babylonian Talmud into Italian. The experiment focuses on translation and demonstrates how the combined use of LLMs and Translation Memory can improve the quality of suggested translations in a Computer-Assisted Translation context. The initial results highlight both the positive contribution of this hybrid technique and the limitations posed by the nature of complex texts like the Babylonian Talmud, whose translation can only be interpretative.
Keywords: Babylonian Talmud; Computer-Assisted Translation; Large Language Models; Artificial Intelligence

ABSTRACT (ITALIANO)
Esperimenti sull'uso dei LLM per la traduzione del Talmud babilonese. In questo articolo, presentiamo un esperimento sull'uso dei modelli di linguaggio di grandi dimensioni (Large Language Models) nella traduzione del Talmud babilonese in italiano. L'esperimento riguarda la traduzione e dimostra come l'uso combinato di LLM e Translation Memory possa migliorare la qualità delle traduzioni suggerite in un contesto di traduzione assistita dal computer. I primi risultati evidenziano sia il contributo positivo di questa tecnica ibrida, sia i limiti posti dalla natura di testi complessi come il Talmud babilonese, la cui traduzione non può essere che interpretativa.
Parole chiave: Talmud Babilonese; Traduzione Assistita dal Computer; Modelli di Linguaggio di Grandi Dimensioni; Intelligenza Artificiale

1. INTRODUCTION
The use of Large Language Models (LLMs) has become pervasive. The fields where such models are being applied are constantly expanding, driven by the curiosity to test their effectiveness on specific datasets and the undeniable evidence of their growing reliability. To evaluate the performance of the various models currently available, specific tasks have been defined to test their linguistic, cognitive, and multimodal capabilities[1]. Some of these tasks are central to research focusing on the linguistic competence of the models and their knowledge of the world, such as GPQA (Graduate-Level Google-Proof Q&A) (Rein et al., 2023) and MMLU-Pro (Massive Multitask Language Understanding) (Wang et al., 2024a). These particular tasks offer insights into the applications of the models' access to knowledge (both factual and linguistic) and the performance that can currently be achieved. It is therefore unsurprising that LLMs have been applied to the task discussed in this article, namely computer-assisted translation (CAT), due to their ability to understand and generate natural language.
While the use of LLMs in Machine Translation is not surprising (for a review, see Wang et al., 2024b), CAT supported by LLMs is another hot research frontier. The synergistic relationship between LLMs and Translation Memory (TM), the core of CAT systems, is central to the work presented in (Mu et al., 2023). The results show that the performance of an LLM trained for translation can be significantly enhanced by integrating TM-derived data into prompts. Although there is literature on the application of machine translation to religious texts (Liebeskind et al., 2024; Zaid & Bennoudi, 2023; Beal, 2022), to the best of our knowledge, no experiments have yet been conducted on the integration of TMs and LLMs in computer-assisted translation of religious texts.
This contribution illustrates an application of LLMs in the support of translating the Babylonian Talmud, carried out within the framework of the Babylonian Talmud Translation Project into Italian. As partly emerges from the experiment described later, undertaking the translation of the Talmud presents significant challenges due to the text's intrinsic complexity and the diverse range of topics it covers. The Talmud is primarily written in Biblical Hebrew and Babylonian Aramaic, using an extremely concise and

[1] For an overview of the main tasks, see the page: https://github.com/huggingface/leaderboards/blob/main/docs/source/en/open_llm_leaderboard/about.md (cons. 24/01/2025).

often cryptic style, rendering a literal translation incomprehensible. Additional words have been included in the translation to enhance readability. The text is structured into thematic "blocks" and further divided into "logical units" (e.g., objections, questions). Additional notes cover Halakhah, Nature, Linguistics, and brief biographies. Each volume also includes appendices containing glossaries of various nature.

2. METHODOLOGY

The translation of the Babylonian Talmud into Italian, carried out as part of the project of the same name since 2012, is supported at every stage by the Traduco software (Giovannetti et al., 2016), a CAT tool specifically developed for the project. Like any CAT tool, Traduco uses a TM as a database to suggest translations.

A TM is essentially a database containing sentence pairs that automatically records translated segments alongside their original source texts during translation. The primary aim of a TM is to reuse previous translations, thus enhancing the speed and consistency of the translation. TMs are particularly effective when working with texts characterized by a significant number of repeated or partially repeated segments (such as standardized expressions), a feature prominently found in highly formulaic and repetitive texts like the Babylonian Talmud.

After 12 years of translations, the TM contains 291,708 pairs of translated segments and a total of 1,633,544 segment pairs if historical versions are also considered. Thanks to this vast amount of available data, the system is able to suggest an exact translation for 35.94% of the segments, classified within the system as a "five-star" translation (net of any adjustments required by the context). The basic idea for this experiment arose precisely from the large amount of data available in the TM and the possibility of combining these data with the translation capabilities of language models.

Specifically, the following strategy was adopted: if a new segment to be translated appears in the TM with at least three suggestions having a similarity score greater than or equal to 50% (Bellandi et al., 2016), the support of an LLM is utilized.

The adopted methodology was inspired by two well-known techniques: Retrieval Augmented Generation, or RAG (Lewis et al., 2020), and few-shot prompting (Brown et al., 2020).

The idea is to integrate the prompt with a set of segments suggested by the TM for the new segment to be translated, providing the model with partial examples of translations[2]. This approach ensures that the model is not left entirely free to translate solely based on its internal knowledge but is guided by examples to preserve, as much as possible, the translational style already adopted, particularly in the choice of lexicon.

Normally, in a RAG system, semantic similarity measures are used to identify contexts. In this case, as well, the suggestions provided by the TM take this criterion into account, as the Translation Memory System already incorporates an embedding model that adjusts the similarity measure used to rank suggestions based on the semantic distance between words. For more details, see (Bellandi et al., 2016).

Below is the prompt template prepared for this experiment, where {*target_segment*} represents the segment to be translated:

> *You are an expert in translations from Hebrew-Aramaic to Italian.*
> *In the CONTEXT you will find examples of translations in the form of pairs*
> *where SOURCE is the Hebrew-Aramaic text and TARGET is the Italian text.*
> *CONTEXT:*
> *SOURCE="Hebrew segment",TARGET="Italian segment"*
> *...*
> *SOURCE="hebrew segment",TARGET="italian segment"*
> *Write only the translation of {target_segment} without writing anything else.*

To better illustrate the methodology adopted, we first present a "running example". The textual segment we are about to translate is "כֵּיוָן דַּאֲמַר לָהּ חֲצִי פְּרוּטָה פְּסָקָה", found in the first chapter of the tractate Qiddushin. For this segment, we have the reference Italian translation obtained after the translation process: "*Dal momento che le ha detto: 'Mezza perutà' ha interrotto l'azione*".

[2] In the context of the Talmud Translation Project, the experts adopted a customised text segmentation in which each string can range from a single word to short sentences. This choice responds to the peculiar and repetitive nature of the text and maximises the effectiveness of TM usage.

The system's TM is unable to provide any five-star suggestions for this segment. In other words, this sentence has not yet been translated in the exact form in which it appears. However, for this segment, the TM contains five suggestions with similarity scores ranging from 68% to 50%. The first of these suggestions provides the following translation: "*Dal momento che le dice: 'Domani' ha interrotto l'azione*". Before testing the application of suggestions on the LLM (we used Llama 3.1 70b), the language model was first asked to translate the segment directly without any assistance, resulting in the following translation: "*Poiché disse a lei mezza peruta, la interruppe*".
At this point, the prompt template described above was instantiated, and the model was queried again, this time providing the five TM suggestions as "context" input. The resulting translation was: "*Dal momento che le dice: 'Mezza perutà' ha interrotto l'azione*".
At a glance, the resulting translation is more accurate than the one obtained by querying the LLM directly. Even the word "פְּרוּטָה" was correctly translated with its accented form, "*perutà*", following the contextual examples provided and in accordance with the project's editorial standards.
Methodologically, the process began by extracting a random sample of 300 segment pairs (source-target) from the TM among those fulfilling the previously introduced eligibility requirements (i.e., the presence of at least 3 suggestions with a similarity score of at least 50%). For each segment pair, the "source" was translated both by directly invoking the LLM and by using the above prompt supplemented with translation suggestions. The "target" part was set aside as the reference segment. Subsequently, to evaluate the quality of the translations produced, three different measures used in Machine Translation were applied: SacreBLEU (Post, 2018), BERTscore (Zhang et al., 2020), and Meteor (Lavie & Agarwal, 2007). Given the particular nature of the text and its translation, a multidimensional evaluation approach was deemed preferable. This is because SacreBLEU focuses on lexical adherence, BERTscore analyzes deep semantic similarity, and METEOR accounts for synonyms and morphological flexibility. The results of the initial tests have been documented in two separate reports, one describing two examples extensively[3] and another more concisely compiling all the translations performed[4].
An analysis of the data obtained (see Table 1) revealed that for 237 of the segments considered (79% of the total), the combined use of LLM+TM resulted in a performance improvement across all three metrics compared to using the LLM alone.

	SacreBLEU	BERTscore	Meteor
LLM	11.0939	0.7283	0.2097
LLM+TM	35.7756	0.8273	0.5082
%increase	222.48%	13.60%	142.37%

Table 1. The arithmetic means of the values calculated for the three metrics and the corresponding percentage increases between LLM+TM and LLM.

To better evaluate the outcome of the experiment on the remaining 21% of the segments (comprising 63 pairs), a domain expert was consulted and asked to classify the translations obtained into one of the following four categories: the translation produced by LLM+TM is the best (16 cases), the translation produced by the LLM alone is the best (4 cases), both translations are acceptable (18 cases), or neither translation is acceptable (25 cases). Excluding these last two classes of results, in which the two techniques respond in agreement, the percentage of cases where LLM+TM produces better translations than LLM stands at 80%, in line with the previous result reported in Table 1.
Additionally, analyzing the remaining cases, it becomes evident that the specific nature of Talmudic text (often requiring an interpretative translation enriched with explanatory additions to improve readability) necessitates considering as reference segments translations that are significantly distant from the literal rendering of the original text. This inevitably results in very low scores on some or all of the metrics considered.
For example, the string "נְדָרִים,", which the model translates as "*voti,*" was translated in the Talmud (and considered as the reference segment in the experiment) with the phrase "*Per quanto riguarda le offerte di*

[3] https://github.com/klab-ilc-cnr/TalmudAI-AIUCD2025/blob/main/Translation.md (cons. 24/01/2025)
[4] https://github.com/klab-ilc-cnr/TalmudAI-AIUCD2025/blob/main/TranslationResults.md (cons. 24/01/2025)

voto,". However, the model with the translation examples provides a better translation ("*nedarìm*"), despite the fact that the measures classify this case as one of the negative ones. This is precisely because this transliteration often appears among the suggested translation examples (see footnote 2 for more details).

3. CONCLUSIONS

Although the results presented in this article are preliminary, it is already possible to draw some conclusions regarding the potential contribution that LLMs can offer in supporting the translation of particularly complex texts, such as the Babylonian Talmud.

The combined adoption of LLM and Translation Memory showed an improvement in the quality of the proposed translations, demonstrating that this hybrid approach can help in preserving the translation consistency and style.

Further analyses are necessary, first of all, to investigate in a more analytical manner how much the nature of the reference translation (especially when highly enriched with explanatory additions) and the number, order and nature of the example translations provided to the model from the TM influence the quality of the suggested translations. Furthermore, the information derived from the different styles in which the translations are written will be utilized, such as the use of bold style to represent literal translations.

Among future developments, additional models will be tested, particularly to verify if good performance can also be achieved with smaller-sized models. This would enable reduced hardware investment and easier integration into existing systems, such as Traduco.

Finally, we also plan to experiment with fine-tuning approaches using the extensive collection of parallel segments produced during the translation of the Babylonian Talmud.

ACKNOWLEDGEMENTS

Scientific publication produced thanks to the agreement between the National Research Council – Institute of Computational Linguistics and the PTTB S.c.a r.l. – Babylonian Talmud Translation Project.

REFERENCES

Beal, T. (2022). Interface of the Deep: Design Cues for Engaging New Media and Machine Translation with Religious Scriptures. In *The Routledge Handbook of Translation and Religion* (pp. 103–120). Routledge, Taylor & Francis Group. https://doi.org/10.4324/9781315443485-9

Bellandi, A., Benotto, G., Di Segni, G., & Giovannetti, E. (2016). Investigating the application and evaluation of distributional semantics in the translation of humanistic texts: A case study. *Proceedings of the 2nd Workshop on Natural Language Processing for Translation Memories (NLP4TM 2016)*, *28*, 6–11. http://www.lrec-conf.org/proceedings/lrec2016/workshops/LREC2016Workshop-NLP4TM_Proceedings.pdf#page=11

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. *Proceedings of the 34th International Conference on Neural Information Processing Systems*.

Giovannetti, E., Albanesi, D., Bellandi, A., & Benotto, G. (2016). Traduco: A collaborative web-based CAT environment for the interpretation and translation of texts. *Digital Scholarship in the Humanities*, *32*(suppl_1), i47–i62. https://doi.org/10.1093/llc/fqw054

Lavie, A., & Agarwal, A. (2007). METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments. In C. Callison-Burch, P. Koehn, C. S. Fordyce, & C. Monz (Eds.), *Proceedings of the Second Workshop on Statistical Machine Translation* (pp. 228–231). Association for Computational Linguistics. https://aclanthology.org/W07-0734/

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Proceedings of the 34th International Conference on Neural Information Processing Systems*.

Liebeskind, C., Liebeskind, S., & Bouhnik, D. (2024). Machine Translation for Historical Research: A Case Study of Aramaic-Ancient Hebrew Translations. *J. Comput. Cult. Herit.*, *17*(2), 20:1-20:23. https://doi.org/10.1145/3627168

Mu, Y., Reheman, A., Cao, Z., Fan, Y., Li, B., Li, Y., Xiao, T., Zhang, C., & Zhu, J. (2023). Augmenting Large Language Model Translators via Translation Memories. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Findings of the Association for Computational Linguistics: ACL 2023* (pp. 10287–10299). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-acl.653

Post, M. (2018). A Call for Clarity in Reporting BLEU Scores. In O. Bojar, R. Chatterjee, C. Federmann, M. Fishel, Y. Graham, B. Haddow, M. Huck, A. J. Yepes, P. Koehn, C. Monz, M. Negri, A. Névéol, M. Neves, M. Post, L. Specia, M. Turchi, & K. Verspoor (Eds.), *Proceedings of the Third Conference on Machine Translation: Research Papers* (pp. 186–191). Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-6319

Rein, D., Hou, B. L., Stickland, A. C., Petty, J., Pang, R. Y., Dirani, J., Michael, J., & Bowman, S. R. (2023). *GPQA: A Graduate-Level Google-Proof Q&A Benchmark* (No. arXiv:2311.12022). arXiv. https://doi.org/10.48550/arXiv.2311.12022

Wang, Y., Ma, X., Zhang, G., Ni, Y., Chandra, A., Guo, S., Ren, W., Arulraj, A., He, X., Jiang, Z., Li, T., Ku, M., Wang, K., Zhuang, A., Fan, R., Yue, X., & Chen, W. (2024a). *MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark* (No. arXiv:2406.01574). arXiv. https://doi.org/10.48550/arXiv.2406.01574

Wang, Y., Zhang, J., Shi, T., Deng, D., Tian, Y., & Matsumoto, T. (2024b). Recent Advances in Interactive Machine Translation With Large Language Models. *IEEE Access*, *12*, 179353–179382. IEEE Access. https://doi.org/10.1109/ACCESS.2024.3487352

Zaid, A., & Bennoudi, H. (2023). AI vs. Human Translators: Navigating the Complex World of Religious Texts and Cultural Sensitivity. *International Journal of Linguistics, Literature and Translation*, *6*(11), 173–182. https://doi.org/10.32996/ijllt.2023.6.11.21

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2019, September 25). *BERTScore: Evaluating Text Generation with BERT*. International Conference on Learning Representations. https://openreview.net/forum?id=SkeHuCVFDr

Metodi di allineamento testuale bilingue per un'edizione genetica digitale dei *Mémoires* di Carlo Goldoni

Matteo Zibardi[1]
[1] Sorbonne Université, CERES, Francia – matteo.zibardi@sorbonne-universite.fr

ABSTRACT (ITALIANO)

Il presente contributo propone i primi risultati ottenuti nell'ambito della realizzazione di un'edizione genetica digitale dei *Mémoires* di Carlo Goldoni, prodotta grazie al ricorso agli strumenti di Trattamento Automatico della Lingua (TAL). Pubblicata a Parigi nel 1787, l'autobiografia è il frutto del rifacimento del testo di diciassette prefazioni, scritte in lingua toscana, poste in testa ai tomi dell'edizione Pasquali delle opere teatrali dell'autore editi tra il 1761 e il 1778. La genesi bilingue dell'opera è stata però ignorata dalla critica goldoniana. Il presente contributo intende colmare tale aporia proponendo delle modalità operative per la realizzazione di un'edizione genetica digitale dei *Mémoires* di Carlo Goldoni. Dal momento che il testo francese è frutto di un profondo rifacimento di quello italiano, un semplice calcolo di similarità coseno tra le frasi delle due versioni non è stato sufficiente. Dunque, sono state estratte ed allineate le frasi contenenti entità nominate condivise dai due stadi testuali. L'allineamento testuale bilingue a partire dalle entità nominate si è rivelato di gran lunga più efficace e ha permesso di individuare molte porzioni comuni tra i due testi, nonostante la diversità linguistica. È stata così creata la prima versione di un'edizione sinottica che sarà presto realizzata nella sua interezza.

Parole chiave: edizione scientifica digitale; Named Entity Recognition; allineamento testuale bilingue; critica genetica; Carlo Goldoni.

ABSTRACT (ENGLISH)

Methods of bilingual text alignment for a genetic digital edition of Goldoni's Mémoires. This contribution proposes preliminary results obtained while realizing a digital genetic edition of Carlo Goldoni's *Mémoires*. This edition has been produced through the use of Natural Language Processing (NLP) tools. Goldoni's work represents a significant case study in the field of genetic criticism: published in Paris in 1787, it is the result of the rewriting of seventeen prefaces, written in Tuscan, placed at the beginning of the Pasquali edition of the author's plays published between 1761 and 1778. However, the bilingual genesis of this work has been neglected in Goldoni's critic bibliography. This contribution aims to describe the methodology employed to realize a digital genetic edition of the *Mémoires*. Assuming that the French texts has not only been traduced from the Italian one, but also modified in some of its parts, the similarity score between the sentences of the two version was not a reliable result. To find a better solution, the sentences in both source and target text containing equivalent named entities have been aligned. The bilingual text alignment based on the recognition of the same named entities has proved itself to be more efficient and permitted to detect a lot of common blocks between the two versions, despite linguistic and textual differences. So, the first draft of a synoptic genetic edition of the *Mémoires* has thus been created.

Keywords: digital scholarly editing; Named Entity Recognition; bilingual textual alignment; genetic criticism; Carlo Goldoni

1. INTRODUZIONE

I *Mémoires de M. Goldoni pour servir à l'histoire de sa vie et à celle de son théâtre* sono un caso di studio altamente significativo per la ricostruzione della vita teatrale europea del XVIII secolo e per l'analisi delle pratiche di scrittura autobiografica. L'opera, pubblicata in lingua francese a Parigi nel 1787, contiene un racconto retrospettivo ed in gran parte apologetico della vita del drammaturgo Carlo Goldoni, trasferitosi nella capitale francese nel 1762. Oltre all'indiscutibile valore documentario, in ottica genetica i *Mémoires* si configurano come il risultato di un «macrotesto autobiografico» (Anglani, 1995: 325), dal momento che Goldoni aveva già iniziato a raccontare della sua vita, in lingua toscana, all'interno del paratesto dell'edizione delle sue opere teatrali: dapprima nella prefazione all'edizione Bettinelli stampata a Venezia nel 1750, e poi nelle diciassette prefazioni in testa ai rispettivi tomi dell'edizione Pasquali, stampata sempre a Venezia dal 1761 al 1778.

Per la stesura dei *Mémoires*, Goldoni riprende gran parte del materiale precedentemente scritto in toscano traducendolo in francese, ma cambia i particolari di alcuni passaggi, elimina delle sequenze e ne aggiunge altre. Ne deriva dunque un'opera che possiede una propria autonomia strutturale e contenutistica, ma che risulta essere allo stesso tempo inscindibile dal proprio avantesto in lingua toscana. Per tale configurazione composita, i *Mémoires* sono stati definiti un'«autobiografia linguistica» (Oppici, 2016: 70), mettendone in luce la genesi plurilingue che richiede di essere analizzata attraverso le categorie della critica genetica. Tuttavia, una ricognizione della bibliografia critica goldoniana mostra una lacuna per quanto riguarda lo studio del plurilinguismo dell'autore, che si può trovare solo in contributi ormai datati (Folena, 1954; Folena, 1984), oppure di breve ampiezza, come la voce «Langue» redatta da D'Onghia (D'Onghia, 2019) e recentemente pubblicata all'interno del *Dictionnaire Goldoni*. Tali riferimenti critici non approfondiscono il rapporto tra il racconto autobiografico e l'impiego di una lingua seconda. Inoltre, gli studi più importanti dedicati ai *Mémoires* (Vianello, 1991; Alonge, 2009; Ferrone, 2011) citano marginalmente le prefazioni Bettinelli e Pasquali, senza considerare l'opera nella totalità della sua genesi testuale plurilingue. Solamente l'edizione Mondadori dei *Mémoires* (Bosisio, 1993) ha colto l'importanza del «macrotesto autobiografico» goldoniano, pubblicando in un unico volume il testo francese accanto alle prefazioni Bettinelli e Pasquali. Tuttavia, la mancanza di un approfondimento critico relativo al rapporto tra le differenti versioni e l'importanza gerarchica affidata ai *Mémoires*, pubblicati per primi nonostante siano l'ultima redazione del testo, intaccano l'eccellente intuizione editoriale e filologica.
A partire da tale aporia critica, il presente contributo intende suggerire delle modalità operative per la realizzazione di un'edizione genetica dei *Mémoires*, supportata dagli strumenti offerti dal Trattamento Automatico della Lingua (TAL). Nello specifico, si dimostrerà come sia possibile allineare i blocchi di testo comuni tra le prefazioni Pasquali e il testo dei *Mémoires,* nonostante la diversità linguistica. Un'edizione sinottica così prodotta permetterà di ottenere uno strumento che favorirà lo studio del rapporto tra le due redazioni, colmando così la mancanza presente negli studi goldoniani. Da questo punto di vista, la critica genetica assume un'importanza fondamentale (Sciarrino & Anokhina, 2018), dal momento che la vicinanza al testo consentirà di analizzare in maniera approfondita il plurilinguismo di Goldoni e, in generale, degli autori bilingui o plurilingui. Nel seguito del contributo verrà descritta la metodologia impiegata che ha permesso di muovere i primi passi nella realizzazione di una vera e propria edizione genetica digitale che vedrà prossimamente la luce. Nonostante i risultati siano ancora parziali, si ritiene comunque opportuno di condividerli con la comunità scientifica per stimolare un dibattito fecondo sulla creazione di edizioni genetiche digitali, supportate dal ricorso agli strumenti di TAL.

2. REALIZZARE UN'EDIZIONE GENETICA DIGITALE BILINGUE

Il testo dell'edizione francese dei *Mémoires* è stato ricavato seguendo scrupolosamente l'*editio princeps* in tre tomi pubblicata a Parigi nel 1787. Per ragioni di compatibilità con i modelli linguistici presenti attualmente, il testo è stato normalizzato in francese contemporaneo attraverso alcuni cambiamenti relativi alla morfologia e all'accentazione di alcune parole. Diverso è invece il caso dell'avantesto. Mancando dei modelli allenati su dati relativi alla varietà diacronica dell'italiano del XVIII secolo, il testo delle prefazioni Pasquali non è stato modificato ed è stato ricavato a partire dall'edizione critica realizzata da Paolo Bosisio e precedentemente citata. Il procedimento prevedeva dunque l'utilizzo di due file .txt: il primo contenente in modo continuo il testo delle diciassette prefazioni Pasquali, mentre il secondo il testo normalizzato dei *Mémoires*.

Per quanto riguarda l'allineamento testuale plurilingue, si è pensato in un primo momento calcolare la similarità tra le frasi del testo italiano con quelle del testo francese utilizzando dei modelli di *text similarity*,[1] che hanno tuttavia offerto dei risultati poco soddisfacenti. Il procedimento è stato semplice: per ogni frase del testo italiano sono stati individuati i tre migliori possibili candidati per l'allineamento che presentassero una similarità coseno maggiore rispetto ad una determinata soglia fissata a 0.4. Il valore basso della soglia, scelto per garantire la possibilità di trovare blocchi di testo comuni anche in presenza di traduzione e riformulazioni, ha però creato un numero molto alto di falsi positivi che avrebbe richiesto uno spoglio manuale estremamente lungo per verificare la correttezza degli allineamenti. In particolare, la

[1] Quello che ha fornito le prestazioni migliori e che è stato poi tenuto in considerazione per i successivi calcoli di similarità è LaBSE, la cui documentazione è fornita all'indirizzo seguente: https://huggingface.co/sentence-transformers/LaBSE (cons. 10/01/2025)

fascia critica è rappresentata dall'intervallo di frasi che hanno una similarità tra 0.50 e 0.65, per cui vengono individuati nel testo francese sia falsi positivi, che frasi soggette ad una parziale riscrittura ma che condividono comunque l'informazione del testo di partenza. In alcuni casi, invece, la frase corrispondente, che pure esiste, non viene per nulla riconosciuta, come nel caso seguente:


```
"l'imer, che pensava a sostener gl'intermezzi, e temea dell'incontro della passalacqua, fatto avea un altro acquisto.": [
      {
         "target": "mais imer croyait que l'intermède soutenait la tragi-comédie, et c'était celle-ci qui soutenait l'intermède.",
         "score": 0.5087771415710449
      },
      {
         "target": "le patron avait là des affaires, il s'y arrêta et mit pied-à-terre, je crus le moment favorable pour m'en aller;",
         "score": 0.44995665550231934
      },
      {
         "target": "je redoutais si fort ce moment terrible, qu'en sortant d'une faute j'en méditais une autre qui pouvait achever ma perte.",
         "score": 0.43072694540023804
      }
   ],
```


Tralasciando i limiti dovuti alla mancanza di un modello linguistico performante per i dati relativi all'italiano del XVIII secolo, l'imprecisione negli allineamenti suggerisce anche la particolarità del corpus preso in analisi. Essendo infatti presenti moltissime libere rielaborazioni di concetti e situazioni presenti nel testo italiano, i *Mémoires* non possono essere considerati alla stregua di una semplice traduzione e ampliamento dell'avantesto. Il corpus solleva di conseguenza anche delle problematiche a livello filologico, essendo molto difficile trovare una definizione di blocco testuale comune laddove a modificare il testo di partenza intervengono sia una traduzione che una riformulazione. Inoltre, è poco frequente la corrispondenza tra singole frasi e gli stessi concetti possono essere espressi da più periodi nel testo italiano, mentre da una sola frase nel testo francese, e viceversa. Come indicazione di metodo, è dunque necessario andare alla ricerca di punti di ancoraggio all'interno delle due versioni, piuttosto che perseguire uno scrupolo filologico volto al riconoscimento di varianti relative a singoli sintagmi o espressioni.
Da questo punto di vista, la bibliografia specifica relativa al tema dell'allineamento testuale non offre particolari spunti risolutori, dal momento che la maggior parte dei test condotti dalle autrici e dagli autori ha come obiettivo l'allineamento di corpora paralleli, vale a dire costruiti a partire da un testo sorgente e dalla sua rispettiva traduzione (Zweigenbaum & al, 2017; Repar & al, 2021; Malandri & al, 2024; Molfese & al, 2024; Reboul, 2024,[2] Xu & al, 2024). È stato dunque necessario elaborare un altro tipo di metodologia per approcciare lo studio sinottico dei due testi. Una possibile soluzione è stata suggerita dalle caratteristiche stesse del corpus. Visto lo spiccato carattere odeporico di entrambe le versioni e constatata la presenza di numerose informazioni sugli ambienti culturali e politici dell'epoca trasmesse da Goldoni, è parso opportuno considerare tali aspetti per imbastire l'operazione di allineamento bilingue. L'intuizione si basa sull'assunto che, per quanto diversi nella forma, le due versioni condividono in ogni caso gran parte della materia autobiografica che tocca necessariamente il racconto delle persone incontrate e dei luoghi visitati. Sarà così altamente significativo per il confronto tra i due testi verificare l'aggiunta, l'eliminazione o la conservazione delle informazioni nel passaggio da un'edizione all'altra.
Si è dunque deciso di ricavare automaticamente in entrambi i testi tutte le entità nominate, vale a dire i nomi di luoghi, persone, organizzazioni etc. per utilizzarle come punti di ancoraggio nell'allineamento

[2] Questo interessante contributo affronta la questione dell'allinamento bilingue utilizzando anch'esso le entità nominate come base per il riconoscimento di sequenze comuni tra il testo dell'Odissea e alcune sue traduzioni francesi. Tuttavia, per quanto possa presentare tagli e aggiunte circoscritte, il corpus scelto rientra sempre nell'ordine della traduzione. Il metodo, dunque, non può essere integralmente replicato per il confronto dei *Mémoires* con il proprio avantesto toscano, viste le numerose inversioni dell'ordine degli elementi citati e le peculiarità specifiche del corpus goldoniano.

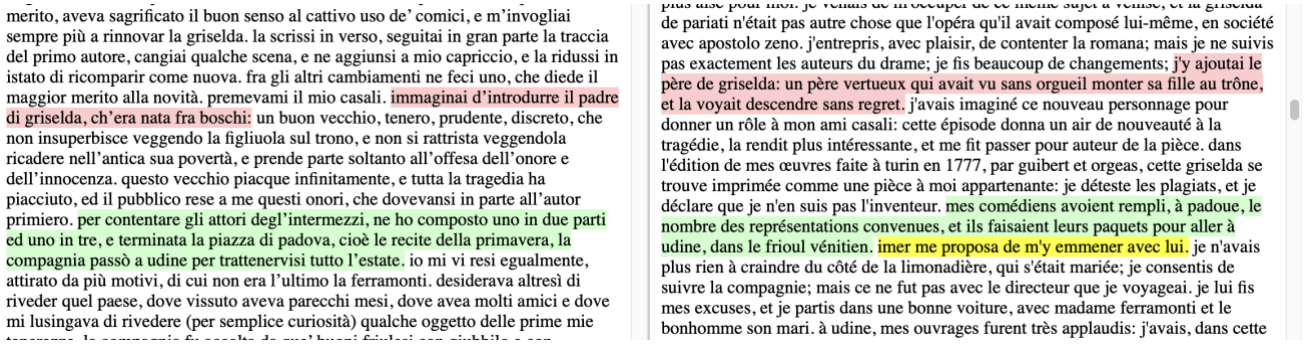
bilingue.[3] Dopo la loro estrazione, è stato necessario filtrare i risultati dall'analisi automatica dal momento che, come indicato dalla documentazione ufficiale (Zaratiana & al, 2024) i modelli di riconoscimento di entità nominate utilizzati non possiedono una precisione del 100%, raggiungendo comunque ottimi livelli di approssimazione. Inoltre, le diverse denominazioni di una medesima entità nominata all'interno dello stesso testo sono state regolarizzare assegnando a ciascun elemento un id unico. A questo punto, per ogni entità nominata riconosciuta nel testo italiano è stata individuata l'entità nominata corrispondente nel testo francese sulla base di un calcolo di similarità coseno utilizzando LaBSE. In questa fase, tutte le lettere maiuscole sono state rese minuscole, dal momento che Goldoni, secondo una prassi tipica dell'epoca, utilizza le maiuscole per designare anche alcuni nomi comuni. Un processo di normalizzazione di questo tipo permette infatti al modello di evitare delle possibili ambiguità o allucinazioni. Tale metodologia risulta essere estremamente funzionale essendo capace di ridurre al minimo i limiti derivanti dalla mancanza di modelli specifici allenati per il trattamento di dati testuali provenienti da testi italiani del XVIII secolo. Infatti, le entità nominate sono molto più facili da accoppiare tra di loro, dal momento che nel testo dei *Mémoires* non si trovano delle vere e proprie traduzioni, quanto più degli adattamenti al sistema linguistico francese che, in ogni caso, non si discosta troppo dall'originale e permette di ottenere dei risultati affidabili.
È stato così possibile accoppiare delle liste di frasi "source" per il testo italiano e "target" per il testo francese contenenti la medesima entità nominata. Per affinare ancora di più la precisione dell'allineamento, i cluster di frasi appena creati sono stati ulteriormente filtrati con la creazione di altri cluster contenenti al proprio interno due o più entità nominate. Il procedimento è molto semplice, dal momento che se una frase del testo italiano o francese è presente in più cluster, significa necessariamente che contiene al proprio interno più entità nominate. Tale procedimento si è reso necessario per filtrare le frasi che contengono al proprio interno nomi di luogo o di persona citati frequentemente, e che costituivano dunque cluster numericamente importanti. Considerando poi solo i cluster che presentavano almeno una frase nella lista "source" e una nella lista "target", sono stati calcolati gli allineamenti migliori sulla base della similarità coseno tra gli embeddings delle frasi. Ciò ha permesso di risolvere degli errori provenienti dal semplice calcolo della similarità condotto in precedenza e minimizzare le tempistiche del controllo umano per verificare la validità degli allineamenti:


```
"passalacqua (passalacqua), imer (imer)": [
      {
         "source": [
            "e la terza un'operetta in musica in sei personaggi, intitolata la fondazion di venezia, in cui cantavano l'imer, l'agnese, la passalacqua, il gandini brighella, il campagnani arlecchino, ed il mio casali cantovvi anch'egli, e si fece onore.",
            "l'imer, che pensava a sostener gl'intermezzi, e temea dell'incontro della passalacqua, fatto avea un altro acquisto."
         ],
         "target": [
            "imer la trouva à son gré, me pria d'en prendre soin, et je m'en chargeai avec plaisir, la trouvant très jolie et très docile.\n madame passalacqua en devint jalouse;",
            "mais il n'était pas trop content de madame passalacqua: sa voix était fausse, sa manière était monotone, et sa physionomie grimacière: imer voulait soutenir les intermèdes, et un musicien de l'orchestre lui en proposa le moyen."
         ]
      }
   ],
```


Calcolando così la similarità all'interno dei cluster creati e confrontando i risultati con quelli derivati dall'analisi della semplice similarità tra le frasi del testo italiano e francese, è possibile notare un notevole incremento prestazionale nel momento in cui l'allineamento viene effettuato a partire dall'estrazione di frasi contenenti entità nominate. Infatti, a fronte di 54 blocchi di testo riconosciuti da entrambi i metodi, il calcolo di similarità a partire da entità nominate ha permesso di individuare ulteriori 86 porzioni testuali

[3] Per l'estrazione delle entità nominate è stato utilizzato il modello GLiNER che rappresenta in questo momento lo stato dell'arte per il compito richiesto. In particolare, urchade/gliner_multi-v2.1 per il francese: https://huggingface.co/urchade/gliner_multi-v2.1 (cons. 10/01/2025) e DeepMount00/GLiNER_ITA_LARGE per l'italiano: https://huggingface.co/DeepMount00/GLiNER_ITA_LARGE (cons. 10/01/2025).

comuni, a fronte delle 14 del semplice calcolo di similarità coseno. Dunque, all'interno dell'allineamento testuale bilingue, l'estrazione di blocchi testuali comuni a partire da entità nominate equivalenti si rivela più efficiente del semplice calcolo di similarità coseno in oltre il doppio dei casi.
Per visualizzare i risultati, è stato creato un file .html che presenta una finestra bipartita contenente a sinistra il testo integrale delle prefazioni Pasquali e a destra quello dei *Mémoires* (Fig. 1), rispettando così l'ordine cronologico di redazione. Sono stati evidenziate in giallo le frasi individuate da entrambi i metodi, in rosso quelle esclusive del calcolo similarità e in verde quelle presenti nei cluster creati. Ciò permette di meglio comprendere il funzionamento del procedimento e, attraverso un'attenta analisi dei risultati, di migliorare in seguito il procedimento. Inoltre, cliccando su un blocco comune nella versione francese, oppure in quella italiana, è possibile allineare i due testi e studiare i blocchi comuni all'interno del loro contesto originale. Tale visualizzazione bipartita rispetta l'autonomia espressiva delle singole redazioni, dal momento che il testo italiano e quello francese posso essere letti autonomamente, mettendo in luce allo stesso tempo il legame che tra di esse intercorre.


merito, aveva sagrificato il buon senso al cattivo uso de' comici, e m'invogliai sempre più a rinnovar la griselda. la scrissi in verso, seguitai in gran parte la traccia del primo autore, cangiai qualche scena, e ne aggiunsi a mio capriccio, e la ridussi in istato di ricomparir come nuova. fra gli altri cambiamenti ne feci uno, che diede il maggior merito alla novità. premevami il mio casali. immaginai d'introdurre il padre di griselda, ch'era nata fra boschi: un buon vecchio, tenero, prudente, discreto, che non insuperbisce veggendo la figliuola sul trono, e non si rattrista veggendola ricadere nell'antica sua povertà, e prende parte soltanto all'offesa dell'onore e dell'innocenza. questo vecchio piacque infinitamente, e tutta la tragedia ha piacciuto, ed il pubblico rese a me questi onori, che dovevansi in parte all'autor primiero. per contentare gli attori degl'intermezzi, ne ho composto uno in due parti ed uno in tre, e terminata la piazza di padova, cioè le recite della primavera, la compagnia passò a udine per trattenervisi tutto l'estate. io mi vi resi egualmente, attirato da più motivi, di cui non era l'ultimo la ferramonti. desiderava altresì di riveder quel paese, dove vissuto aveva parecchi mesi, dove avea molti amici e dove mi lusingava di rivedere (per semplice curiosità) qualche oggetto delle prime mie

de pariati n'était pas autre chose que l'opéra qu'il avait composé lui-même, en société avec apostolo zeno. j'entrepris, avec plaisir, de contenter la romana; mais je ne suivis pas exactement les auteurs du drame; je fis beaucoup de changements; j'y ajoutai le père de griselda: un père vertueux qui avait vu sans orgueil monter sa fille au trône, et la voyait descendre sans regret. j'avais imaginé ce nouveau personnage pour donner un rôle à mon ami casali: cette épisode donna un air de nouveauté à la tragédie, la rendit plus intéressante, et me fit passer pour auteur de la pièce. dans l'édition de mes œuvres faite à turin en 1777, par guibert et orgeas, cette griselda se trouve imprimée comme une pièce à moi appartenante: je déteste les plagiats, et je déclare que je n'en suis pas l'inventeur. mes comédiens avoient rempli, à padoue, le nombre des représentations convenues, et ils faisaient leurs paquets pour aller à udine, dans le frioul vénitien. imer me proposa de m'y emmener avec lui. je n'avais plus rien à craindre du côté de la limonadière, qui s'était mariée; je consentis de suivre la compagnie; mais ce ne fut pas avec le directeur que je voyageai. je lui fis mes excuses, et je partis dans une bonne voiture, avec madame ferramonti et le bonhomme son mari. à udine, mes ouvrages furent très applaudis: j'avais, dans cette

Figura1. Esempio di blocchi testuali comuni nell'edizione sinottica dei *Mémoires*.

La codifica non stabilisce una gerarchia tra il testo dei *Mémoires* e quello delle prefazioni Pasquali. Le porzioni di testo sono inserite all'interno di tag con un id alfanumerico contenente la sigla della lingua e un numero, una classe per mettere distinguere i vari colori e l'evento onclick che grazie ad una semplice funzione chiamata alignText permette di allineare i due blocchi testuali. Di seguito due esempi completi presenti all'interno del file:

> predicava dall'insigne pulpito del duomo di detta città il padre jacopo cataneo, agostiniano scalzo milanese.
>
> je vais le jour des cendres à la cathédrale pour entendre le père cattaneo, augustin réformé, et je trouve son sermon admirable;

Per questo motivo, un'eventuale codifica in formato XML in cui siano confrontati i due testi non dovrà seguire le linee guida della TEI per quanto riguarda l'apparato critico e la rappresentazione delle varianti,[4] dal momento che quest'ultima segue un principio verticale di subordinazione rispetto ad un testo principale. Piuttosto, risulterebbe filologicamente più adeguato l'utilizzo dell'elemento <seg>[5] per circondare il blocco testuale scelto, unitamente all'attributo @corresp,[6] che presenterà un valore grazie a cui sarà possibile associarlo alla propria sequenza parallela. La natura arbitraria dell'elemento è consona rispetto al carattere eterogeneo dei blocchi paralleli, che sono stati riscontrati spesso al livello della frase; non corrispondono in nessun caso a dei paragrafi comuni e nemmeno a delle porzioni più ampie. Per questo, è da escludere l'assegnazione dell'attributo @corresp a elementi come <p>[7] e [8] che individuano al contrario lunghe sequenze di testo. Per quanto riguarda la visualizzazione dei risultati,

[4] https://tei-c.org/release/doc/tei-p5-doc/en/html/TC.html#index-body.1_div.13_div.4 (cons. 03/04/2025).

[5] https://tei-c.org/release/doc/tei-p5-doc/en/html/ref-seg.html (cons. 04/04/2025).

[6] https://tei-c.org/release/doc/tei-p5-doc/it/html/ref-att.global.linking.html (cons. 04/04/2025).

[7] https://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-p.html (cons. 04/04/2025).

[8] https://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-div.html (cons. 04/04/2025).

potrebbe inoltre essere esplorata la possibilità di utilizzare EVT 3.0.[9] Sarebbe tuttavia necessario implementare manualmente dei parametri specifici e personalizzati per l'allinamento dei segmenti e la messa in evidenza dei blocchi testuali comuni, dal momento che un numero molto elevato di edizioni genetiche digitali che utilizzano il software EVT si basano sulla logica dell'apparato critico e dei diversi testimoni. Al contrario, il corpus del «macrotesto autobiografico» goldoniano solleva delle problematiche filologiche che richiedono un altro tipo di codifica, più neutra e meno gerarchicamente definita.

3. CONLCUSIONI

È stata presentata in questo contributo la prima versione di un'edizione genetica digitale dei *Mémoires* di Carlo Goldoni, realizzata operando un allineamento testuale bilingue dei blocchi di testo comuni ottenuti grazie all'estrazione e al filtraggio delle entità nominate nei due testi componenti il corpus. La visualizzazione sinottica ottenuta permette già di iniziare a formulare alcune considerazioni critiche circa il rapporto stilistico ed espressivo che intercorre tra le due versioni. Resta ancora da verificare la possibilità di allineare le porzioni di testo comuni che non contengono entità nominate, in modo tale da creare uno strumento che possa poi essere riutilizzato per lo studio della genesi plurilingue di alcune opere e non solo. In ogni caso, si è dimostrato che attraverso un procedimento semplice, facilmente replicabile e che non richiede il *fine-tuning* di uno o più modelli specifici sia possibile allineare blocchi di testo anche molto diversi tra di loro, ma che condividono le stesse informazioni relative alle entità nominali.
Infine, in ottica contrastiva, un'ulteriore implementazione potrebbe prevedere una visualizzazione differente per i blocchi testuali esclusivi dell'una o dell'altra redazione, in modo tale che l'utente possa agevolmente comprendere lo sviluppo diacronico del testo e il rapporto stilistico e contenutistico tra le due versioni. Una volta ottenuta un'edizione genetica digitale completa, verrà messo a libera disposizione della comunità scientifica uno strumento indispensabile per colmare una lacuna presente negli studi goldoniani. Più in generale, la condivisione della metodologia impiegata per la realizzazione tecnica dell'allineamento testuale bilingue permetterà anche ad altri specialisti di realizzare edizioni genetiche per lo studio di scrittori bilingui o plurilingui. Sono già attualmente disponibili in rete delle ottime edizioni genetiche digitali, che si limitano però all'analisi di opere la cui genesi interessa una sola lingua.[10] Nel solco di questi progetti già avviati, la futura edizione dei *Mémoires* goldoniani intende essere un punto di riferimento per la realizzazione di edizioni genetiche bilingui o plurilingui ricorrendo agli strumenti di trattamento automatico della lingua.

BIBLIOGRAFIA

Alonge, R. (2009). Goldoni Mémoires: A Parigi si vive meglio. Problemi di Critica Goldoniana. XVI : numero speciale : Terzo centenario della nascita di Carlo Goldoni e Secondo centenario della morte di Carlo Gozzi : tomo terzo, 2009, 1000–1012. https://doi.org/10.1400/130551

Anglani, B. (1995). I « Mémoires»: Bilanci e prospettive. Alberti C. & Pizzamiglio G. (Eds.). Carlo Goldoni 1793-1993: Atti del Convegno del bicentenario, Venezia, 11-13 aprile 1994, 325–339.

D'Onghia L. (2016) Langue. Dictionnaire Goldoni. Comparini L. & Fabiano A. (Eds) Parigi: Garnier. 109–114.

Ferrone, S. (2011). La vita e il teatro di Carlo Goldoni. Venezia: Marsilio.

Folena, G. (1958). L'esperienza linguistica di Carlo Goldoni. Lettere Italiane, X (1), 21–54.

Folena, G. (1984). L'italiano in Europa. Esperienze linguistiche del Settecento. Torino: Einaudi.

Goldoni, C. (1993). Memorie. Bosisio, P. & Ranzini, P. (Eds.). Milano: Mondadori.

Malandri, L., Mercorio, F., Mezzanzanica, M., & Pallucchini, F. (2024). SeNSe: Embedding alignment via semantic anchors selection. International Journal of Data Science and Analytics. https://doi.org/10.1007/s41060-024-00522-z

Molfese, F., Bejgu, A., Tedeschi, S., Conia, S., & Navigli, R. (2024). CroCoAlign: A Cross-Lingual, Context-Aware and Fully-Neural Sentence Alignment System for Long Texts. Proceedings of the 18th

[9] https://iris.unito.it/handle/2318/1896873 (cons. 04/04/2025).

[10] Si pensi al lavoro sulle varianti de *Il nome della rosa*: https://www.variantidellarosa.it (cons. 10/01/2025) oppure alla sezione "Génétique" del portale *eBalzac*: https://www.ebalzac.com/genetique (cons. 10/01/2025).

Conference of the European Chapter of the Association for Computational Linguistics. Graham, Y. & Purver, M. (Eds.). San Giuliano: Association for Computational Linguistics, 2209-2220. https://aclanthology.org/2024.eacl-long.135/

Oppici, P. (2016). I Mémoires di Goldoni come autobiografia linguistica. Regards croisés France-Italie: Langues, écritures et cultures: Atti della giornata di studi organizzata dalle Università di Macerata e Clermont-Ferrand. Bisconti D. & Fabiani, D. (Eds.). Macerata: EUM, 69-80.

Reboul, M. (2023). Étude semi-automatique des traductions d'Homère avec "Odysseus" in Notre Homère. Deloince-Louette, C. & Salha, A. (Eds.). Grenoble: UGA Éditions, 163-182. https://doi.org/10.4000/books.ugaeditions.32951

Repar, A., Martinc, M., Ulčar, M., & Pollak, S. (2021). Word-embedding based bilingual terminology alignment. Electronic lexicography in the 21st century. Proceedings of the eLex 2021 conference. 5–7 July 2021. Kosem, I. et al. (Eds.). Brno: Lexical Computing CZ. https://doi.org/10.5281/zenodo.5547982

Sciarrino, E., & Anokhina, O. (2018). Plurilinguisme littéraire: De la théorie à la genèse. Genesis, 46, 11-34. https://doi.org/10.4000/genesis.2554

Vianello M. (1991). L'avvocato in commedia, Goldoni e l'autobiografia. Studi Veneziani, 22, 325–358.

Xu, Y., Hu, L., Zhao, J., Qiu, Z., XU, K., Ye, Y., & Gu, H. (2024). A Survey on Multilingual Large Language Models: Corpora, Alignment, and Bias. Frontiers of Computer Science. https://doi.org/10.1007/s11704-024-40579-4

Zaratiana, U., Tomeh, N., Holat, P., & Charnois, T. (2024). GLiNER: Generalist Model for Named Entity Recognition using Bidirectional Transformer. Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Duh, K., Gomez, H., & Bethard, S. (Eds.). Città del Messico: Association for Computational Linguistics, 5364-5376 https://doi.org/10.18653/v1/2024.naacl-long.300

Zweigenbaum, P., Sharoff, S., & Rapp, R. (2017). Overview of the Second BUCC Shared Task: Spotting Parallel Sentences in Comparable Corpora. Proceedings of the 10th Workshop on Building and Using Comparable Corpora. Sharof, S., Zweigenbaum, P., Rapp, P. (Eds.) Vancouver: Association for Computational Linguistic, 60-67. https://doi.org/10.18653/v1/W17-2512

Eastern Law in Western Words: Analyzing Roman Legal Terminology in Medieval Charters

Tamás Kovács[1], Angelos Nikolaou[2], Johannes Laroche[3], Georg Vogeler[4]

[1] Department of Digital Humanities, University of Graz, Austria - tamas.kovacs@uni-graz.at
[2] Department of Digital Humanities , University of Graz, Austria - angelos.nicolaou@uni-graz.at
[3] Department of Digital Humanities , University of Graz, Austria – johannes.laroche@uni-graz.at
[4] Department of Digital Humanities , University of Graz, Austria – georg.vogeler@uni-graz.at

ABSTRACT (ENGLISH)

Monasterium.net provides a dataset of over 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 machine learning methods from the fields of natural language processing (NLP). We propose a method for analyzing the reception of Roman law in medieval Central Europe through examining diplomatic sources. Corpus-wide token encoding and subsequence generation via strategic omission of n positions are implemented in this research using Leave-N-Out grams (LNO-grams). The system's array-based encoding enhances this process by generating integer-encoded patterns, where each row represents a different skip combination. This unified representation allows the Intersection over Union (IoU) metric to consider all possible matching patterns, capturing both local variations and long-range dependencies. We demonstrate our new method with a case study of Roman law's reception in medieval Central Europe, showcasing the evolution and regional differences in legal language.

Keywords: Roman Law Reception, Digital Diplomatics, Legal Terminology, Text Analysis

ABSTRACT (ITALIANO)

Il diritto orientale in parole occidentali: Analizzare la terminologia giuridica romana nelle carte medievali

Monasterium.net fornisce un dataset di oltre 600.000 carte europee. Per utilizzare questo set di dati per questioni diplomatiche su scala europea, invece di concentrarsi su studi individuali, regionali o specifici per le collezioni, sono necessari metodi automatici. Nel progetto ERC "From digital to distant diplomatics", combiniamo gli approcci tradizionali della diplomazia con metodi di apprendimento automatico provenienti dal campo dell'elaborazione del linguaggio naturale (NLP). Proponiamo un metodo per analizzare la ricezione del diritto romano nell'Europa centrale medievale attraverso l'esame delle fonti diplomatiche. La codifica dei token a livello di corpus e la generazione di sottosequenze attraverso l'omissione strategica di n posizioni sono implementate in questa ricerca utilizzando i Leave-N-Out grams (LNO-grams). La codifica basata su array del sistema migliora questo processo generando modelli codificati con numeri interi, dove ogni riga rappresenta una diversa combinazione di salti. Questa rappresentazione unificata permette alla metrica Intersection over Union (IoU) di considerare tutti i possibili pattern di corrispondenza, catturando sia le variazioni locali che le dipendenze a lungo raggio. Dimostriamo il nostro nuovo metodo con un caso di studio sulla ricezione del diritto romano nell'Europa centrale medievale, mostrando l'evoluzione e le differenze regionali nel linguaggio giuridico.

Parole chiave: Ricezione del diritto romano, Diplomatica digitale, Terminologia giuridica, Analisi del testo

1. INTRODUCTION

This study presents a new computational approach to analyzing how Roman law was received in medieval Central Europe, using diplomatic sources. Traditional historical scholarship has extensively explored the influence of Justinian's Corpus Iuris Civilis on medieval legal practice, yet our understanding of its practical implementation across different temporal and geographical contexts remains limited. The proposed research addresses this gap by applying sophisticated computational analysis to medieval charters from Central European archives, focusing on Austrian documents from the period 1200-1400 CE. Central Europe's key Roman law reception period falls within this research timeframe, providing ample documented evidence for meaningful computational study.

Medieval charters relied heavily on formulaic language, using pre-established templates for legal actions that represented conventionalized pairings of form and function. However, these formulas present a moving target for modern interpretation, exhibiting significant variation across time and regions. This variability stems from several factors, including the evolution of legal practices, regional dialects, and individual scribal preferences. One legal concept could be expressed differently in various historical documents. For example, Papal documents use "appellatione remota", German Episcopal charters use "appellatione cessante", and French ecclesiastical courts used "sine appellationis obstaculo." Such variations complicate the identification and analysis of formulaic language in an extensive corpus of documents.

2. STATE OF ART

This inherent variability poses a challenge for traditional computational analysis methods. Methods based on frequency, counting identical word sequences, cannot easily identify formulas written differently. Similarly, inconsistent spelling, grammar, and word order impede the ability of N-gram approaches to identify formulaic expressions. N-grams, which analyze sequences of *n* consecutive words, are sensitive to even minor variations in the text (Manning et al., 2008) Medieval scribes sometimes add case-specific details to formulas, like "*causa que vertitur inter monasterium sancti Benedicti et homines ville de Capraria*" (the case between St. Benedict's monastery and Capraria villagers). Standard n-gram methods cannot recognize the resulting patterns (Korkiakangas, 2024; Korkiakangas & Passarotti, 2011). This is important because these insertions were common practice (Koolen & Hoekstra, 2022), reflecting a need for formula adaptation in specific legal scenarios.

Although skip-grams offer more flexible matching (Mikolov et al., 2013), their limited local context restricts their capacity to model the long-range dependencies vital to understanding legal formulas. Analyzing word sequences with gaps, skip-grams handle variability to a degree, but are limited by their context window size. Intricate legal formulas, involving complex conditional statements and their consequences, can surpass standard skip-gram window sizes. Long formulas, such as "*si non omnes his exequendis potueritis interesse, duo vestrum ea nichilominus exequantur,*" may extend beyond typical window sizes. The method may overlook key relationships between elements such as "*si non omnes*" and "*duo vestrum.*"

3. METHOD

The methodological innovation at the core of this research is FLAME (Formulaic Language Analysis in Medieval Expressions, https://github.com/kreeedit/FLAME (in progress)), a computational system specifically designed for analyzing formulaic language in medieval legal documents. FLAME's novel approach, LNO-grams, identifies legal formulas regardless of variations in expression, spelling, or word order. The value of this method is apparent in medieval texts, given the significant scribal variations in legal terminology. The system's adaptable design enables it to recognize comparable legal ideas despite variations in wording or structure.

FLAME uses a complex two-stage procedure. First, it implements a corpus-wide token encoding scheme that translates words into integer sequences for efficient pattern matching while retaining contextual sensitivity. This approach makes it possible to recognize similar legal terms regardless of context or spelling variations. The encoding method accounts for both individual words and their contextual relationships, thus enabling more subtle pattern recognition than keyword methods. Secondly, it uses a Leave-N-Out pattern generation system which, for a text sequence of length s, creates subsequences by removing n strategic positions. This method maintains the fundamental framework of legal formulas while allowing for the diverse adaptations made by medieval scribes. Despite different wording and structures, the system can still pinpoint similar legal concepts.

The system uses array-based encoding to improve this process; it creates integer-encoded patterns, with each row showing a unique skip combination. The unified representation lets the IoU metric assess all potential matching patterns immediately, encompassing both local and long-range aspects. Although FLAME presently employs IoU, its architecture is adaptable to more refined similarity evaluations, encompassing the weighting of diverse match categories as complete matches, partial matches, word order discrepancies, and lexical replacements. The adaptable nature of similarity assessment enables researchers to tailor their analyses to specific research questions or document features.

The visualization aspect of FLAME has shown to be especially helpful in historical analysis. The system creates

interactive HTML that compares similar texts and visualizes similarity scores with heatmaps across the corpus. Researchers can use these tools to explore matching document sections and discover “bridge words” connecting similar legal terms. Researchers can use the visualization system’s temporal analysis features to study how legal terminology evolves across institutions and over time. Distant reading reveals stable and variable parts of charter formulas, aiding the study of their changes. Researchers can interactively explore data within the visualizations, ranging from large-scale institutional patterns down to fine-grained textual differences.

This research uses computational and traditional historical analysis in creative ways. The research analyzes charters available on monasterium.net from Central European archives. The digital availability of these documents through monasterium.net has made it possible to conduct large-scale analysis of medieval legal documentation, though the computational processing of these materials presents its own challenges. Synchronic patterns in legal terminology adoption across institutions are identified through our use of time series clustering analysis. Viewing specific legal terms as temporal data points, this method reveals previously unseen patterns in how Roman law was received.

Several key steps are involved in the systematic processing of the charter corpus. Documents are initially classified by their source institution, date of creation, and document type. This categorization facilitates the study of adoption trends across various medieval institutions. Second, spelling variations and abbreviations in the medieval Latin texts are standardized through preprocessing. Finally, in step three, FLAME detects and trace patterns of legal terms used in the standardized texts, and the results are analyzed using both computational and traditional historical methods to ensure that identified patterns are historically meaningful.

This analysis might identify different chancery groups based on their particular legal terminology patterns (Härtel, 2015). Early adoption of Roman legal concepts and terminology might be seen in Episcopal chanceries (Willoweit, 2000). This early embrace may be due to their links with Italian legal education and deep understanding of Justinian’s principles (Huschner, 2003). Royal courts could adopt a more measured, methodical approach to implementing new legal terms, combining Roman law with existing practices (Bates, 1995). This measured approach may indicate a calculated strategy to modernize the legal system while preserving established practices. Roman law’s adoption by municipalities could have been delayed and selective, prioritizing practical application. Cities with strong trade links might have adopted Roman legal terms earlier than isolated communities, showing regional differences in adoption patterns.

Analyzing these clusters may reveal if Roman law terminology spread institutionally or geographically in medieval Europe, thus challenging current assumptions about legal innovation dissemination. A temporal analysis of Roman legal term adoption across institutions could pinpoint peaks, thereby indicating whether widespread legal reforms, influential scholars, or other factors drove the process. Discovered patterns may or may not align with documented legal reforms and new schools, possibly revealing the legal transformation coordination. The timeline of these patterns might also reveal links to major historical shifts, including the creation of new universities and the exchange of legal experts between different regions. This analysis will help determine the most significant factors —institutional, geographical, or otherwise—behind the adoption of Roman legal terminology.

FLAME demonstrates significant potential in analyzing medieval legal texts, but limitations exist, suggesting avenues for further development. Currently, the system’s structural analysis ignores legal meaning, which may lead to functionally distinct formulas being grouped together due to structural similarities. This limitation is resolvable through the integration of semantic analysis tools and domain-specific knowledge bases. Its computational cost rises with corpus size, and it demands careful parameter adjustments depending on the document type. Ongoing development of more efficient processing algorithms and more sophisticated similarity metrics is needed to overcome these technical challenges. Future work will concentrate on improving FLAME’s analytical skills by adding semantic analysis tools and specialized legal knowledge, thus enhancing its ability to differentiate between structurally similar but functionally distinct legal formulas.

However, the proposed method offers substantial contributions to the fields of digital humanities method and medieval legal history. First, it showcases how computational methods effectively analyze large medieval legal document collections; this offers a repeatable model for similar research across other times and places. The research’s method is adaptable to analyzing other historical documents, especially those using formulaic language or standardized expressions. Second, it reveals new understandings of how Roman law was adopted in medieval

Central Europe, showing a more intricate, institutionally led process than previously thought. This research challenges traditional views of Roman law dissemination, suggesting fresh approaches to understanding the interplay between institutions and legal transformations. A third benefit is a new computational tool for analyzing formulaic language in historical documents; its uses extend beyond legal texts. The adaptable pattern-matching and similarity assessment methods of the FLAME system could prove useful in analyzing other formulaic historical documents.

ACKNOWLEDGEMENTS

All of the contributing authors are part of the ERC project "From Digital to Distant Diplomatics", located at the Department of Digital Humanities of the University of Graz. Tamás Kovács has developed the concept and the code of the presented approach. The code development benefitted from support by Angelos Nikolaou. Johannes Laroche contributed to the work with data evaluation and diplomatics research, 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

Bates, D. (1995). Le rôle des évêques dans l'élaboration des actes ducaux et royaux entre 1066 et 1087. *Les Évêques Normands Du XIs. : Colloque de Cerisy-La-Salle, 30 Septembre-3 Octobre 1993*, 103–115.

Härtel, R. (2015). Urkundenlandschaften zwischen Donau, Rhein und Adria. In *Andreas Schwarcz/Katharina Kaska (Hgg.), Urkunden – Schriften – Lebensordnungen. Neue Beiträge zur Mediävistik. Vorträge der Jahrestagung des Instituts für Österreichische Geschichtsforschung aus Anlass des 100. Geburtstags von Heinrich Fichtenau (1912–2000) (Wien, 13.–15. Dezember 2012)* (Vol. 63, pp. 193–211). Beck.

Huschner, W. (2003). *Transalpine Kommunikation im Mittelalter: Diplomatische, kulturelle und politische Wechselwirkungen zwischen Italien und dem nordalpinen Reich (9.-11. Jahrhundert)* (Vol. 52). Hahnsche Buchhandlg.

Koolen, M., & Hoekstra, F. G. (2022). Detecting Formulaic Language Use in Historical Administrative Corpora. In F. Karsdorp, A. Lassche, & K. Nielbo (Eds.), *Proceedings of the Computational Humanities Research Conference 2022* (pp. 127–151).

Korkiakangas, T. (2024). *A linguist's viewpoint: Formulaic language as a challenge for historical linguistics*. https://doi.org/10.5281/ZENODO.10461847

Korkiakangas, T., & Passarotti, M. (2011). Challenges in Annotating Medieval Latin Charters. *Journal for Language Technology and Computational Linguistics*, *26*(2), 105–116. https://doi.org/10.21248/jlcl.26.2011.150

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge university press.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *arXiv:1301.3781 [Cs]*. http://arxiv.org/abs/1301.3781

Shannon, C. E. (1948). A Mathematical Theory of Communication. *The Bell System Technical Journal*, *27*, 379–423, 623–656.

Willoweit, D. (2000). Grundherrschaft und Territorienbildung. Landherren und Landesherren in deutschsprachigen Urkunden des 13. Jahrhundert. In *Strukturen und Wandlungen der ländlichen Herrschaftsformen vom 10. Zum 13. Jahrhundert. Deutschland und Italien im Vergleich, hg. V. Gerhard DILCHER and Cinzio VIOLANTE* (Vol. 14, pp. 215–233). Duncker & Humblot.

Il corpus del Digesto: approcci e metodi computazionali per la creazione di risorse linguistiche

Alessandra Cinini[1], Paola Marongiu[2], Eva Sassolini[3]
[1]Istituto di Linguistica Computazionale “A. Zampolli”, (CNR - Pisa), Italia, alessandra.cinini@cnr.it
[2]Istituto di Linguistica Computazionale “A. Zampolli”, (CNR - Pisa), Italia, paola.marongiu@cnr.it
[3]Istituto di Linguistica Computazionale “A. Zampolli”, (CNR - Pisa), Italia, eva.sassolini@cnr.it

ABSTRACT[1] (ITALIANO)

Il contributo vuole discutere della creazione, elaborazione e uso di corpora testuali bilingui (latino-italiano) paralleli a supporto della traduzione specialistica in ambito giuridico.
L’occasione che ci ha permesso di affrontare i diversi aspetti sottesi a questo tipo di iniziative è stato il progetto di traduzione in italiano del Digesto di Giustiniano, del quale il nostro gruppo di ricerca è partner tecnologico nel progetto PRIN 2022 PNRR “TESTO”. Parallelamente alla creazione del corpus bilingue sincronizzato sono stati implementati una serie di strumenti di supporto alla traduzione da offrire agli esperti di dominio: memorie di traduzione (*Translation Memory*) disponibili online in modalità contrastiva dei libri già tradotti; annotazione linguistica dei testi con l’obiettivo di predisporre un glossario terminologico estratto che rappresenti un punto di partenza per la creazione di un lessico giuridico bilingue.
Chiavi: Linguistica Computazionale; sincronizzazione di testi; risorse linguistiche di dominio; estrazione di termini

ABSTRACT (ENGLISH)

The Corpus of the Digest: Computational Approaches and Methods for the Creation of Linguistic Resources.
The paper aims to describe the creation, development and use of parallel bilingual (Latin-Italian) text corpora to support specialised translation in the legal field. The PRIN 2022 PNRR project ‘TESTO’ on the Italian translation of Justinian's Digest, in which our research group is a technology partner, allowed us to address the various aspects underlying this type of initiative. In addition to the creation of the parallel bilingual corpus, we implemented a series of translation support tools to be offered to domain experts: translation memories of the texts already translated, available online in contrastive mode; linguistic annotation of the texts to prepare an extracted terminological glossary, identified as a starting point for the creation of a bilingual legal lexicon.
Keywords: Computational linguistics; texts alignment; domain linguistic resources; term extraction

1. INTRODUZIONE

Il progetto TESTO ha come obiettivo primario il completamento della traduzione in lingua italiana del Digesto di Giustiniano. Il contesto è quindi quello della traduzione specialistica in ambito giuridico, affrontata da un gruppo/team di traduttori con competenze di dominio. La traduzione ha evidenziato in passato e presenta anche oggi diversi aspetti peculiari che ne hanno dilatato i tempi. Infatti, da un lato le grandi dimensioni del corpus testuale (50 libri per migliaia di pagine), dall’altro le lingue coinvolte e la natura specialistica dei contenuti rappresentano ancora oggi una sfida dal punto di vista metodologico. L’obiettivo dei traduttori e del prof. Sandro Schipani, che ha guidato per anni il progetto, è sempre stato quello di creare un ponte comunicativo tra culture diacronicamente molto distanti, permettendo allo stesso tempo la valorizzazione di profili comuni. La traduzione del Digesto richiede la comprensione del testo originario per garantire la salvaguardia della struttura logico-argomentativa e delle sfumature linguistiche anche nel testo in traduzione (Addis 2023). Nello scenario descritto la presenza di risorse di supporto alla traduzione (memorie di traduzione e/o lessici terminologici bilingui) è fortemente auspicabile, ma di difficile reperimento poiché sistemi di supporto consolidati adatti al compito specifico non sono facilmente disponibili. Con questo obiettivo l’Istituto di Linguistica Computazionale “A. Zampolli” del CNR (CNR-ILC) ha offerto il proprio contributo tecnologico nella realizzazione di un corpus parallelo bilingue a partire dai libri già tradotti, che ha rappresentato poi la base dalla quale poter derivare altre risorse linguistiche a supporto della traduzione. Si è trattato di un processo incrementale che vede oggi il corpus parallelo coprire circa l’80% dell’intera opera. Le problematiche legate alla traduzione di un testo come il Digesto hanno richiesto la ricerca di soluzioni specifiche i cui risultati verranno analizzati per determinare quali strumenti software possano supportare adeguatamente il processo di traduzione, senza però sostituire il traduttore umano nei

[1] Nel quadro di un’elaborazione comune, Eva Sassolini, come responsabile delle attività scientifiche del progetto, ha curato l’impianto, la revisione complessiva dell’articolo e la redazione dei §§ 1 e 4, Alessandra Cinini, come responsabile della banca dati, ha invece curato i §§ 2 e 3, Paola Marongiu, vincitrice di un assegno di ricerca sul progetto, ha curato i §§ 2.1, 3.1 e 4.1.

domini specialistici. Inoltre, sarà discusso l'approccio utilizzato che combina l'intervento manuale agli strumenti computazionali.

2. LA RISORSA BILINGUE: CONTESTO E PROBLEMATICHE DI TRADUZIONE

Il Digesto di Giustiniano (*Digesta* nel suo nome latino), insieme con il *Codex Iustiniani*, le *Iustiniani Institutiones* e le *Novellae Constitutiones*, fanno parte del *Corpus Iuris Civilis*, un progetto realizzato sotto l'ordine dell'imperatore Giustiniano (527–566 d.C.) tra il 529 e il 534 d.C., e che raccoglie un ricco numero di testi della giurisprudenza romana (Banchich et al. 2015; Dingledy, 2016). Il Digesto rappresenta la parte più importante del *Corpus Iuris Civilis*, poiché offre la più ricca collezione di testi giuridici romani. Il Digesto comprende infatti estratti selezionati da 1528 scritti prodotti da 39 giuristi romani, coprendo 800 anni di storia giuridica romana (Dingledy, 2016: 5; Ribary & McGillivray, 2020). Gli scritti dei giuristi sono organizzati in 50 libri, suddivisi in cinque gruppi a seconda dell'argomento: Diritto Pubblico (Libro I); Diritto Privato (Libri II–XLVII); Diritto Penale (Libro XLVIII); Procedura d'Appello e Tesoro (Libro XLIX); Diritto Municipale, Diritto Specializzato e Definizioni (Libro L). I libri sono suddivisi in 432 titoli, ogni titolo è diviso in leggi e queste ultime in paragrafi.
Lo studio del Digesto per i giuristi di oggi serve per comprendere e sviluppare un sistema di principi sui quali si basa il diritto di ogni Stato, indipendentemente dai rispettivi sistemi giuridici (Schipani 2005: xxvii).

2.1 Il testo latino e la traduzione

Il tipo di latino utilizzato nel Digesto è altamente tecnico-specialistico e appartiene ai cosiddetti linguaggi di dominio, o lingue speciali (Cortelazzo 1994; Mazzini 2009). Di conseguenza, presenta caratteristiche specifiche a livello lessicale, sintattico e stilistico. Nella traduzione di testi specialistici, il traduttore si trova sempre a mediare tra una traduzione *source-oriented* (che punta al rispetto del testo e della struttura della lingua di partenza) e una *target-oriented* (che è invece orientata al fruitore del testo in traduzione, alla lingua e alla cultura target). Nel primo caso si adottano traduzioni estremamente fedeli al testo fonte, e quindi si opta per l'utilizzo di calchi o prestiti per esprimere concetti tecnici che non sono presenti nella cultura e/o nel sistema giuridico della lingua target. Nel secondo caso invece concetti specifici del testo fonte vengono interpretati nella lingua target, utilizzando ad esempio neologismi, parafrasi e rielaborazioni più o meno invasive della sintassi della lingua di partenza (Paolucci 2013; Paolucci 2017:325;328). Al contesto descritto si aggiungono altri due fattori di complessità. Il primo è legato alla lingua del testo fonte, ovvero il latino, una lingua storica e quindi non più parlata, utilizzata per descrivere un sistema legale non più esistente; il secondo riguarda lo status peculiare del Digesto, che si colloca a metà strada tra un'opera letteraria e un testo di dominio.
Particolare attenzione è posta inoltre sulla traduzione dei termini tecnici e sulla coerenza terminologica lungo tutta l'opera tradotta. Per quanto riguarda il primo aspetto, nonostante si cerchi di rispettare la sintassi di base del testo latino, la lingua esprime i termini/concetti in modo più conciso rispetto all'italiano, utilizzando talvolta delle ellissi o sottintendendo il referente; inoltre esistono concetti non presenti di fatto nella lingua italiana, che devono essere quindi resi attraverso una parafrasi. Ne sono mostrati esempi in Figura 1.

Tipo aggiunta	Definizione	Esempio
Morfosintattica	Veicola informazioni grammaticali che nel latino vengono espresse a livello morfologico, o che sono richieste dalla costruzione sintattica in italiano	(LAT) *Si heres* ***praecepto fundo*** *rogatus sit hereditatem restituere [...]* (ITA) Se l'erede, <**a cui**> **sia stato prelegato un fondo**, sia stato richiesto <mediante fedecommesso> di dare l'eredità [...]
Stilistica	Aggiunge elementi non necessari alla grammaticalità della frase in italiano, ma utili al miglioramento della traduzione dal punto di vista stilistico	(LAT) *Ius civile est, quod in totum* ***a naturali vel genitum*** *recedit [...]* (ITA) Il diritto civile... non si discosta in tutto **dal diritto naturale o <dal diritto>** delle genti [...]
Di completamento	Aggiunge pochi elementi/ dettagli che aiutano il lettore nella comprensione del (con)testo, a volte includendo elementi dedotti dal contesto precedente	(LAT) *[...] ab intestato* ***rogatus****que* ***sit*** *restituere hereditatem [...]* (ITA) [...] e **sia stato richiesto** <**per fedecommesso**> di dare l'eredità [...]
Informativa	Aggiunge contenuto per favorire la comprensione del lettore, esplicitando informazioni non presenti nel testo latino	(LAT) [...] *nisi liberalitatem tantum ad priorem fideicommissarium heres* ***voluit*** *pertinere.* (ITA) [...] a meno che l'erede **<che pur ha ugualmente diritto di trattenerla>** non abbia voluto **<e dichiarato>** che <questa sua> liberalità spetti solo al primo fedecommissario.

Figura 1. Tipologie di aggiunte individuate nel Digesto

In questi casi, i traduttori adottano un approccio interpretativo che si esplica anche con aggiunte, cioè delle inserzioni testuali segnalate graficamente attraverso delle parentesi uncinate (< >) che esplicitano le informazioni non direttamente riconducibili al testo latino (o greco). Questa pratica è finalizzata a facilitare la comprensione del testo, ma produce nei testi un disallineamento non trascurabile. Per questa ragione è stata condotta un'approfondita analisi quantitativa e qualitativa di tali interventi allo scopo di individuare strategie per il loro trattamento nella fase di allineamento automatico dei testi. Dall'analisi condotta è emerso che le aggiunte più rilevanti dal punto di vista dell'approccio alla traduzione sono certamente quelle informative. È infatti attraverso queste ultime che il traduttore specialista veicola concetti tecnici, spesso connessi al contesto legale romano, che quindi non trovano un diretto corrispondente in italiano.

3. GLI STRUMENTI IMPLEMENTATI

Nella linguistica computazionale, i corpora rivestono un ruolo centrale. In particolare, quelli paralleli sono fondamentali per diverse applicazioni, tra le quali la traduzione automatica, l'estrazione di terminologie e la disambiguazione semantica (Fathi, 2015). Per la costruzione del corpus bilingue lo studio della letteratura in ambito di sincronizzazione dei testi ha mostrato come le più diffuse procedure software di allineamento automatico di testi implementino algoritmi basati su metodi statistici, che tipicamente lavorano a livello di frase o sequenza di caratteri. Quando però il parallelismo stringente tra testo e traduzione si perde a favore di un'estensione/semplificazione dei contenuti, gli stessi algoritmi di allineamento producono un numero maggiore di porzioni non allineate, sulle quali è necessario intervenire con strategie diverse. Tipicamente, un tale approccio produce buoni risultati solo per lingue tipologicamente affini e per testi *well-behaved* (bene educati), ma non è adatto al caso del corpus di dominio descritto. Anche quello che utilizza il metodo linguistico (o *rule-based*) non era utilizzabile in assenza di un lessico bilingue in grado di stabilire corrispondenze traduttive univoche tra parole/lemmi. La situazione descritta ha orientato le scelte verso un metodo 'misto', in cui l'allineamento automatico viene corretto stabilendo delle 'ancore' linguistiche in corrispondenza di una lista di parole/elementi per i quali sono stati identificati degli equivalenti bilingui (Marinai et al. 1992). Per la parallelizzazione del corpus del Digesto sono state prima generate liste allineate di parole target e corrispondenti elementi source, tra i quali è stata individuata e annotata una rete di ancore, in modo da permettere la parallelizzazione dei contesti relativi, visualizzabili a richiesta in modalità contrastiva. Successivamente, la componente aleatoria del legame tra ancore (non si tratta di veri traducenti) ha richiesto un puntuale lavoro di revisione manuale delle corrispondenze individuate automaticamente, e ha rappresentato un test di valutazione delle procedure di estrazione di candidati termini. Il corpus testuale allineato, che in forma di concordanze contrastive è possibile consultare online[2], si compone dei primi 36 libri ed è strutturato in 72 archivi indicizzati (uno per ciascun libro e lingua) contenenti 1.044.055 occorrenze per la lingua italiana e 635.260 occorrenze per la lingua latina, per un totale di quasi 1.700.000 parole. Questa funzionalità, assimilabile alle memorie di traduzione, è stata utilizzata da subito come supporto al lavoro dei traduttori per perseguire l'uniformità dei criteri di traduzione e la coerenza terminologica. Andando oltre lo stretto ambito dei traduttori, sfruttando le corrispondenze tra unità testuali che sono traduzione una dell'altra è possibile infine disporre di risorse bilingue da poter riutilizzare per l'addestramento di sistemi di traduzione automatica di testi di dominio. Nella maggior parte dei casi, la corrispondenza è già possibile a livello di frase: nei primi 36 libri del Digesto le frasi allineate sono 24.213, mentre le unità non allineate a livello di frase risultano 1.358. Sono inoltre in fase di studio strategie per sfruttare le unità già allineate per una sperimentazione sulla traduzione automatica basata su un approccio di Machine Learning (Deep Learning) da applicare ai libri ancora da tradurre.

3.1 Studi preparatori per il glossario

Le problematiche nell'allineamento dei testi hanno influito sulla costruzione di un sistema automatico in grado di estrarre una corrispondenza bilingue tra termini. La costruzione di un lessico terminologico bilingue porta con sé diverse sfide. Mentre per le lingue moderne più diffuse la procedura di estrazione di termini nella lingua source può avvalersi sia di risorse terminologiche già esistenti da modificare/integrare da cui partire, sia di tecniche di NLP (lemmatizzazione e *term extraction*), la presenza di una lingua storica riduce la disponibilità di tali risorse e strumenti. Inoltre, per quanto riguarda l'individuazione dei candidati traducenti, l'operazione può risultare complessa nei casi in cui un termine non trovi un corrispettivo nella lingua di arrivo e sia stata utilizzata per esso un'espressione polirematica e/o anche più estesa (si vedano gli esempi in figura 1). Sulla base dell'analisi delle traduzioni, abbiamo rilevato 4 livelli di complessità nella riconduzione dei termini dalla lingua source alla lingua target.

[2] http://dbtvm1.ilc.cnr.it/digesto

Il primo livello riguarda i casi più semplici di polisemia. Si veda l'esempio 1: recuperando i contesti del lemma *insula* emergono dai testi due traduzioni diverse: "isola" (1a) e "casamento" (1b), con diversi ambiti di applicazione.

1) ***Insula***
   a) (LAT) *ULPIANUS libro nono ad edictum. Insulae Italiae pars Italiae sunt et cuiusque provinciae.* (Dig. 5, I, 9)
   (ITA) ULPIANO, nel libro nono All'editto. Le isole dell'Italia sono parte dell'Italia, come quelle di ogni provincia <lo sono della stessa>.
   b) (LAT) *Idem scribit, si aream emero et* ***insulam*** *in ea aedificavero, recte me Publiciana usurum.* (Dig. 6, II, 11, 9)
   (ITA) Lo stesso <Pomponio> scrive che, se avrò comprato l'area, ed avrò edificato su di essa un **casamento**, correttamente mi servirò della Publiciana.

I casi di polisemia includono anche quei lemmi che utilizzati in un contesto non specialistico assumono generalmente uno o più sensi comuni, ma che quando compaiono nel discorso legale specializzano il proprio significato per indicare un concetto distintivo di questo ambito. Si veda a questo proposito l'esempio 2, che illustra il caso del verbo latino *caveo*. Nel discorso non specialistico, questo lemma si può tradurre con "stare in guardia, guardarsi da (rispetto/contro qualcosa o qualcuno)", "stare all'erta", oppure "evitare" (2a); tuttavia, nel contesto giuridico e in particolare nel Digesto, questo termine viene utilizzato con il senso tecnico di "decretare, disporre, stabilire", nel caso di leggi ed editti (2b), oppure "prestare (stipulazione di) garanzia" (2c) (OLD, s.v. "caveo").

2) ***Caveo***
   a) (LAT) *Si fullo vestimenta polienda acceperit eaque mures roserint, ex locato tenetur, quia debuit ab hac re* ***cavere****.* (Dig. 19, II, 13, 6)
   (ITA) Se un tintore abbia ricevuto dei vestiti per pulirli ed i topi li abbiano rosicchiati, è tenuto <con l'azione> da locazione, poiché avrebbe dovuto **cautelarsi** riguardo a ciò.
   b) (LAT) *Oratione divi Marci* ***cavetur****, ut, si senatoris filia libertino nupsisset, nec nuptiae essent:* (Dig. 23, II, 16)
   (ITA) Nell'orazione del divo Marco <Aurelio letta in senato> **si dispone** che, se la figlia di un senatore avesse sposato un liberto, <queste> non sarebbero state nozze:
   c) (LAT) *Tutor et curator, ut rem salvam fore pupillo* ***caveant****, mittendi sunt in municipium, quia necessaria est satisdatio.* (Dig. 2, VIII, 8, 4)
   (ITA) Il tutore e il curatore, affinché **prestino garanzia** al pupillo a salvaguardia dell'integrità del suo patrimonio, devono essere fatti andare al municipio, perché è una dazione di garanti necessaria.

Il caso di *caveo* "prestare (stipulazione di) garanzia" (2c) è utile anche per illustrare un'ulteriore complicazione nella compilazione del glossario, che è rappresentata dalla presenza di espressioni polirematiche, sia in latino che in italiano. Alcune di queste infatti, specialmente in un contesto specialistico, assumono grande rilevanza e diffusione. Nell'esempio 2c un unico termine nella lingua source viene tradotto con una polirematica nella lingua target. Esistono anche casi nei quali la polirematica si trova nella lingua source, e viene tradotta con un unico termine nella lingua target. Questo è il caso di *aes alienum* (lingua source) che viene tradotto come "debito" (lingua target), illustrato nell'esempio 3.

3) ***Aes alienum***
   (LAT) *Si* ***aere alieno*** *dominico exhauriatur peculium servi, res tamen in causa peculiaria manent:* (Dig. 15, I, 4, 5)
   (ITA) Se il peculio del servo si esaurisca per **debiti** nei confronti del padrone, le cose, tuttavia, restano ugualmente nel peculio:

Infine, sono stati individuati casi di polirematiche a incastro o concatenate, illustrate nell'esempio 4.

4) ***Ab intestato***
   a) (LAT) *Haec oratio pertinet ad alimenta, quae testamento vel codicillis fuerint relicta sive ad testamentum factis sive* ***ab intestato****.* (Dig. 2, XV, 8, 2)
   (ITA) Questa orazione riguarda gli alimenti che siano stati lasciati nel testamento o nei codicilli, sia testamentari sia **in assenza di testamento**.

b) (LAT) *Si quis ex his personis, quae ad successionem* ***ab intestato*** *non admittuntur, de inofficioso egerit (nemo enim eum repellit) et casu optinuerit, non ei prosit victoria, sed his qui habent ab intestato successionem:* (Dig. 5, II, 6, 1)
(ITA) Se una delle persone, che non sono ammesse alla successione **intestata**, abbia agito per testamento inofficioso (infatti, nessuno la respinge), e per un caso abbia vinto, questa non giova a lei, ma a coloro che hanno titolo alla successione intestata:

La collocazione *ab intestato* significa letteralmente "da (parte di) qualcuno che non ha fatto testamento". Questa viene tradotta solitamente nel testo italiano attraverso la poliremática "in assenza di testamento" (4a), o addirittura si conserva la locuzione latina *ab intestato*, tuttora utilizzata nel linguaggio giuridico moderno. Questa espressione compare in casi di successione nei quali il defunto non ha potuto, per qualche motivo, lasciare testamento. Quando si accompagna a sostantivi che si riferiscono alla cessione dei beni, come *successio* "successione" o *hereditas* "eredità", *ab intestato* forma, con tali sostantivi, un'ulteriore espressione specialistica del dominio giuridico, che viene tradotta con "successione/eredità legittima" o "intestata", in quanto in assenza di testamento si applica la legge che regola le successioni.
In questo scenario l'intervento della correzione/revisione manuale del lessico/glossario è imprescindibile. Ma è proprio qui che la linguistica computazionale può offrire un supporto fattivo e importante, non solo per cercare strategie in grado di ridurre l'intervento umano allo stretto necessario, ma anche per verificare le ipotesi con campagne di estrazione mirate e puntuali sui casi più ambigui. Sulla base delle analisi condotte, abbiamo optato per un processo incrementale, in cui si parte dalla lemmatizzazione del testo latino e dalla selezione di termini validati e/o da validare da parte degli esperti, per poi definire un primo insieme dei termini source. Per quelli target, date le problematiche descritte e l'esperienza maturata in fase di allineamento, per il collegamento delle 'ancore' linguistiche si sta studiando un modo di individuare il 'senso' di una traduzione identificandolo tra un nucleo di parole ricorrenti. Questo procedimento si basa sull'ipotesi distributiva di Harris (1968), che afferma che parole semanticamente simili tendono a comparire in contesti linguistici simili. Si tratterebbe quindi di estendere la ricerca del termine traducente a un insieme di più parole che hanno legami semantici stabili.

4. PROSPETTIVE FUTURE

Il Progetto è tuttora in corso, così come il processo di traduzione. Nell'ambito delle attività svolte si ritengono rilevanti sia il protocollo di trattamento, allineamento, standardizzazione del formato di rappresentazione del corpus bilingue, sia l'estrazione di risorse linguistiche di dominio. Ogni aspetto di tali attività ha richiesto uno studio specifico e un continuo lavoro di riallineamento/adattamento degli strumenti, in particolare per l'allineamento dei testi e per l'estrazione di terminologia bilingue.
Quest'ultimo aspetto resta ancora una sfida aperta perché la traduzione non è conclusa e i libri che sono ancora in traduzione potrebbero introdurre cambiamenti nella terminologia, vista la strutturazione ad argomento dell'opera.

4.1 Un confronto tra modelli neurali generativi e traduzione umana

La traduzione automatica di norma è poco applicata nei contesti di traduzione di dominio. È tuttavia intento del progetto esplorare ambiti non usuali. Con questa intenzione si è deciso di provare comunque un approccio automatico, ora che i modelli di Machine Learning hanno preso il posto di quelli statistici. In tempi recentissimi, lo sviluppo del settore dell'Intelligenza Artificiale (IA) ha portato all'introduzione di modelli estremamente avanzati che possono essere applicati alla traduzione automatica. La svolta in questo senso è rappresentata dalla Traduzione Automatica Neurale (NMT, Neural Machine Translation), che consiste nell'utilizzo di reti neurali artificiali: sottoposte a grandi quantità di dati linguistici in fase di training, queste rielaborano l'informazione appresa in maniera complessa attraverso i vari strati della rete. Nel campo della NMT sono attualmente estremamente popolari i Large Language Models (LLM), tra i quali uno degli esponenti più in vista è il modello GPT, sviluppato dall'azienda OpenAI (2023), i quali generano automaticamente una risposta a partire da un prompt impartito dall'utente (Angelucci 2024:187–190). Il motivo del successo di questi modelli risiede nell'abbandono della statica e limitativa rappresentazione vettoriale delle parole a favore di quella contestuale. Le innovazioni prodotte dalle ultime ricerche hanno cambiato le tecniche disponibili, dotandole della capacità di valutare il contesto della parola, ovvero utilizzando informazioni legate alle parole adiacenti ad essa. Al fine di confutare le affermazioni degli esperti della traduzione, che restano dubbiosi sull'utilizzo di questi strumenti, sono stati condotti alcuni esperimenti pilota su GPT-4, confrontando l'output del modello con la traduzione realizzata dagli specialisti su diversi passaggi del Digesto. Per effettuare gli esperimenti è stato utilizzato il modello GPT-4, in accesso gratuito, al quale è stato sottoposto il seguente prompt: "Traduci il seguente passaggio dal latino all'italiano: [PASSAGGIO IN LATINO]". Al modello sono state sottoposte frasi dalla lunghezza e complessità

variabili. GPT raggiunge già degli ottimi risultati nei compiti di traduzione, anche quando la lingua di partenza è il latino (Volk et al., 2024). Tuttavia, il nostro obiettivo era testare le performance del modello sulla traduzione di un testo specialistico, di materia giuridica e di età antica. I primi risultati hanno evidenziato come anche i modelli più recenti trovino difficoltà in tutti quei casi di utilizzo delle aggiunte che abbiamo ampiamente descritto in precedenza (sezione 2.1). A conferma di questo riportiamo qui di seguito alcuni passaggi del Digesto (esempi in a); le traduzioni degli specialisti (esempi in b); e le traduzioni di GPT-4 (esempi in c).

5) ***Manus***
    a) (LAT) *Manumissiones quoque iuris gentium sunt. est autem manumissio de* ***manu*** *missio, id est datio libertatis:* (Dig. 1, I, 4)
    b) (ITA: Dig.) Anche le manomissioni appartengono al diritto delle genti. La mano-missione è la dismissione dalla **<soggezione a quella potestà chiamata> "mano"**: si tratta, cioè, della concessione della libertà.
    c) (ITA: chatGPT) Anche le manomissioni appartengono al diritto delle genti. Ora, la manomissione è la liberazione dalla **mano**, cioè la concessione della libertà.
6) ***Status/libertinus***
    a) (LAT) *Homo liber, qui se vendidit, manumissus non ad suum statum revertitur, quo se abdicavit, sed efficitur libertinae condicionis.* (Dig. 1, V, 21)
    b) (ITA: Dig.) L'uomo libero, che si vendette, se viene manomesso non ritorna al proprio **stato <originario>**, dal quale ha abdicato, ma assume la **condizione di liberto**.
    c) (ITA: chatGPT) Un uomo libero, che si è venduto come schiavo, una volta manomesso, non ritorna al suo **stato originario**, che ha rinunciato, ma diventa di **condizione libertina**.

Gli esempi 1 e 2 offrono una panoramica di pregi e difetti del modello GPT, se confrontato con la traduzione realizzata dagli specialisti. In particolare, l'esempio 2 permette di confrontare la traduzione di *status* e *libertinus*. Il primo termine si riferisce allo stato giuridico del soggetto, un concetto che esiste anche nella lingua italiana, e non ha quindi bisogno di essere ulteriormente specificato attraverso una perifrasi. È interessante notare anche che il traduttore opta per un'aggiunta che specifica lo stato giuridico al quale ritorna il soggetto (<originario>, esempio 2b); d'altra parte GPT-4 inserisce lo stesso tipo di informazione anche se questa non è direttamente reperibile nel testo, probabilmente su indicazione del verbo *revertitur* da *revertor* "ritornare", che suggerisce un'idea di ripristino di una situazione precedente. Diverso è il caso di *libertinus*, aggettivo che si riferisce alla condizione di liberto. "Condizione di liberto" è infatti la formula scelta dai traduttori per questo aggettivo nell'esempio 2b. Diversamente, il modello GPT-4 opta per "libertino", che però risulta essere una scelta errata, in quanto in italiano questo aggettivo ha subito un mutamento semantico, e fa oggi riferimento ad atteggiamenti licenziosi e a una condotta disordinata.

RINGRAZIAMENTI

Progetto di ricerca PRIN 2022 PNRR P20224NJLK – SH2 – TESTO "Translating, Encoding, Sharing The Origins. From the Littera Florentina to an open-access Italian translation of Justinian's Digest", nell'ambito del Piano Nazionale Ripresa e Resilienza (PNRR) M4C2 – Investimento 1.1 "Progetti di Ricerca di Rilevante Interesse Nazionale (PRIN)- finanziato dall'Unione europea – NextGeneration EU" – CUP B53D23032480001.

BIBLIOGRAFIA

Addis, F. (2023). Sistema, codice e formazione nel canone di Sandro Schipani. Roma e America: diritto romano comune: rivista di diritto dell'integrazione e unificazione del diritto in Europa e in America Latina: 44, 2023, 239–254.

Angelucci, M. (2024). Computabilità e traduzione interlinguistica: considerazioni sull'impiego dell'Intelligenza Artificiale in ambito traduttivo. DigItalia, 19(1), 183–197.

Banchich, T. M., Marenbon, J. & Reid, C. J. (2015). The Revival of Roman Law and Canon Law. In Miller Jr, F. D., & Biondi, C. A. (eds.) A Treatise of Legal Philosophy and General Jurisprudence: Volume 6: A History of the Philosophy of Law from the Ancient Greeks to the Scholastics. Dordrecht: Springer. 251–265.

Dingledy, F. W. (2016). The Corpus Juris Civilis: A Guide to its History and Use. Legal Reference Services Quarterly. DOI: 10.1080/0270319X.2016.1239484.

Fawi, Fathi. (2015). Costituzione di un corpus giuridico parallelo italiano-arabo. In Proceedings of the Second Italian Conference on Computational Linguistics CLiC-it 2015, 125–129.

Harris Z. S. (1968) Mathematical Structures of Language, New York, Wiley.

Marinai E., Peters C., Picchi E. (1992). A Project for Bilingual Reference Corpora. Acta Linguistica Hungarica. vol. 41. n.1/4. 191–204. http://www.jstor.org/stable/44308294.

OLD = Glare, P. G. W. (ed.). 1997[1982]. Oxford Latin dictionary. Oxford: Oxford University Press.

OpenAI. 2023. GPT-4 technical report.

Paolucci, S. (2013). Strategia estraniante e strategia addomesticante nella traduzione dei testi giuridici. Linguistica 53(2), 73–89.

Paolucci, S. (2017). Foreignising and domesticating strategies in translating legal texts. International Journal of legal discourse, 2(2), 243-263.

Pezzato E. (2022). Brevi note sulle inscriptiones del Digesto. CULTURA GIURIDICA E DIRITTO VIVENTE, 10, pp. 1-14. https://dx.doi.org/10.14276/2384-8901/3403.

Ribary, M., & McGillivray, B. (2020). A Corpus Approach to Roman Law Based on Justinian's Digest. Informatics, 7(4), 44. https://doi.org/10.3390/informatics7040044.

Volk, M., Fischer, D. P., Fischer, L., Scheurer, P., Ströbel, P., Sprugnoli, R., & Passarotti, M. (2024). LLM-based machine translation and summarization for Latin. Third Workshop on Language Technologies for Historical and Ancient Languages -- LT4HALA (at LREC/COLING), Torino, 25 May 2024.

Preliminary Results for the Explanation of Neural Network-based Handwriting Identification in Historical Manuscripts

Riccardo De Cesaris[1], Valerio Caravani[2], Arianna Pastorini[3], Serena Ammirati[4], Paolo Merialdo[5]
[1] Università degli Studi "Roma Tre", Italy – riccardo.decesaris@uniroma3.it
[2] Università degli Studi "Roma Tre", Italy – valerio.caravani@uniroma3.it
[3] Università di Bologna, Italy – arianna.pastorini2@unibo.it
[4] Università degli Studi "Roma Tre", Italy – serena.ammirati@uniroma3.it
[5] Università degli Studi "Roma Tre", Italy – paolo.merialdo@uniroma3.it

ABSTRACT (ENGLISH)
This work tackles the "black-box" problem in Digital Paleography—specifically in Handwriting Identification—by introducing a tailored explainability approach designed to enhance the interpretability of AI-based systems in manuscript analysis. As a novel contribution, we apply Explainable Artificial Intelligence (XAI) techniques to the case study of "Vat.lat.653", an 11th-century manuscript preserved in the Vatican Apostolic Library. Leveraging a Convolutional Neural Network (CNN)-based classifier, we adapt LIME, a well-established explanation method, to identify the most influential features on each page with respect to the model's decisions. The present study represents one of the first attempts to systematically integrate XAI methods into paleographic analysis of historical handwritten texts, providing tools and methodology to assess the reliability and consistency of automated systems, helping promote their use in this domain. Although our findings provide insights into the potential of XAI methods for paleographic analysis, they remain preliminary and require further validation in order to fully assess the effectiveness and generalizability of the proposed approach.

Keywords: Explainable AI; Handwriting Identification; Digital Paleography

ABSTRACT (ITALIANO)
Risultati Preliminari per l'interpretazione dell'Identificazione della Grafia effettuata con Reti Neurali su Manoscritti Storici.

Il presente lavoro affronta il problema "black-box" nella Paleografia Digitale—in particolare nell'Identificazione della Grafia—introducendo un approccio progettato su misura per migliorare l'interpretabilità dei sistemi basati su IA nell'analisi dei manoscritti. Come contributo originale, applichiamo tecniche di Explainable Artificial Intelligence (XAI) al caso di studio del manoscritto "Vat.lat.653", un documento dell'XI secolo conservato nella Biblioteca Apostolica Vaticana. Utilizzando un classificatore basato su Reti Neurali Convoluzionali (CNN), adattiamo LIME, un metodo consolidato per la generazione di spiegazioni, al fine di identificare le caratteristiche più influenti in ogni pagina rispetto alle decisioni del modello. Questo studio rappresenta uno dei primi tentativi di integrare sistematicamente metodi XAI nell'analisi paleografica di manoscritti storici, fornendo strumenti e metodologie per valutare l'affidabilità e la coerenza dei sistemi automatici, promuovendone l'adozione in questo ambito. Sebbene i risultati offrano spunti sul potenziale delle tecniche XAI applicate all'analisi paleografica, rimangono preliminari e richiedono ulteriori validazioni per una valutazione completa dell'efficacia e della generalizzabilità dell'approccio proposto.

Parole Chiave: Intelligenza Artificiale Interpretabile; Identificazione della Grafia; Paleografia Digitale

1. INTRODUCTION
In recent years, the exploitation of Deep Learning techniques for the automated analysis of both historical and modern manuscripts has gathered increasing interest, offering the potential of both simplifying and significantly improving the work of paleographers and domain experts (Hu, Wang, Li, & Wang, 2021; Nieddu, Firmani, Merialdo, & Maiorino, 2021). Among the various tasks that can be addressed using such approaches there is "Handwriting Identification", which consists in partitioning texts into parts which can be attributed to different scribes, each one characterized by his own writing style.

However, the adoption of Deep Learning models raises the critical challenge known as the "black-box problem" (Stokes, 2015), one of the most debated issues in the deployment of AI systems in real-world applications. Briefly, this problem concerns the lack of transparency in model decision-making, where

outcomes are typically expressed only through performance metrics, limiting their interpretability and usefulness for non-technical users. Addressing this limitation requires developing explanation strategies capable of making model predictions both understandable and meaningful to experts in the target domain.

In this work, we present an original contribution to the field of Digital Paleography by integrating Explainable Artificial Intelligence (XAI) techniques into the automated analysis of historical manuscripts. Among the numerous approaches (Pradhan, Lahiri, Galhotra, & Salimi, 2022) for interpreting learning-based models, our methodology leverages LIME (Ribeiro, Singh, & Guestrin, 2016), a well-known "feature-based" explainability technique. LIME's straightforward strategy succeeds in explaining a prediction provided by the original classifier leveraging an inherently interpretable model (with a known decision function) whose purpose is to approximate the behavior of the original classifier within a local region around the instance to be explained itself. For this reason, LIME can be defined as "model-agnostic", meaning it can be applied to any type of classifier, regardless of its complexity or architectural details.

Through the adaptation of LIME to the Handwriting Identification problem, we aim to identify the portions of the manuscript that most influence a CNN-based classifier (Lastilla, Ammirati, Firmani, Komodakis, Merialdo, & Scardapane, 2022) trained to attribute pages to individual scribes: in doing so, we are able to provide concrete and visual explanations for model's decisions. While the proposed approach shows promising results—with the generated explanations effectively reflecting the model's decision-making process for the task of interest—its broader effectiveness and generalizability need further validation across additional experiments and case studies.

2. CASE STUDY

Our case study focuses on "Vat.lat.653"[1], an 11th-century manuscript preserved in the Vatican Apostolic Library. The manuscript comprises 538 pages (or 269 folios) and was selected from the collection of manuscripts employed in (Lastilla, Ammirati, Firmani, Komodakis, Merialdo, & Scardapane, 2022), primarily due to the high classification performance achieved on its scribes, which forms a strong basis for a reliable explainability analysis. "Vat.lat.653" contains the *Expositio in Epistulas Pauli* by Aimon of Auxerre and was copied in the so-called "minuscola romanesca", a variant of Carolingian minuscule typical of Rome and its surroundings from the 10th to the 12th century; it was produced in the Monastery of St. Scholastic in Subiaco. In 522 out of the 538 pages, paleographers identified four distinct scribal hands contributing to the manuscript's composition. Assuming each page was written by a single scribe, the page distribution is reported below:

- **Scribe 0**, 42 pages: from folio 1 *recto* (r) to folio 21 *verso* (v);
- **Scribe 1**, 58 pages: from folio 237 *recto* (r) to folio 265 *verso* (v);
- **Scribe 2**, 202 pages: from folio 22 *recto* (r) to folio 121 *verso* (v);
- **Scribe 3**, 220 pages: from folio 122 *recto* (r) to folio 231 *verso* (v);

Each scanned page measures 902x1279 pixels, with the text organized into two columns per page, each containing approximately thirty lines.

3. METHODOLOGY

The original classifier (Lastilla, Ammirati, Firmani, Komodakis, Merialdo, & Scardapane, 2022) involved in our experiments consists of a ResNet18 (He, Zhang, Ren, & Sun, 2016) backbone encoder followed by two fully-connected linear layers. The model has been fine-tuned[2] on 380x380-pixel crops extracted from the "Vat.lat.653" pages with the objective of performing a multi-label classification task. The system achieved an overall accuracy greater than 86% on the Test Set, providing a sufficiently reliable foundation for the subsequent explainability analysis.

As outlined in the previous sections, the aim of the proposed explainability strategy is to evaluate the relevance of page sub-regions that the model relies on at inference time. In LIME terminology, these sub-

[1] Available Online at: http://www.mss.vatlib.it/guii/console?service=present&term=@5Vat.lat.653_ms&item=1&add=0&search=1&filter=&relation=3&operator=&attribute=3040

[2] The initial weights of the backbone encoder are the result of the experiments conducted in (Lastilla, Ammirati, Firmani, Komodakis, Merialdo, & Scardapane, 2022)

regions are referred to as "super-pixels": in our implementation, each super-pixel measures 50x50 pixels and enables the analysis of feature associated with short text sequences (2-3 consecutive characters) as well as features spanning across two consecutive lines of text. Given an instance to be explained, LIME approximates the behavior of the original classifier within a local region around the instance by training an inherently interpretable model.

To achieve this, LIME generates a set of random samples (referred to as "neighboring points") by applying random modifications to the features of the original input instance. These perturbed samples are fed into the original classifier to obtain their corresponding outputs. Finally, using the generated neighboring points and their associated responses from the original black-box model, LIME trains an interpretable model that serves as a local approximation of the original one. In our context, LIME locally approximates the original classifier's behavior for 380x380-pixel crops extracted from the instance to be explained, assigning an importance score to each super-pixel. The scores are normalized within the range [-1, 1]: for each super-pixel, a positive value indicates that the corresponding super-pixel contributes positively to the correct classification, while negative values suggest that its information misleads the model's prediction.
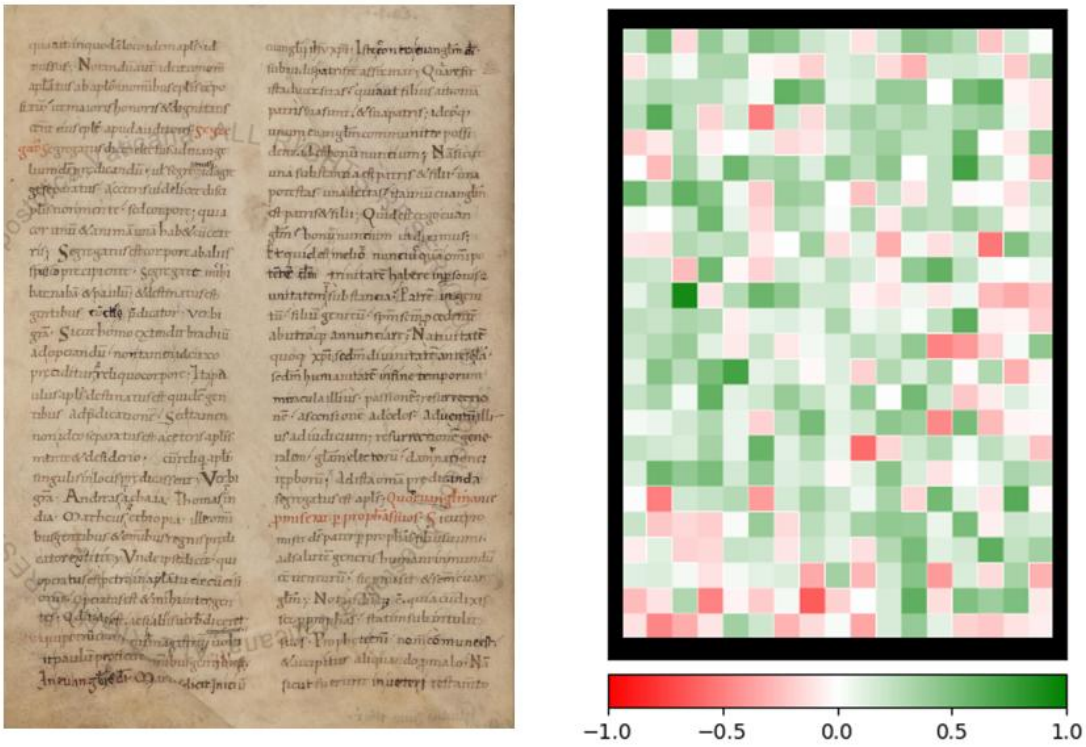


Figure 1. (i) Folio 1 *verso* – (ii) Its attribution Map

To identify the most important super-pixel at the level of a single page, we collected a set of overlapping 380x380 crops covering the entire page using a sliding window mechanism.
For the instance to be explained, its "attribution map" (Figure 1) is constructed by aggregating the attribution scores of its "super-pixels": specifically, for each "super-pixel" present on the page, its attribution score is calculated as the average of the scores obtained from all crops that include the corresponding sub-region. To reduce possible fluctuations (Zhang, Song, Sun, Tan, & Udell, 2019; Tan, Tian, & Li, 2024) in the importance score values caused by the random perturbation process, the "attribution map" is computed by averaging the results over 3 iterations.

In the attribution maps, green regions correspond to super-pixels deemed relevant by LIME, red regions indicate misleading super-pixels whose content negatively affects the prediction, and white areas represent information that appears to have no significant impact on the model's decision.

This approach enables domain experts to visually identify the sequences considered most influential by the classifier, assess their paleographic relevance, investigate potential parallels between automated and human methodologies, and refine the process of identifying different writing styles within a document.

Furthermore, the identified features can be aggregated across multiple pages to evaluate the presence of character sequences that consistently influence the prediction of a specific scribe.

4. EXPERIMENTAL RESULTS & VALIDATION

The proposed methodology was validated through two orthogonal procedures:

- **Training Set Masking & Re-Training**: starting from the explanations computed for the Training Set instances, the "super-pixels" estimated as the most relevant were removed. The system was then re-trained using the "masked" version of the Training Set instances, and the system's accuracy on the Test Set was finally re-computed.
- **Test Set Masking & Accuracy Re-Computation**: starting from the explanations computed for the Test Set instances, the "super-pixels" estimated as the most relevant were removed. The system accuracy on the Test Set was then re-computed.

Training Set Masking & Re-Training: for each Training Set instance, we incrementally removed the most relevant crops, masking the top 5%, 10% and 20% of the total area of a given instance.
After performing this "masking process", we re-trained the original classifier and subsequently measured again the overall accuracy on the Test Set. The results obtained (Table 1) are reported below:

Masked Area	Overall Accuracy on Test Set
0% (baseline)	86,36%
5%	78,41%
10%	73,86%
20%	64,77%

Table 1. Accuracy Measures for baseline and re-trainings

Test Set Masking & Accuracy Re-Computation: for each Test Set instance, we incrementally removed the most relevant crops (Atanasova, Simonsen, Lioma, & Augenstein, 2020), masking from the top 1% to the top 20% (with a step size of 1%) of the total area of a given instance. After performing this "masking process", we provided the modified Test Set as input to the original classifier and subsequently re-measured the overall accuracy on these instances. Figure 2 highlights how the system's performance (Test Accuracy) degrades as the proportion of masked relevant information increases.

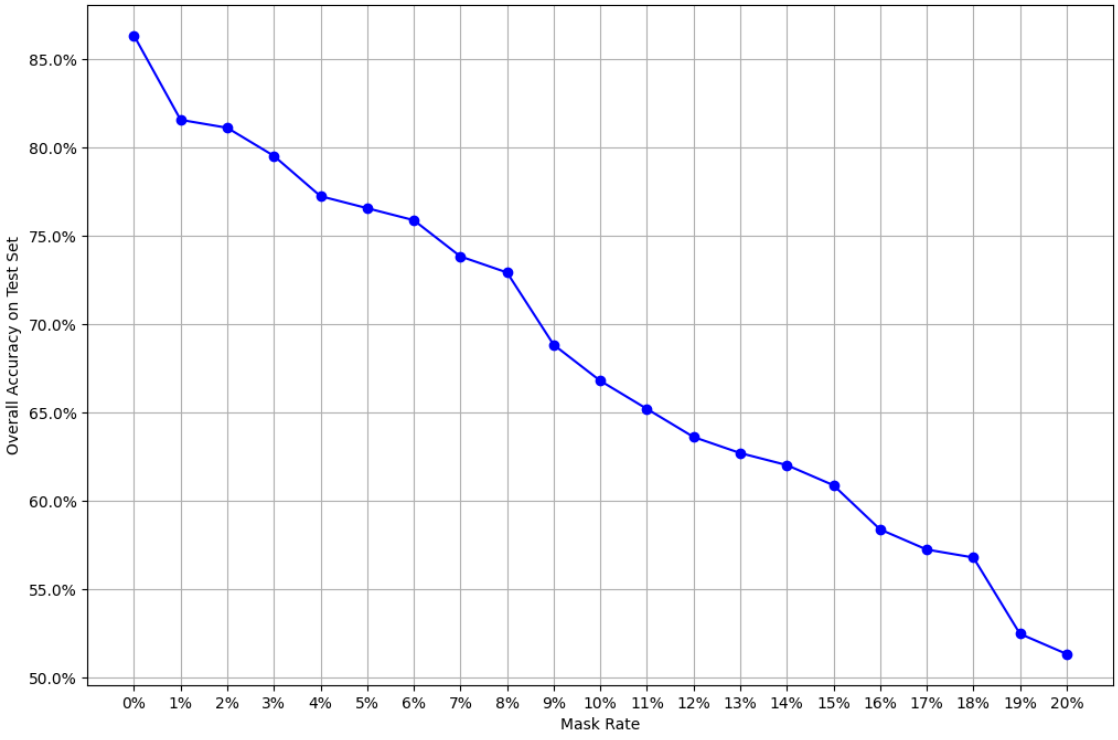


Figure 2. Relationship between Mask Rates and Test Accuracies (in percentage)

The experiments conducted and the corresponding outcomes highlight the effectiveness of the proposed explainability approach in identifying the features critical to the model decision-making process. More precisely, the observed performance degradation upon removing the information deemed relevant by LIME further confirms that the system genuinely relies on these features to make its predictions.

5. CONCLUSIONS AND FUTURE WORK

The approach proposed in this work tackles the "black-box" problem within the field of Digital Paleography, specifically concerning the Handwriting Identification task. Our technique enables the identification of a limited set of relevant features for each page under study, providing new insights into the interpretability of AI-based systems in manuscript analysis. By applying this method across multiple pages, it is also possible to uncover additional distinctive features that may have a greater impact on the classifier's prediction for a given scribe.

The results obtained show promising potential, yet they remain preliminary. As part of future work, further validation is needed to confirm the effectiveness and reliability of the proposed approach. The technique will be extended and tested on different models and datasets to evaluate its generalizability. Additionally, we aim to leverage the insights gained from the explainability process itself to improve the classifier's performance and refine its predictions in subsequent iterations.

REFERENCES

Atanasova, P., Simonsen, J.G., Lioma, C., Augenstein, I. (2020). A Diagnostic Study of Explainability Techniques for Text Classification. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). https://doi.org/10.18653/v1/2020.emnlp-main.263

He, K., Zhang, X., Ren, S., Sun, J. (2016). Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/CVPR.2016.90

Hu, P., Wang, W., Li, Q., & Wang, T. (2021). Touching text line segmentation combined local baseline and connected component for Uchen Tibetan historical documents. Information Processing & Management, Volume 58, Issue 6. https://doi.org/10.1016/j.ipm.2021.102689

Lastilla, L., Ammirati, S., Firmani, D., Komodakis, N., Merialdo, P., Scardapane, S. (2022). Self-supervised learning for medieval handwriting identification: A case study from the Vatican Apostolic Library. Information Processing & Management, Volume 59, Issue 3. https://doi.org/10.1016/j.ipm.2022.102875

Nieddu, E., Firmani, D., Merialdo, P., Maiorino, M. (2021). In Codice Ratio: A crowd-enabled solution for low resource machine transcription of the Vatican Registers. Information Processing & Management, Volume 58, Issue 5. https://doi.org/10.1016/j.ipm.2021.102606

Pradhan, R., Lahiri, A., Galhotra, S., Salimi, B. (2022). Explainable AI: Foundations, Applications, Opportunities for Data Management Research. In Proceedings of the 2022 International Conference on Management of Data (SIGMOD '22). https://doi.org/10.1145/3514221.3522564

Ribeiro, M.T., Singh, S., Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). https://doi.org/10.1145/2939672.2939778

Stokes, P.A. (2015). Digital Approaches to Paleography and Book History: Some Challenges, Present and Future. Frontiers in Digital Humanities 2, 5. https://doi.org/10.3389/fdigh.2015.00005

Tan, Z., Tian, Y., & Li, J. (2024). GLIME: general, stable and local LIME explanation. Advances in Neural Information Processing Systems, *36*. https://proceedings.neurips.cc/paper_files/paper/2023/hash/71ed042903ed67c7f6355e5dd0539eec-Abstract-Conference.html

Vatican Apostolic Library: Website of the Vatican Apostolic Library. https://www.vaticanlibrary.va/en/

Zhang, Y., Song, K., Sun, Y., Tan, S., & Udell, M. (2019). "Why should you trust my explanation?" Understanding Uncertainty in LIME explanations. arXiv preprint arXiv:1904.12991. https://arxiv.org/abs/1904.12991

From Documents to Data: Digital Technologies in the Study of Notarial Charters - poster proposal -

Franziska Decker[1], Sandy Aoun[1], Giuseppe Consolo[1]
[1] University of Graz, Austria – {firstname}.{lastname}@uni-graz.at

ABSTRACT (ENGLISH)

Our contribution presents an ongoing research effort that investigates late medieval notarial charters using advanced computational methods. Drawing from the Monasterium.net[1] archive, we build a curated corpus of 14th–15th-century documents, primarily from Southern Italy, Austria, and Germany. We aim to identify and analyze notarial practices and documentary typologies through text mining, information extraction, and network analysis. A custom annotation scheme is being developed to deploy machine learning and rule-based models tailored to the historical and linguistic complexity of the sources. By extracting named entities and relationships and visualizing them, we explore the social networks and documentary formulas embedded in the charters. The work contributes to the development of digital diplomatics and supports the evolution of Monasterium.net as a platform for computational cultural heritage research.
Keywords: Digital Diplomatics; Information Extraction; Network Analysis; Late Medieval History

ABSTRACT (ITALIANO)

Dai documenti ai dati: le tecnologie digitali per lo studio delle carte notarili.

Il nostro contributo presenta un progetto in corso che esplora le carte notarili tardo-medievali attraverso l'impiego di metodologie digitali e computazionali avanzate. Il corpus selezionato è tratto dal portale Monasterium.net e comprende documenti del XIV e XV secolo, provenienti in particolare dall'Italia meridionale, dall'Austria e dalla Germania. L'obiettivo è analizzare le pratiche notarili e le tipologie documentarie mediante tecniche di text mining, NER (Named Entity Recognition) e network analysis. Per affrontare la complessità storica e linguistica delle fonti abbiamo sviluppato uno schema di annotazione personalizzato, adatto sia a modelli di apprendimento automatico che a modelli basati su regole. L'estrazione automatica di entità e relazioni, e la loro successiva visualizzazione, consente di ricostruire formule documentarie e reti sociali implicite nei documenti. Il progetto intende contribuire allo sviluppo della diplomatica digitale e alla trasformazione di Monasterium.net in una piattaforma sempre più adatta alla ricerca computazionale nel campo del patrimonio culturale.
Parole chiave: Diplomatica Digitale; Information Extraction; Network Analysis; Basso Medioevo

1. INTRODUCTION

The study of medieval notarial charters has long been a cornerstone for understanding administrative and social practices in pre-modern Europe (Schmoeckel & Schubert, 2009; Ostos Salcedo & Pardo Rodríguez, 1997; Trenchs, 1989; Schuler, 1976; Amelotti & Costamagna, 1975; Cencetti, 1964).
Diplomatics have long identified a well-structured normative and formal system in notarial writing characterized by a rigorous codification of documentary formulas. The evolution of the *instrumentum publicum* attested in the 12th century confirms the evidential value of the notarial document, conferring a central role in the juridical and administrative production to the notary (Zabbia, 2009).
Notarial charters are composed of highly formalized sections and free parts where the notary adapts the form to the needs of the individual act. This textual structure, studied in the work of Sabatini (1965; 1968), allows us to identify phenomena of linguistic micro-variation, which reveal significant differences in documentary practices depending on the geographical and institutional context (Greco, 2017). The analysis of documentary clauses allows identifying distinctive features of documentary typologies and notarial circles, providing valuable data for reconstructing professional networks and interactions between social actors (Capriolo, 2017).
In this context, our work aims to integrate computational methods with the diplomatists traditional approach to analyze late-medieval notarial documentation preserved on Monasterium.net (hereinafter called MOM-CA[2]). Through text mining and network analysis, we intend to identify correlations between

[1] *Monasterium.net,* https://www.monasterium.net (23/01/2025).
[2] The abbreviation stands for "Monasterium - Collaborate Archive".

the individuals involved in the charters' creation, textual formulas, and different documentary typologies, thus contributing to a deeper understanding of notarial practices in the fourteenth and fifteenth centuries. The research aligns with the themes of the AIUCD conference, which emphasizes computational approaches to cultural heritage, including text analysis, network visualization, and the application of cutting-edge digital tools.

2. METHODOLOGY

2.1 CORPUS CREATION

The documents for this study will be sourced from MOM-CA, a collaborative digital archive with over 600,000 entries of European charters, which are mainly but not exclusively from German, Italian, Austrian, Slovakian, Czech Republic, and Hungarian archives. The collection for our research will include documents from the 14th and 15th centuries, with a preliminary focus on German, Austrian, and Southern Italian charters.

Notarial charters suffer from underrepresentation in multiple ways. Firstly, their numbers are much lower in comparison to non-notarial sealed charters, especially in regions above the Alps (e.g. they constitute only 4% of the total archival material scrutinized by Weileder, 2019; see also Leipert et al., 2019). Furthermore, as for the late Middle Ages we are presented with a plethora of documents difficult to handle manually, the scholarly focus has rather been on regal and papal charters, leading to only few transcriptions of notarial charters in contrast to their vast amount. Additionally, if provided, many of the available charter abstracts focus only on the main points of a charter's content but lack information about its scribe, i.e. the notary, where applicable.

To build the corpus of our research study, our data undergoes multiple filtering steps. The first step consists of filtering the entire MOM-CA database by retaining all the charters that have an assigned date definitely pointing to the 14th and 15th century, and by excluding all the entries that have assigned dates certainly belonging to other centuries. Charters which have unclear information or no date value are also excluded and stored for potential manual control. The second step consists of filtering the charters belonging to the 14th and 15th century based on terms and keywords unique to notarial documents which can be found in their respective abstracts or full texts. Charters with existing annotations for the CEI-XML[3] tags `<cei:notariusDesc>`, `<cei:notariusSign>`, or `<cei:notariusSub>` are also preserved at this stage. Implementation-wise, both filtering steps make use of the Python library pandas. Finally, the last step consists of filtering the extracted late medieval charters by the language of their abstract or full text transcription. This is done by applying a language detection tool which has been developed within the authors' project[4] and has been fine-tuned on MOM-CA data.

Since this work is part of a larger project called 'From Digital to Distant Diplomatics' (DiDip), the examined corpus will undergo changes with progress in the research areas of other team members carrying out work for instance on handwritten text recognition, providing more full text charter transcriptions for the corpus. Also, object detection and classification tasks from the field of computer vision using visual clues on charters like seals and notary signs have already yielded promising results (Leipert et al. 2019; Nicolaou et al., 2023) and will enable the identification of further notarial material, especially when only image data exists. Lastly, while progressing in our research, new common textual formulas might become apparent which can broaden the adopted search terms (Brugnoli, 2011).

2.2 CORPUS ANNOTATION

To extract information from the charters, we plan to create ground truth data, which would enable us to build and evaluate computational models. Annotation guidelines employed in the context of extracting named entities from German charters abstracts are quite generic (Aoun et al. 2024), while the BeNASch[5] project, which targets challenges particular to historical German, focuses on complex entity nesting. Both approaches do not fully fit our data and research case study. Hence, we are developing a suitable custom annotation scheme, which defines relevant fine-grained named entity types and pertinent relationships between them. While toolkits such as INCePTION and Prodigy provide interfaces for semantic

[3] *CEI (Charters Encoding Initative),* https://www.cei.lmu.de/.
[4] https://huggingface.co/ERCDiDip/langdetect.
[5] https://dhbern.github.io/BeNASch/.

annotation, an application[6] which caters more to diplomatists needs is also being developed in the DiDip project.

2.3 CORPUS ANALYSIS

We intend to analyze our corpus by tackling the information extraction (IE) challenge of the computational linguistics discipline. Based on our designed annotation scheme, our initial objective would be to tackle the named entity recognition and relation extraction IE tasks. Different methodologies can be tried out in this context, ranging from a rule-based setup to highly intricate yet quite opaque deep learning-based models (Ehrmann et al., 2023).
The very nature of our data, diplomatics research questions, and singular annotation scheme motivates us to test the applicability of a rule based approach against that of a machine learning based one.
Computationally, a rule based approach would rely on the extraction of linguistic features and the usage of regular expressions; while in the machine learning realm, frameworks such as spaCy and Flair, which allow for training custom models, can be exploited.
At a more advanced stage, network analysis and data visualisation can be carried out by leveraging the Python library NetworkX as well as the Gephi open-source software.

2.4 ANTICIPATED CHALLENGES AND SOLUTIONS

Given the work's early stage, several challenges are anticipated. One major challenge for our project is based on the nature of the MOM-CA archive itself. Collecting documentary data from numerous European archives, said data and its quality are highly heterogeneous and barely standardized. Most of its records are accompanied by only minimal data because merely a unique identifier (the charter's archive signature) and an assigned date are sufficient for the upload to MOM-CA. While charter images and abstracts (*regesta*), as well as metadata can frequently be found as additional data, only a minority of the documents come with a transcription or edition of their full text or even annotations.
Even after filtering charters that already have transcriptions and abstracts, the next challenge in such data would be to disambiguate and link the named entities with a low error rate (Jeller, 2016).
The NER task itself is challenging too, as it is performed on historical languages (true even for the majority of the charter abstracts) with no standardized orthography or grammar, and only few to no out-of-the-box models are available for the specific historical languages. Also, many manual annotations will be necessary for potentially training models, which will require the development of custom annotation schemes.
This research's challenges pose a chance to bridge the gap between computational methods and traditional diplomatics. It benefits from the interdisciplinarity of the DiDip project members and presents an exemplary use case for the expected audience of the new generation of the MOM-CA platform, which is currently in development.

3. EXPECTED CONTRIBUTIONS

This work's purpose is to contribute to the relatively new field of digital diplomatics (Vogeler, 2004; Vogeler, 2007; Duranti, 2009; Ambrosio, 2020), as well as the enrichment of MOM-CA as a resource for computational text analysis and broader application in cultural heritage studies. By automatically extracting named entities and relevant relationships between them, and subsequently displaying them through the use of network analysis and data visualisation tools, we aim to understand the communities and contextual tendencies of individuals who chose to consult specific notaries.
Thus, this study serves as a practical use case highlighting the specific needs of diplomatists in regard to the functionalities of the next generation of MOM-CA. This ensures that the platform evolves in response to scholarly requirements.

ACKNOWLEDGEMENTS

The contributing authors are part of the ERC project "From Digital to Distant Diplomatics" (Grant No. 101019327), affiliated with the Department of Digital Humanities at the University of Graz. Tamás Kovács is developing the custom annotation tool to aid in the corpus annotation part of the presented work. He, along with the project leader Georg Vogeler and colleague Johannes Laroche are providing valuable

[6] https://github.com/kreeedit/ANNIE.

feedback on the research concept and annotation scheme. Florian Atzenhofer-Baumgartner has provided help in the corpus creation part.
The authors 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

Ambrosio, A. (2020). La diplomatica e il digitale. Il fondo della Biblioteca della Società Napoletana di Storia Patria. Research Trends in Humanities Education & Philosophy, 1–15.

Amelotti, Mario, Costamagna, Giorgio (1975). Alle Origini Del Notariato Italiano. Studi Storici sul Notariato Italiano, 2.

Aoun, Sandy et al. (2024). Information Extraction from German Medieval Charters Abstracts. 19th Annual International Conference of the Alliance of Digital Humanities Organizations (DH 2024), Washington, D.C., August 6-10, 2024. https://doi.org/10.5281/zenodo.14671230.

Brugnoli, A. (2011). Insediamento, territorio e formule notarili: Una verifica (Verona, IX-XII secolo). Reti Medievali Rivista, 12/1. http://dx.doi.org/10.6092/1593-2214/310.

Capriolo, G. (2017). Pratiche redazionali nel Regno di Napoli in età aragonese: Realtà territoriali a confronto. Scrineum Rivista, 14, 501–518. https://doi.org/10.13128/Scrineum-21996.

Cencetti, G. (1964). Il notaio medievale italiano. Atti del XIII Congresso Nazionale del Notariato, Genova, IX–XXIII. http://www.rmoa.unina.it/id/eprint/5638.

Duranti, L. (2009). From Digital Diplomatics to Digital Records Forensics. Archivaria, 68, 39–66.

Ehrmann, Maud et al. (2023). Named entity recognition and classification in historical documents: A survey. ACM Computing Surveys, 56/2, 1–47. https://doi.org/10.1145/3604931.

Greco, P. (2017). La formula documentaria della defensio nelle carte notarili latine della Langobardia minor (IX secolo): Uno studio linguistico. Philologica Jassyensia, 13/2, 71–88. https://hdl.handle.net/11588/699461.

Jeller, Daniel (2016). Urkunden als Netzwerke. https://doi.org/10.58079/nmde.

Klie, Jan-Christoph et al. (2018). The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation. In Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations, Santa Fe, New Mexico. Association for Computational Linguistics, 5–9.

Leipert, Martin et al. (2020), The Notary in the Haystack – Countering Class Imbalance in Document Processing with CNNs. Proceedings of the 14th IAPR International Workshop on Document Analysis Systems, Wuhan, China, July 26–29, 2020. Lecture Notes in Computer Science, 12116, 246–261. https://doi.org/10.48550/arXiv.2007.07943.

Nicolaou, Anguelos et al. (2023). Efficient Annotation of Medieval Charters. Document Analysis and Recognition. Proceedings of the ICDAR 2023 Workshops. San José, CA, USA, August 24–26, 2023. Vol. 1. Lecture Notes in Computer Science, 14193, 284–295. https://arxiv.org/abs/2306.14071.

Ostos Salcedo, Pilar, Pardo Rodríguez, María Luisa (1997). Estudios sobre el notariado europeo (siglos XIV–XV).

Sabatini, F. (1965). Esigenze di realismo e dislocazione morfologica in testi preromanzi. Rivista di cultura classica e medioevale, 7, 972–998.

Sabatini, F. (1968). Dalla «scripta latina rustica» alle «scriptae romanze». Studi Medievali, Serie III, 9, 320–358.

Schuler, Peter-Johannes (1976). Geschichte des südwestdeutschen Notariats. Von seinen Anfängen bis zur Reichsnotariatsordnung von 1512.

Schmoeckel, Mathias, Schubert, Werner (Eds.) (2009). Handbuch zur Geschichte des Notariats der europäischen Traditionen. Rheinische Schriften zur Rechtsgeschichte, 12.

Trenchs, José (Ed.) (1989). Notariado público y documento privado. De los orígenes al siglo XIV. Actas del VII Congreso Internacional de Diplomática.

Weileder, Magdalena (2019). Spätmittelalterliche Notariatsurkunden. Prokuratorien, beglaubigte Abschriften und Delegationsurkunden aus bayerischen und österreichischen Beständen. Archiv für Diplpmatik. Beihefte, 18.

Vogeler, Georg (2004). Ein Standard für die Digitalisierung mittelalterlicher Urkunden mit XML. Bericht von einem internationalen Workshop in München 5./6. April 2004. Archiv für Diplomatik, 50, 23–34.

Vogeler, Georg (2007). Digital Diplomatics – Digitale Diplomatik: Historical research with medieval charters in a digital world - Die historische Arbeit mit Urkunden in der digitalen Welt. International Conference, Munich, 28.2.-2.3.2007.

Zabbia, Marino (2009). Notai e modelli documentari: Note per la storia della lunga fortuna di una soluzione efficace. Circolazione di uomini e scambi culturali tra città (secoli XII-XIV). A. L. Trombetti Budriesi (Ed.). 23–40. https://hdl.handle.net/2318/86408.

Verso la svolta computazionale della critica dantesca

Fara Autiero[1], Vittorio Celotto[2], Gennaro Ferrante[3], Chiara Fusco[4], Sandra Gorla[5]
Giuseppe Andrea Liberti[6], Mariangela Palomba[7], Serena Picarelli[8], Stefano Angelo Rizzo[9], Silvia Tripodi[10]
[1] Scuola Superiore Meridionale, f.autiero@ssmeridionale.it
[2] Università di Napoli Federico II, vittorio.celotto@unina.it
[3] Università di Napoli Federico II, gennaro.ferrante@unina.it
[4] Scuola Superiore Meridionale, chiara.fusco@unina.it
[5] Università di Napoli Federico II, sandra.gorla@unina.it
[6] Università di Napoli Federico II, giuseppeandrea.liberti@unina.it
[7] Sapienza Università di Roma, mariangela.palomba@uniroma1.it
[8] Università di Napoli Federico II, serena.picarelli@unina.it
[9] Università di Catania, stefano.angelo.rizzo@gmail.com
[10] Università di Napoli Federico II, silvia.tripodi@unina.it

ABSTRACT (ITALIANO)

Gli studi sulla Commedia di Dante affrontano da sempre sfide significative: l'assenza del testo originale, il successo precoce e ampio del poema, la varietà della trasmissione materiale e paleografica e la ricchezza di illustrazioni e commenti richiedono un costante dialogo tra discipline storiche e filologiche. Integrare metodi computazionali è oggi cruciale per risolvere questioni irrisolte e approfondire la comprensione della ricezione e interpretazione del poema. Basandosi sui risultati dell'Edizione Nazionale dei Commenti danteschi e dell'Illuminated Dante Project, il Naples Dante Project adotta un approccio multidisciplinare, concentrandosi su tre ambiti: medium (manoscritti, libri, frammenti), testo (Commedia, commenti) e immagine (miniature, illustrazioni, disegni). Il progetto funge da aggregatore digitale, integrando dataset sui vettori materiali, versioni testuali e caratteristiche iconografiche ed esegetiche in un archivio IIIF potenziato da tecnologie avanzate come reti neurali, imaging multimodale, modellazione semantica dei dati e computer vision. Recenti finanziamenti e collaborazioni rendono oggi possibile una vasta campagna di digitalizzazione, necessaria in virtù dell'ampliamento dei corpora, e l'innovazione dell'architettura dati. Le prime applicazioni di riconoscimento automatico della scrittura su manoscritti danteschi hanno inoltre segnato una svolta nella prassi ecdotica di tipo stemmatico. Queste acquisizioni recenti orientano il progetto verso tre obiettivi principali: 1. ILLUDANT: un catalogo delle illustrazioni della Commedia (manoscritti miniati, libri a stampa, disegni dal XIV al XVI secolo). 2. ECOMMEDIA: recensio integrale della tradizione manoscritta assistita dall'Intelligenza Artificiale per una nuova edizione critica che combina metodo stemmatico e approccio computazionale. 3. COMMENTI DANTESCHI ANTICHI: database interoperabile di commenti in volgare e latino sulla Commedia fino al XV secolo.

Parole chiave: Semantic Web; Multi-Modal Imaging; HTR; Handwriting Recognition; NLP

ABSTRACT (ENGLISH)

Toward the computational turn of Dante studies for AIUCD2025

Research on Dante's Commedia has always faced significant challenges. The absence of the original text, the poem's early and widespread success, the variety of its material and palaeographic transmission, and the wealth of illustrations and commentaries has constantly required a dialogue between historical and philological disciplines. Today, integrating computational methods into this dialogue is crucial for resolving long-standing questions and deepening understanding of the poem's reception and interpretation. Building on the achievements of the Edizione Nazionale dei Commenti danteschi and the Illuminated Dante Project, Naples Dante Project adopts a multidisciplinary approach, focusing on three key areas: medium (manuscripts, books, fragments), text (Commedia, commentaries), and image (illuminations, book illustrations, drawings). It serves as a digital aggregator, integrating datasets on material vectors, text versions, and iconographic and exegetical features into a IIIF-compliant image archive enhanced by advanced technologies such as Deep Neural Networks, Multi-modal Imaging, Semantic Data Modeling, and Computer Vision. Recent funding and collaborations have expanded digitisation efforts, enriched corpora, and restructured data architecture. Preliminary Handwritten Text Recognition on Dante manuscripts has led to a breakthrough in editorial practices based on the stemmatic method. These advancements support

three main objectives: 1. Illustrated Dante: A catalogue of Commedia illustrations, encompassing illuminated manuscripts, printed books, and in-series drawings from the 14th to 16th century; 2. eCommedia: An AI-assisted recensio of the Commedia's manuscript tradition, enabling a new critical edition that merges computational and traditional methods; 3. Commenti Danteschi Antichi: A database of Commedia vernacular and Latin commentaries produced by the end of 15th century, and grounded in reliable philological editions.

Keywords: Semantic Web; Multi-Modal Imaging; HTR; Handwriting Recognition; NLP

1. INTRODUZIONE

La Commedia di Dante godette di grande popolarità subito dopo la morte del poeta nel 1321. Copiata inizialmente a Ravenna e probabilmente trascritta da amanuensi bolognesi, si diffuse rapidamente in Toscana e nel resto dell'Italia centrale e settentrionale. La tradizione manoscritta diretta del poema, risalente al XIV e XV secolo, comprende oggi 580 manoscritti con almeno una cantica e 131 frammenti. Stimando un tasso di decimazione dell'85% per i manoscritti medievali, si calcola una produzione iniziale di circa 5.000 esemplari nei primi due secoli, rendendolo il testo volgare più diffuso del Medioevo occidentale, accanto ai classici latini. Questa tradizione sfruttò tutti i formati, materiali, impaginazioni e stili di scrittura disponibili, rispondendo alle esigenze di lettori laici e religiosi. Amanuensi con preparazione variabile lavorarono alla Commedia in contesti professionali e semiprofessionali, facendone un riferimento per la trasmissione materiale e grafica dei testi volgari. Tuttavia, le informazioni su questa varietà di evidenze materiali restano largamente scollegate. Gli apparati illustrativi ed esegetici presenti nei manoscritti testimoniano, inoltre, l'esigenza di spiegazioni verbali e visive per un testo semanticamente complesso. Tra XIV e XV secolo, numerosi commentatori professionisti interpretarono il poema, generando manoscritti ricchi di chiose: circa il 44% dei codici include apparati illustrativi, e l'8% presenta cicli iconografici estesi che supportano visivamente la lettura. Questo vasto metatesto richiede ancora definizione, classificazione e studio approfondito.
Ad oggi, manca un'edizione critica capace di gestire l'intera varia lectio della tradizione manoscritta. La ricostruzione del testo senza l'originale rappresenta una sfida per la critica testuale, in particolare con il metodo stemmatico. Nonostante i contributi di studiosi, le edizioni critiche si sono concentrate su tradizioni parziali e un numero limitato di manoscritti. Recentemente, Paolo Trovato ha ampliato la genealogia della tradizione (Tonello & Trovato, 2022), ma i loci critici analizzati rappresentano ancora una parte ridotta del testo. Le tecnologie digitali possono accelerare la classificazione e l'analisi di questa tradizione complessa, permettendo di studiarla per la prima volta nella sua interezza. In particolare, l'allineamento automatico delle trascrizioni integrali da HTR allo scopo di individuare lacune e vistosi errori (vd. sotto) può fornire argomenti più efficaci sulle relazioni genealogiche tra i testimoni. Un database esteso a tutta la tradizione manoscritta, collegato a un archivio digitale di facsimili e aggiornato tramite esperti e crowd-sourcing, semplificherebbe l'inserimento dei dati e fornirebbe informazioni verificabili. Tecniche avanzate di imaging diagnostico permetterebbero l'analisi dei manoscritti per dati invisibili all'occhio umano. L'intelligenza artificiale e il deep learning potrebbero migliorare la filologia materiale, testuale e attributiva, con progressi nel riconoscimento iconografico, nell'identificazione delle mani dei copisti, nella selezione delle varianti testuali e nell'analisi linguistico-stilistica di testi dall'attribuzione incerta, grazie a tecnologie come l'NLP.
Il Naples Dante Project, nato da un progetto PRIN 2022 e sintesi dei progetti danteschi sviluppati a Napoli nell'ultimo decennio, mira a integrare questa complessità in un'unica piattaforma online in cui le componenti del poema, dai supporti materiali alle varianti testuali e linguistiche, saranno dinamicamente collegate a riproduzioni digitali ad alta definizione, utilizzando Deep Learning, imaging multimodale, modellazione semantica dei dati e Computer Vision.

2. VERSO LA SVOLTA COMPUTAZIONALE DELLA CRITICA DANTESCA

Illuminated Dante Project (2015-in corso)

Lanciato nel 2015 con il finanziamento STAR, Illuminated Dante Project (IDP) ha costituito la base pionieristica del Naples Dante Project. Intende offrire un catalogo digitale dei manoscritti della Commedia con illustrazioni del poema e combina un archivio digitale ad alta risoluzione con un database codicologico

e iconografico. Online dal 2019, IDP ospita attualmente 101 dei 274 pezzi del corpus, diventando un punto di riferimento globale per la codicologia digitale dantesca e l'iconografia antica, ampiamente utilizzato nella ricerca e nell'educazione. Ospitato dall'Università di Napoli Federico II (www.dante.unina.it/idp), IDP aderisce ai principi F.A.I.R., con metadati conformi a TEI condivisi con Manus online (Istituto Centrale del Catalogo Unico-ICCU) e manifesti JSON per IIIF. In collaborazione con Biblissima, IDP è pioniere nell'adozione dell'API IIIF in Italia, portando l'Università Federico II a ospitare la Conferenza Annuale IIIF nel 2023. Il database MySQL, sviluppato da Luigi Tessarolo, ridefinisce le descrizioni iconografiche utilizzando categorie indicizzabili come "Attribuzione stilistica", "Tipologia decorativa" e "Iconografia", analizzando le relazioni immagine-testo e i soggetti. L'interfaccia di IDP supporta l'esplorazione per biblioteche di conservazione, descrizioni dettagliate o una mappa geografica delle posizioni dei manoscritti. Nel corso di un decennio, ha fornito accesso aperto e di alta qualità a risorse digitalizzate, promuovendo avanzamenti negli ambiti codicologico, paleografico, filologico e iconografico. I risultati della ricerca (2015-20) sono stati presentati a conferenze e pubblicati su riviste peer-reviewed (Ferrante, 2019; Autiero & Ferrante & Gorla & Picarelli, 2022; Ferrante, 2024a; Ferrante, 2024b).

Nuovi corpora I. Frammenti di Commedia (2021-in corso)

Nel 2021, grazie ad un finanziamento FRA, l'area di ricerca di IDP si è aperta a Fragments of Commedia (FraC), con l'obiettivo di digitalizzare e catalogare frammenti della Commedia databili tra XIV e XV secolo e di esaminarne filologicamente la lezione. Il progetto vuole rivalutare il ruolo dei frammenti, generalmente tralasciati dagli studi precedenti, nella comprensione della trasmissione materiale del poema e nelle relazioni genealogiche tra i suoi testimoni. Il team di paleografi della scuola napoletana ha stabilito il protocollo di ricerca, progettato il corpus e avviato la digitalizzazione, coprendo tutti i frammenti riutilizzati (la cosiddetta tradizione diretta) della Commedia. Questo sforzo ricostruisce testimoni perduti e indaga le circostanze del loro riutilizzo, con descrizioni che affrontano sia lo stato attuale dei frammenti che la loro provenienza originale. Dei 131 frammenti riutilizzati identificati, 40 sono già stati digitalizzati. Lo studio preliminare ha messo in discussione le idee precostituite sull'origine dei frammenti riutilizzati e ha rivelato il significato di materiali trascurati, in particolare i palinsesti. Tra le scoperte vi è un commento latino sconosciuto e un testimone monastico, che gettano luce sulla circolazione di Dante in contesti religiosi. I risultati sono stati presentati a eventi chiave, tra cui Entre fragments i prestatges / Entre fragmentos y anaqueles (Barcelona, 4-5/07/2024) e Naples Dante Project *Per una svolta computazionale della critica dantesca* (Napoli, 18-19/12/2024).

Nuovi corpora II. Illustrated Dante-ILLUDANT (2022-in corso)

Il campo di indagine sulle immagini si è inoltre allargato con Illustrated Dante (ILLUDANT), basato su una proposta FIS 2022, al fine di indagare i primi momenti della mise-en-images del poema: a una prima fase, che coincide in linea di massima con la tradizione manoscritta ed è caratterizzata da una sperimentazione iconografica varia, influenzata da periodo, area geografica, stile o bottega, segue una seconda, in cui si evidenzia una selezione di temi dalla tradizione manoscritta e l'offerta di nuove letture, veicolate dalla tradizione a stampa dei primi due secoli (XV-XVI). ILLUDANT mira a individuare le tendenze iconografiche e a proporre una classificazione su larga scala delle illustrazioni presenti nei manoscritti, nelle edizioni a stampa e nei disegni in serie del poema dantesco.
Lo scopo del progetto consiste nella creazione di un archivio di immagini compatibile con IIIF e un database delle illustrazioni della Commedia tra XIV e XVI secolo basato sui principi dei Linked Open Data (cf. sotto). Includerà il catalogo di IDP, le illustrazioni dei libri a stampa e i disegni in serie del XVI secolo, anonimi e attribuiti a Giovanni Stradano, Alessandro Allori, Federico Zuccari, Jacopo Ligozzi, Ludovico Cigoli. Con il contributo degli storici del libro della scuola napoletana, il team ha definito il protocollo di ricerca e delineato il corpus dei libri a stampa e dei disegni (secc. XV-XVI): dei 15 incunaboli della Commedia, solo 6 sono illustrati, mentre 23 edizioni del XVI secolo contengono un apparato illustrativo. ILLUDANT ha selezionato gli esemplari di stampa da digitalizzare in base alla qualità, integrità, allineamento al progetto editoriale e caratteristiche speciali come presenza eventuale di colorazione artistica e postillati.
Il progetto si pone la sfida di rivoluzionare l'approccio all'analisi iconografica tramite l'integrazione delle più recenti acquisizioni legate all'Intelligenza Artificiale. Sulla base del corpus si prevede di addestrare un modello di riconoscimento delle immagini basato su CNN, con lo scopo di esplorare nuove strade per una classificazione iconografica su larga scala e ad alta precisione.

Nuovi corpora III. Commenti Danteschi Antichi-CoDA (2023-in corso)

Il Corpus CoDA, definito da un progetto PRIN2022, è stato integrato recentemente nell'ecosistema del Naples Dante Project. Esso mira a creare un database dei commenti antichi alla Commedia di Dante in volgare e in latino prodotti entro la fine del XV secolo. Una collaborazione con CNR-OVI e Roma La Sapienza è stata avviata per stabilire un protocollo di ricerca che includa: 1. La creazione della Bibliografia dei Commenti Danteschi (BCD) accessibile tramite PLUTO. 2. Lo sviluppo del Corpus OVI-CoDA utilizzando il software GattoWeb 3.3 con lemmatizzazione selettiva per interrogazioni lessicografiche analitiche. Il corpus include attualmente 67 testi (più della metà dell'intero corpus). Grazie ad una recente collaborazione con CNR-ILC, è stato creato un modello per codificare i testi in formato XML TEI-P5, con l'obiettivo di ampliare il sistema di query a scopi esegetici e ermeneutici, oltre che lessicografici. CoDA potenzia il preesistente Dartmouth Dante Project, offrendo una documentazione all'avanguardia sulla prima esegesi dantesca, migliorando la qualità filologica dei testi e le innovazioni tecnologiche per ottimizzare le possibilità di interrogazione.

Transizione computazionale I. Acquisizione digitale (2017-in corso)

La durabilità e la referenzialità di un progetto come IDP è stata possibile anche grazie ai numerosi accordi istituzionali e scientifici che hanno accompagnato il suo sviluppo. In particolare quello con la Direzione Generale Biblioteche e diritto d'autore (2017–2023), l'ICCU e le varie biblioteche statali e non statali (ad es. Trivulziana) e importanti istituzioni internazionali (ad es. Bibliothèque nationale de France, Oxford Bodleian Libraries) per digitalizzare manoscritti e frammenti della Commedia miniati, concedendo licenze di riproduzione e riuso gratuito in cambio di metadati strutturati. L'evoluzione al Naples Dante Project e l'ampliamento dei corpora implicherà il rinnovo e l'estensione di questo tipo di accordi. Dal 2017, il team di ricerca napoletano collabora con l'infrastruttura View dell'Università di Lovanio per applicare tecniche di imaging diagnostico - Imaging Multispettrale, XRF e Reflectance Transformation Imaging (RTI) – ai manoscritti di IDP. Nel 2019, il Microdome Multispettrale (MSMD) è stato testato sul manoscritto Dante Filippino (CF 2.16), insieme ad una selezione di codici miniati conservati alla Biblioteca dei Girolamini. A novembre 2024, una parte del team si è unita al gruppo di tecnici di Lovanio a Stoccarda per acquisire immagini di un manoscritto della Commedia conservato presso la Württembergische Landesbibliothek.

Transizione computazionale II. Architettura dati (2023-in corso)

Contemporaneamente alla stipula di accordi e collaborazioni e alla definizione di nuovi corpora, si è reso necessario un riesame completo dell'architettura dei dati per allinearla ai modelli astratti e ai principi di Linked Open Usable Data. La nuova architettura del Naples Dante Project organizzerà i dati in cinque modelli descrittivi: manoscritti danteschi (Dante_ms), libri (Dante_bk), frammenti (Dante_frg), disegni (Dante_drw) e iconografia (Dante_icon), collegati con trascrizioni testuali (Dante_txt). Per gestire framework diversificati di database e archivi di immagini (database relazionali e No-SQL, IIIF) e formati di output (XML-TEI, ALTO, JSON, RDF), è stato adottato il framework CADMUS back-end, sviluppato da Daniele Fusi con l'Università di Venezia. Questo sistema consente: a. Gestione di dati eterogenei tramite modelli indipendenti dall'implementazione. b. Modelli dinamici per un'espansione senza soluzione di continuità. c. Un ambiente unificato per database centralizzati, interfacce grafiche, editing, validazione e ricerche in tempo reale. Tale architettura, concepita per essere replicabile nelle sue componenti, può aprire la strada a future iniziative umanistiche computazionali.

Transizione computazionale III. Modellazione semantica (2024-in corso)

Nel suo originario sviluppo del PRIN 2022, NDP ha avviato una collaborazione con il Dipartimento di Matematica e Informatica di Catania per sviluppare una gerarchia concettuale per modelli relativi ai media danteschi (codicologici, bibliografici e figurativi) e all'indicizzazione iconografica (immagine). Il framework semantico mira a integrarsi con i cinque modelli di dati definiti dal gruppo di lavoro sull'Architettura dati (cf. sopra), utilizzando gli standard CIDOC CRM e LRMoo. Una mappatura preliminare del modello Dante_ms ha portato ai seguenti avanzamenti: a. Modellazione gerarchica delle opere e delle espressioni: la distinzione di Work (F1), Expression (F2), Manifestation (F3) e Item (F5) secondo LRMoo chiarisce diversi aspetti della Commedia, con sottoclassi per edizioni critiche, varianti testuali e dettagli macrotestuali (es. Cantiche e Canti). b. Descrizioni avanzate di manoscritti e testi: i Linguistic Objects (E33), collegati a Strings (E62), rappresentano colofoni, dediche e rubriche, applicabili ai manoscritti. c. Unità composite: LRMoo accoglie manoscritti con progetti codicologici o strati testuali multipli. d.

Interoperabilità con CADMUS: l'integrazione basata su dati JSON consente una elaborazione automatizzata dei dati conforme ai principi F.A.I.R.

Transizione computazionale IV. Handwriting Text Recognition (2024-in corso)

L'analisi automatica di manoscritti medievali e moderni tramite deep learning sta acquisendo importanza per migliorare e velocizzare la ricerca paleografica. Tuttavia, i progressi nel riconoscimento degli scribi sono stati limitati dalla mancanza di grandi quantità di dati sufficientemente annotati, poiché il processo di annotazione è complesso e dispendioso in termini di tempo. Sfruttando la classificazione paleografica dei codici danteschi già acquisita dalla bibliografia e in corso di acquisizione per la catalogazione dei manoscritti e dei frammenti di NDP, è stata recentemente avviata una collaborazione con Paolo Merialdo e il gruppo di ricerca in Paleografia Computazionale di Roma 3, che ha ottenuto risultati promettenti con i dati della Biblioteca Vaticana (Lastilla & Ammirati & Firmani & Komodakis & Merialdo & Scardapane, 2022). La collaborazione mira ad addestrare un tool di features paleografici basato su CNN utilizzando dati di Ground Truth (GT) sulla paleografia dantesca fornita dal gruppo napoletano. Un esperimento preliminare prevede il pre-addestramento della CNN su dati GT provenienti da 17 manoscritti di tre copisti di primo Trecento attivi digraficamente nelle tipologie della textualis e della cancelleresca.

Transizione computazionale V. HTR-Dante (2024-in corso)

NDP ha recentemente adottato tecniche avanzate di Deep Learning per trascrivere, allineare e riconoscere la più antica varia lectio della Commedia. Col supporto del PRIN 2022, del PNRR-Changes 2023-25 e del Dipartimento di Eccellenza 2023-27, tra ottobre e dicembre 2024 Thibault Clérice, ricercatore all'INRIA di Parigi e esperto di tecniche avanzate di Handwritten Text Recognition e Natural Language Processing, è stato invitato presso il Laboratorio di Digital Humanities dell'Università Federico II (https://www.fedhlab.unina.it), avviando ufficialmente il progetto HTR_Dante, con l'obiettivo di addestrare modelli HTR sulla piattaforma eScriptorium. Si tratta del primo progetto su larga scala per un'opera letteraria medievale italiana. Quanto all'annotazione per il training di Kraken - il pacchetto OCR alla base dell'ambiente eScriptorium - HTR_Dante si conforma alle linee guida generali del gruppo di ricerca HTRomance dell'INRIA, e intende rispettare le consegne, per ora ancora ufficiose, della recentemente costituita Commissione Nazionale per la Trascrizione e l'Edizione dei Testi in Italiano. Il gruppo di lavoro ha definito protocolli specifici per l'HTR di codici danteschi che hanno dovuto affrontare situazioni speciali come la risoluzione dello scriptio continua, la gestione delle lettere maiuscole per le iniziali delle terzine, e le variazioni formali dei copisti. Per il training dei modelli HTR sono stati isolati quattro tipi di scrittura come peculiari della tradizione manoscritta della Commedia: a. Scrittura cancelleresca (Chancery script). b. Scrittura mercantesca (Merchant script). c. Gotica/Testuale (Gothic/Textualis script). d. Scrittura umanistica (Humanistic script). In questa ricerca preliminare, la scrittura cancelleresca è stata la prima ad essere analizzata utilizzando sei manoscritti del gruppo stemmatico a_0 delineato da Tonello (Tonello, 2018). Dopo aver risolto gli iniziali problemi di trascrizione automatica attraverso l'addestramento del modello, l'HTR realizzato con eScriptorium e il modello adattato CATmuS Medieval 1.5 ha raggiunto un'accuratezza superiore al 90%. Il workflow sarà esteso agli altri tre tipi di scrittura, con l'obiettivo di ottimizzare l'HTR per la tradizione dantesca e verificare le ipotesi stemmatiche più recenti.

Transizione computazionale VI. Riconoscimento automatico delle varianti (2024-in corso)

La fertile collaborazione in situ con Thibault Clérice (INRIA), Angelo Del Grosso e Federico Boschetti (CNRILC) ha aperto un ulteriore campo di indagine per automatizzare la collazione dei testi processati con HTR al fine di identificare errori significativi nella varia lectio integrale della tradizione e verificare le relazioni genealogiche tra i testimoni. Basandosi sul principio metodologico lachmanniano che le lacune testuali sistematiche possono rivelare congiunzioni significative tra i testimoni di un testo (Chiesa, 2022), Clérice ha sviluppato un algoritmo per rilevare lacune nei codici finora trascritti dal gruppo di lavoro HTR_Dante, utilizzando la versione XML dell'edizione critica di Petrocchi del 1966 come Ground Truth. Entro la fine di novembre 2024, l'algoritmo è stato perfezionato e applicato con successo a due manoscritti della Commedia, identificando tutte le loro lacune testuali (da singoli versi a più terzine) e variazioni significative mai riconosciute negli studi filologici precedenti. Questo rappresenta un progresso rivoluzionario nella metodologia della critica testuale, consentendo la ricostruzione dei testi poetici diffusi in ampie tradizioni manoscritte. I risultati preliminari, presentati alla conferenza internazionale di dicembre 2024 (Naples Dante Project. Per una svolta computazionale della critica dantesca), hanno dimostrato il potenziale di questo approccio. Il team prevede di applicare il modello a 20 manoscritti, perfezionarne le

capacità e presentare i risultati alla conferenza Computational Humanities Research 2025. Un'ulteriore fase consisterà nell'esame automatico delle varianti sul più ampio numero possibile di trascrizioni digitali dei codici della Commedia (obiettivo ideale: 400 testimoni su 580 della tradizione non frammentaria) con comunicazione dei risultati su riviste di alto profilo internazionale (es. Nature)

Transizione computazionale VII. Dante Critical Texts (2024-in corso)

Sempre nell'ambito del progetto PRIN 2022, e nell'ottica dello studio comparato del testo criticamente emendato della Commedia, NDP ha inaugurato Dante Critical Texts (DCT), uno strumento progettato per confrontare i testi della Commedia di Dante a partire da edizioni fondate su un esame di tipo stemmatico della tradizione manoscritta. Il progetto mira a rendere accessibili i principali risultati della filologia dantesca (emendatio), dall'edizione di Karl Witte del 1862 fino alle più recenti. DCT offre tre principali funzioni: a. Search: consente agli utenti di cercare occorrenze verbali. b. Texts: permette l'analisi comparativa tra le edizioni, evidenziando divergenze testuali. c. Editions: fornisce descrizioni concise e dettagliate delle varie edizioni. La funzione Texts include attualmente l'edizione di Petrocchi del 1966 (Petrocchi, 1966-67), quella di Malato del 2021 (Malato, 2021) e i testi critici dell'Inferno editi da Vandelli (Vandelli, 1921) e Tonello-Trovato (Tonello & Trovato, 2022). La pubblicazione del Purgatorio di Vandelli (Vandelli, 1921) è prevista a breve. La funzione Editions include già una descrizione dell'edizione di Petrocchi del 1966.

Il quadro delle ricerche appena delineato dimostra che Naples Dante Project sta puntando a integrare la complessità del suo campo di indagine in un unico aggregatore, in cui i vettori materiali del poema, le varianti testuali e gli apparati iconografici ed esegetici siano classificabili con precisione, collegabili in modo dinamico a riproduzioni ad alta risoluzione, oltre che ricercabili in contesto LOD. L'adozione di tecniche di Deep Learning, Imaging Multimodale, Modellazione Semantica dei Dati e di Computer Vision per interconnettere caratteristiche materiali, testuali, meta-testuali, linguistiche e figurative, offriranno una rappresentazione articolata e completa dell'opera dantesca. La divulgazione dei risultati, dagli archivi digitali ai modelli di intelligenza artificiale, in open access e tramite un'unica piattaforma, garantirà una piena condivisione dei risultati del progetto non soltanto con la comunità accademica e i ricercatori ma soprattutto con l'utenza non specialistica, in un'ottica di comunità di ricerca inclusiva e globale che possa confrontarsi con la tradizione testuale e iconografica della Commedia.

RINGRAZIAMENTI

Il progetto è finanziato da: PRIN 2022 - NAPLES DANTE PROJECT; Spoke 3 del PNRR PE 5 Changes - Digital Libraries, Archives and Phylology; Dipartimento di Eccellenza 2023-2027, Dipartimento di Studi Umanistici, Università degli Studi di Napoli Federico II.

BIBLIOGRAFIA

Autiero, F., & Ferrante, G., & Gorla, S., & Picarelli, S. (2022). Dante e Petrarca allo (stesso) scrittoio. Per lo sviluppo di un'ontologia di IDP a partire dall'istanza manoscritti di Itinera. Culture digitali. Intersezioni: filosofia, arti, media. Preceedings della 11a conferenza nazionale AIUCD. Ciracì, F., & Miglietta, G., & Gatto, C. (Eds.). 169-171.

Chiesa, P. (2022). Elementi di critica testuale. Bologna: Patròn.

Ferrante, G. (2019). Illuminated Dante Project. Per un archivio digitale delle piú antiche illustrazioni della Commedia. I. Un case study quattrocentesco (mss. Italien 74, Riccardiano 1004 e Guarneriano 200). Dante visualizzato. Carte ridenti ii: XV secolo. Prima parte. Ciccuto, M., & Livraghi, L.M.G. (Eds.) Firenze: Cesati, 229-255.

Ferrante, G. (2024). Premessa. La Commedia di Dante nell'antica tradizione manoscritta. Per uno studio integrato e computazionale dei codici, dei testi e delle immagini. Ferrante, G. (Ed). Roma-Padova: Editrice Antenore, 7-11.

Ferrante, G. (2024). Verso il Naples Dante Project. Recensio digitale integrale dei codici della Commedia di Dante. "A guisa d'uom che 'n dubbio si raccerta". Vecchie e nuove prospettive per la biografia e l'opera dantesca. Bardi, R., & Canova, L. (Eds.). Firenze: Cesati, 145-160.

Lastilla, L., & Ammirati S., & Firmani D., & Komodakis N., & Merialdo P., & Scardapane S. (2022). Selfsupervised learning for medieval handwriting identification: A case study from the Vatican Apostolic Library. Information Processing & Management, 59, 3, 1-20.
Malato, E. (Ed.). 2021. Dante Alighieri. Inferno. Roma: Salerno Editrice.
Petrocchi, G. (Ed). 1966-67. Dante Alighieri. La Commedia secondo l'antica vulgata. Milano: Mondadori.
Tonello, E. (2018). Sulla tradizione tosco-fiorentina della 'Commedia' di Dante (secoli XIV-XV). Padova: libreriauniversitaria.it.
Tonello, E., & Trovato, P. (Eds.). 2022. Dante Alighieri. Commedia. Inferno. Padova: libreriauniversitaria.it.
Vandelli, G. (Eds.). 1921. Dante Alighieri. La Divina Commedia. Firenze: Bemporad.Tonello, E., & Trovato, P. (Eds.). 2022. Dante Alighieri. Commedia. Inferno. Padova: libreriauniversitaria.it.
Vandelli, G. (Eds.). 1921. Dante Alighieri. La Divina Commedia. Firenze: Bemporad.

Philo-1-preview
Un modello T5-Base per l'emendazione dei testi antichi

Giuseppe Ferrara[1]
[1] Università degli Studi di Siena, Italia - giuseppe.ferrara@student.unisi.it

ABSTRACT (ITALIANO)
In questo paper, si presenta *Philo-1-preview*, un *Large Language Model* di tipo *Seq2Seq* di 297 milioni di parametri ottenuto mediante un'operazione di *fine-tuning* condotta su *PhilTa*, una variante *encoder-decoder* multilingue di T5-Base pre-addestrata sul greco antico e sul latino, e specializzato nell'individuazione e nella correzione degli errori di maiuscola, minuscola, pronuncia, dittografia e aplografia, combinati o isolati, che si possono riscontrare nel testo tradito di molte opere antiche. Il suo sviluppo è stato guidato da un approccio al problema diverso da quelli già sperimentati in progetti precedenti, quali *Bentley* e *Ianus*, di cui *Philo-1-preview* e il suo sistema di *front-end Ianus AI* rappresentano i diretti discendenti, e *Latin BERT* e *Logion*: l'emendazione non viene più trattata come un processo meccanico da risolvere con un algoritmo *rule-based* o come un *task* di *fill-mask*, ma come un *task* di generazione di testo (*text-to-text*) combinato con un'operazione di *denoising*. I *dataset* utilizzati per il *fine-tuning* del modello sono stati ricavati dai *corpora* di testi greci e latini presenti su *Perseus Digital Library*, dopo alcune opportune operazioni di *pre-processing* volte a isolarne le parti effettivamente scritte nelle lingue di interesse e a normalizzarne i criteri di trascrizione. In fase di valutazione, *Philo-1-preview* ha mostrato prestazioni molto elevate nel compito per cui è stato progettato, con un *BLEU score* di 67.50 e un *cross-entropy loss* di 0.078. Il confronto con *Latin BERT* e *Logion* ha evidenziato che *Philo-1-preview* fornisce correzioni dei testi antichi di qualità superiore a quelle degli altri modelli, almeno per gli errori su cui è stato condotto il *fine-tuning*. Per superare i limiti di questo modello, si è già iniziato a lavorare allo sviluppo di un modello BART che sia in grado di identificare e correggere, in autonomia, tutte le tipologie di errori riscontrabili nella tradizione delle opere antiche.
Parole chiave: filologia digitale; Large Language Model; Ianus AI; Philo-1-preview; emendazione

ABSTRACT (ENGLISH)
Philo-1-preview. A T5-Base model for the emendation of ancient texts
In this paper, I present *Philo-1-preview*, a *Seq2Seq Large Language Model* with 297 million parameters obtained through a fine-tuning process conducted on *PhilTa*, a multilingual encoder-decoder variant of T5-Base pre-trained on Ancient Greek and Latin. It is specialized in identifying and correcting majuscule and minuscule scripts, pronunciation, dittography and haplography errors, whether combined or isolated, which can be found in the textual transmission of many ancient works. Its development has been guided by an approach to the problem different from those already tested in previous projects, such as *Bentley* and *Ianus*, of which *Philo-1-preview* and its front-end system *Ianus AI* are the direct descendants, and *Latin BERT* and *Logion*. In this new approach, emendation is no longer treated as a mechanical process to be solved with a rule-based algorithm or as a fill-mask task, but rather as a text-generation (text-to-text) task combined with a denoising operation. The datasets used for the model's fine-tuning were derived from the Greek and Latin text corpora in the *Perseus Digital Library*, after some pre-processing steps aimed at isolating the parts actually written in the languages of interest and to normalize their transcription criteria. During the evaluation phase, *Philo-1-preview* showed very high performance in the task for which it was designed, with a BLEU score of 67.50 and a cross-entropy loss of 0.078. A comparison with *Latin BERT* and *Logion* revealed that *Philo-1-preview* provides higher-quality corrections of ancient texts than the other models, at least for the types of errors on which it was fine-tuned. In order to overcome the limitations of this model, work has already begun on the development of a BART model capable of autonomously identifying and correcting all the types of errors that can be found in the textual tradition of ancient works.
Keywords: digital philology; Large Language Model; Ianus AI; Philo-1-preview; emendation

1. INTRODUZIONE
In questo contributo, si presenta *Philo-1-preview* la versione beta di un *Large Language Model* (LLM) di tipo *Seq2Seq* basato sull'architettura T5 (*Text-to-Text Transfer Transformer*) (Raffel et al., 2019) pensato

per semplificare l'emendazione dei testi antichi, greci e latini.[1] Lo sviluppo di tale modello si è reso necessario dopo aver constatato la scarsa diffusione di tecnologie che siano in grado di assistere, con profitto, i filologi classici in questa fase del processo di edizione di un'opera antica. I LLMs sono stati finora impiegati principalmente per agevolare processi di analisi testuale caratteristici del *Natural Language Processing* (NLP), come il *PoS Tagging*, il *Dependency Parsing* e la lemmatizzazione, mentre il loro impiego per la ricostruzione degli originali dei testi antichi è limitato a pochi casi isolati, alcuni dei quali esulano dall'ambito delle lingue classiche:

1. *MsBERT*, un modello BERT pre-addestrato su un *dataset* comprensivo di trascrizioni di manoscritti ebraici, del *Pre-Print Rabbinic Corpus* e del *Comprehensive Rabbinic Corpus*, per l'integrazione dei testi ebraici trasmessi in forma frammentaria (Shmidman et al., 2024);
2. una RNN addestrata sui *Coptic SCRIPTORIUM Corpora* per l'integrazione delle lacune dei manoscritti coptici (Levine et al., 2024);
3. *Pythia* (Assael et al., 2019) e *Ithaca* (Assael et al., 2022), due modelli, rispettivamente un LSTM di tipo *Seq2Seq* e un modello basato sull'architettura *Transformer*, per l'integrazione delle iscrizioni greche pervenute mutile;
4. *Latin BERT* (Bamman & Burns, 2020), un modello BERT pre-addestrato su vari *corpora* di testi latini (*Corpus Thomisticum*, *Internet Archive*, *The Latin Library*, *Patrologia Latina*, *Perseus Digital Library*, *Vicipaedia*) per l'integrazione delle lacune in essi presenti;
5. *Logion* (Cowen-Breen et al., 2023; Graziosi et al., 2023), un modello BERT pre-addestrato su una versione ampliata del *corpus* di testi greci premoderni di Singh et al. (2021) per l'emendazione dei testi greci.

Tra questi modelli, quelli specializzati nei domini del greco antico e del latino sono caratterizzati da livelli di *top-1 accuracy* non ancora soddisfacenti nei *task* di emendazione o integrazione: 32.6% per *Pythia* (Assael et al., 2022: 282); 61.8% per *Ithaca* (Assael et al., 2022: 282); 33.1% per *Latin BERT* (Bamman & Burns, 2020: 4–6); 58.3% per *Logion* (Cowen-Breen et al., 2023: 8).

Philo-1-preview si inserisce nel panorama dei modelli attualmente disponibili apportando alcune novità significative agli approcci già sperimentati. L'impostazione adottata per sviluppare il modello affronta il problema della ricostruzione del testo originario di un'opera antica come un *task* di generazione di testo (*text-to-text*) combinato con un'operazione di cancellazione del rumore presente nell'*input* del modello (*denoising*). *Latin BERT* e *Logion* trattano il medesimo problema come un *task* di *fill-mask* che rispecchia il *task* di *Masked Language Modeling* (MLM) usato durante il loro pre-addestramento (Bamman & Burns, 2020: 4–6; Cowen-Breen et al., 2023: 7–8): in altri termini, essi cercano di predire la parola più adatta a colmare una lacuna artificialmente introdotta, all'interno della frase, in corrispondenza di una parola considerata errata. Da un lato, tale approccio permette di utilizzare direttamente i modelli pre-addestrati per l'emendazione dei testi antichi senza necessariamente fine-tunarli, dal momento che il compito per cui vengono impiegati è coerente con la metodologia utilizzata durante il *pre-training* per far acquisire al modello le competenze linguistiche desiderate. Dall'altro, un simile approccio richiede uno sforzo implementativo aggiuntivo, rendendo indispensabile lo sviluppo di un sistema che permetta al modello di individuare correttamente le parole errate presenti in una data frase, nel caso in cui non le si voglia segnalare esplicitamente tramite il *token* [MASK]. Trattare, invece, l'emendazione di un testo antico come un *task* di *denoising* attraverso un modello *Seq2Seq* basato sull'architettura T5 permette di implementare l'individuazione automatica, da parte del modello, degli errori presenti in quel testo.

Lo sviluppo di *Philo-1-preview* rappresenta il risultato di un lavoro di ricerca più ampio articolato in tre tappe:

1. l'implementazione, nel 2020, di *Bentley*, un *software rule-based* per la correzione degli errori di minuscola presenti nei testi latini (Ferrara, 2023: 85);
2. lo sviluppo, nel 2023, di *Ianus*, una nuova versione di *Bentley* che estende la correzione degli errori di minuscola anche al greco antico e introduce la correzione degli errori di maiuscola e pronuncia in entrambe le lingue (Ferrara, 2023);
3. l'implementazione, iniziata nel 2024, di *Ianus AI*, un *software* per un'emendazione dei testi antichi assistita dall'Intelligenza Artificiale.

[1] Una volta terminata la fase di *beta testing*, il modello e il suo sistema di *front-end Ianus AI* verranno distribuiti sotto licenza GNU GPLv3 rispettivamente su https://huggingface.co/giuseppeferrara/Philo-1-preview (cons. 10/04/2025) e https://github.com/giuseppeferrara/ianus-ai.git (cons. 10/04/2025).

L'evoluzione metodologica segnata dall'ultima tappa di questo percorso, che ha sostituito gli algoritmi deterministici implementati nelle due versioni precedenti del *software* con i LLMs, si è resa indispensabile per migliorare la capacità predittiva di *Ianus* attraverso la valutazione del contesto della frase per l'elaborazione delle congetture e per rendere la nuova versione del *software* capace di individuare autonomamente le corruttele presenti nel testo tradito.

2. ARCHITETTURA DEL MODELLO, DATASET E FINE-TUNING

Philo-1-preview è un modello di 297 milioni di parametri ottenuto attraverso un'operazione di *fine-tuning* condotta su *PhilTa* (Riemenschneider & Frank, 2023), una variante *encoder-decoder* multilingue del modello T5-Base (Raffel et al., 2019) pre-addestrata su vari *corpora* di testi greci e latini (*Open Greek & Latin Project*, *Greek Medieval Texts* di CLARIN, *Patrologia Graeca*, *Corpus Corporum*, *Project Gutenberg*). Il *fine-tuning* ha avuto come obiettivo dotare il modello della capacità di individuare e correggere autonomamente gli errori presenti nel testo tradito delle opere antiche.

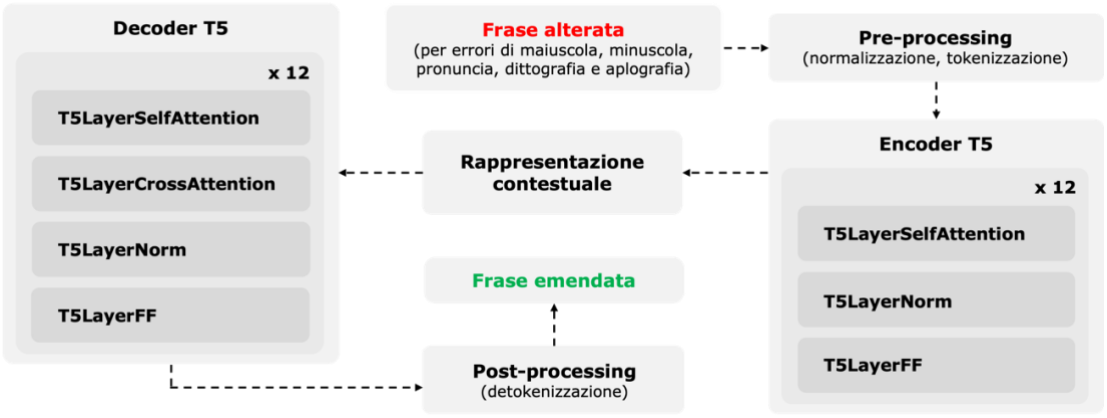


Figura 1. Architettura e flusso logico di *Philo-1-preview* e di *Ianus AI*, il suo sistema di *front-end*.

L'architettura di *Philo-1-preview*, integrata con il flusso logico del modello e del suo sistema di *front-end*, è riportata in Fig.1. La frase alterata per la presenza di errori di maiuscola, minuscola, pronuncia, dittografia e aplografia (di singole lettere) viene pre-processata ed entra nell'*encoder* del modello, che ne produce una rappresentazione contestuale tenendo conto delle relazioni tra i singoli *token*. Questa rappresentazione viene inviata al *decoder* del modello che predice la sequenza di *token* della frase emendata. Quest'ultima verrà convertita in una stringa testuale grazie alla detokenizzazione che ha luogo durante il *post-processing*.

I *dataset* utilizzati nelle fasi di *training*, validazione e valutazione sono stati ricavati dai *corpora* di testi greci e latini presenti su *Perseus Digital Library*, dopo alcune opportune operazioni di *pre-processing*:

1. eliminazione degli elementi paratestuali in essi presenti (tag XML, commenti editoriali, note, apparati critici);
2. normalizzazione dei testi, articolata in:
    a) sostituzione di tutte le lettere maiuscole con le corrispondenti minuscole;
    b) eliminazione delle lettere ramiste da quelli latini;
    c) sostituzione della iota sottoscritta con quella ascritta in quelli greci;
    d) rimozione dei segni diacritici nelle lettere interessate da possibili fraintendimenti da quelli greci.

In questo modo, si sono ottenuti due file CSV, uno per il greco antico e uno per il latino, contenenti, come *record*, coppie di frasi, di cui una alterata (per errori di maiuscola, minuscola, pronuncia, dittografia e aplografia isolati o combinati) e l'altra corretta, di lunghezza massima pari a 15 parole. Le regole per produrre le frasi nella loro versione alterata sono state, in parte, ricavate dall'algoritmo di *Ianus* e, in parte, formulate *ex novo*. Il *dataset* complessivo (con 1028111 *record*) è stato suddiviso in *dataset* di *training* (90% dei *record* totali) e *dataset* di valutazione (10% dei *record* totali) (Abraham et al., 2024: 390).

Gli iperparametri da utilizzare durante il *fine-tuning*, riportati in Tab.1, sono stati trovati attraverso un'operazione di ottimizzazione condotta con *Optuna* e articolata in 10 *trial* effettuati su un *subset* pari al 10% del *dataset* finale. La loro ricerca e il *training* effettivo del modello, della durata di circa 13 ore, sono stati svolti su una GPU NVIDIA A100.

Iperparametro	Valore
Numero di epoche	5
Learning rate	0.0002123818291492695
Batch size	16
Weight decay	0.2984303550950661

Tabella 1. Iperparametri utilizzati durante il processo di *fine-tuning* di *PhilTa*.

3. VALUTAZIONE ED EFFICACIA DEL MODELLO

Le *performance* di *Philo-1-preview* nel *task* di interesse sono state valutate attraverso due metriche: il *BLEU* (*Bilingual Evaluation Understudy*) *score* (Papineni et al., 2001), utilizzato per ottenere una misura quantitativa della qualità delle correzioni proposte dal modello mediante il confronto tra la frase prodotta in *output* e la frase corretta di riferimento; il *cross-entropy loss*, che, confrontando la distribuzione di probabilità prevista dal modello con quella reale dei *token* della frase corretta, fornisce una misura degli scarti del modello stesso (Rao & McMahan, 2019: 46–48). Per massimizzare l'attendibilità della prima delle due metriche, si sono ignorati i *token* di *padding*.

Metrica	*PhilTa*	*Philo-1-preview*	Variazione di *performance*
BLEU score	2.03	67.50	+65.47
Cross-entropy loss	3.25	0.078	-3.17

Tabella 2. Confronto tra i valori delle metriche di valutazione delle *performance* di *PhilTa* e *Philo-1-preview*.

Più in dettaglio, la valutazione di *Philo-1-preview* si è articolata in due fasi. La prima ha avuto come obiettivo la verifica del corretto aggiornamento dei pesi del modello prodotto dal processo di *fine-tuning*. A tale scopo, si è effettuato un confronto tra le *performance* di *PhilTa* e quelle di *Philo-1-preview* sul medesimo *dataset* di valutazione (i risultati sono riportati in Tab.2). *Philo-1-preview* ha ottenuto un *BLEU score* di 67.50, contro il 2.03 di *PhilTa*, dimostrando di aver effettivamente acquisito, attraverso il processo di *fine-tuning*, la capacità di produrre correzioni del testo analizzato di qualità molto elevata. Il *cross-entropy loss* del modello è stato di 0.078, contro il 3.25 di *PhilTa*.
La seconda fase ha avuto come obiettivo la valutazione dell'effettivo contributo di *Philo-1-preview* al processo di emendazione dei testi antichi, attraverso il suo confronto con i modelli *gold-standard* sopra menzionati: *Latin BERT* e *Logion*. Si ricordi, per inciso, che questi ultimi, al contrario di *Philo-1-preview*, non sono stati fine-tunati, ma soltanto pre-addestrati. Questo, tuttavia, non invalida il confronto e i risultati ottenuti per due ragioni: in primo luogo, perché gli sviluppatori di *Latin BERT* e *Logion* presentano esplicitamente questi modelli come destinati, anche o esclusivamente, all'emendazione dei testi antichi nella loro versione pre-addestrata; in secondo luogo, perché la valutazione, nel loro caso, è stata effettuata non sul *task* di generazione di testo e di *denoising*, come per *Philo-1-preview*, ma sul *task* di *fill-mask* per cui essi stessi sono stati ottimizzati. A questo scopo, si è resa necessaria una modifica dei *record* del *dataset* di valutazione per poterne riadattare la struttura al compito di *gap-filling* e poterli utilizzare come *input* di *Latin BERT* e *Logion*: le frasi alterate contenute nelle coppie di frasi selezionate (per il resto rimaste inalterate) sono state modificate sostituendo le loro parole errate con il *token* speciale [MASK].

Modello	*BLEU score*
Philo-1-preview	67.50
Latin BERT	0.76
Logion	0.81

Tabella 3. Confronto tra i *BLEU score* di *Philo-1-preview*, *Latin BERT* e *Logion*.

I risultati del confronto, riportati in Tab.3, mostrano che *Philo-1-preview* è, tra i modelli disponibili, quello che fornisce correzioni dei testi antichi più qualitative, almeno per gli errori di maiuscola, minuscola, pronuncia, dittografia e aplografia (di singole lettere). Nella correzione di queste categorie di errori, intesa come predizione delle parole più adatte a colmare le lacune introdotte al posto delle parole errate, *Latin BERT* e *Logion* hanno ottenuto un *BLEU score* molto basso, rispettivamente di 0.76 e 0.81.

4. IL SISTEMA DI *FRONT-END* E LA SUA *GUI*

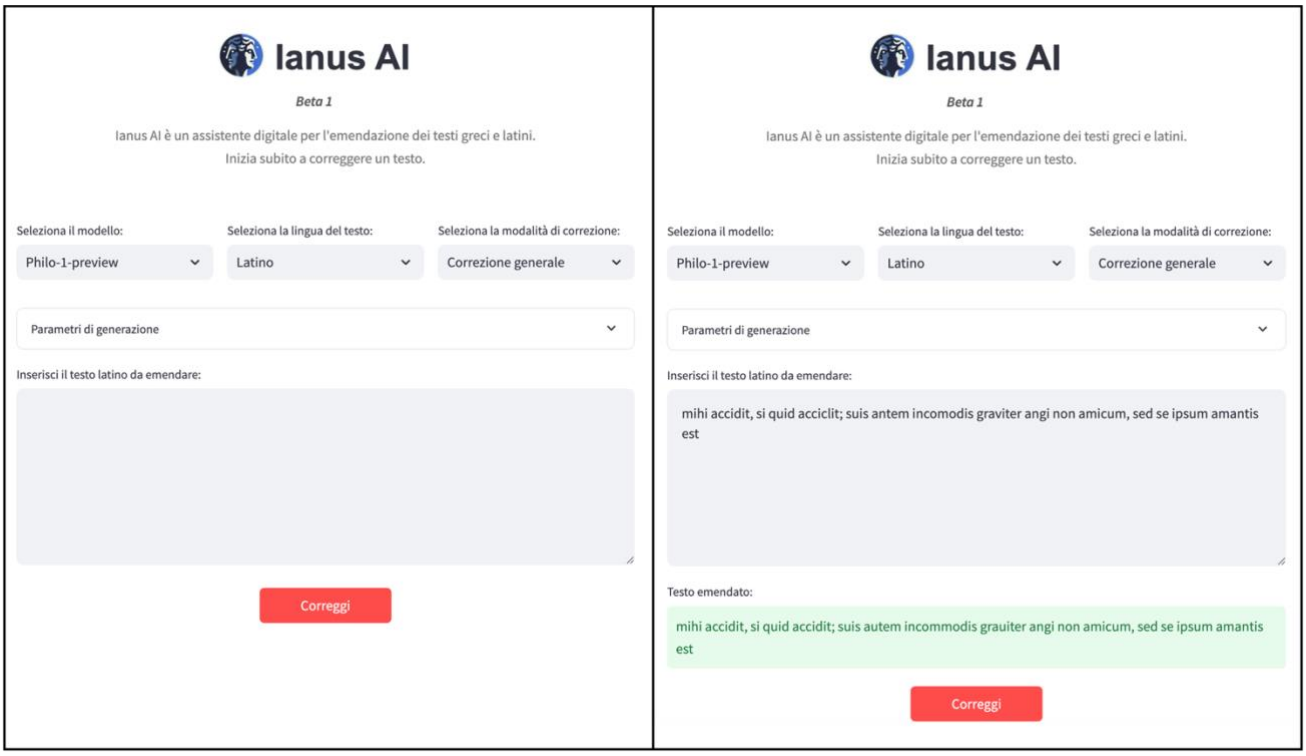


Figura 2. Interfaccia grafica di *Ianus AI*, il sistema di *front-end* di *Philo-1-preview*.

Per rendere il modello facilmente interrogabile, è stato realizzato anche un sistema di *front-end* dotato di una *Graphic User Interface* (GUI) semplice e intuitiva (vd. Fig.2), utilizzando il *framework open-source Streamlit*. I principi che hanno guidato lo sviluppo della GUI sono stati:

1. creare una gerarchia visiva tra gli elementi dell'interfaccia;
2. implementare un design pulito e minimalista;
3. ridurre la curva di apprendimento dell'utente per massimizzare l'usabilità dell'applicazione;
4. responsività.

Il primo principio si è tradotto nell'utilizzo, per le varie componenti dell'interfaccia, di dimensioni tipografiche diverse, finalizzate a rendere facilmente distinguibili gli elementi descrittivi da quelli funzionali. Il secondo principio si è tradotto nell'utilizzo di un *layout* semplice e arioso attraverso una corretta gestione degli spazi bianchi, per far focalizzare l'attenzione dell'utente sui *task* principali dell'applicazione. Il terzo principio è stato implementato attraverso un sistema di *feedback* offerti all'utente sotto forma di avvisi e messaggi di errore e attraverso un'interfaccia lineare in grado di guidare l'utente durante l'intero processo di correzione. Il quarto principio si è tradotto nello sviluppo di un'interfaccia dotata di un *layout* fluido e flessibile, in grado di adattarsi, in autonomia, alle differenti risoluzioni degli schermi dei dispositivi, senza sacrificare la funzionalità dell'applicazione.
L'utente, dopo aver selezionato il modello da utilizzare per l'emendazione, la lingua del testo immesso e la modalità di correzione (generale, che corregge autonomamente l'intero testo immesso, o mirata, che corregge soltanto le parole del testo segnalate al modello dall'utente), inserisce l'*input* testuale corrotto in un *widget TextArea* e, cliccando sul pulsante *Correggi*, innesca le operazioni di *pre-processing*, inferenza e *post-processing*, che porteranno all'elaborazione del testo emendato. Quest'ultimo verrà mostrato in un riquadro apposito che comparirà, dinamicamente, sotto il *widget TextArea*.

5. CONCLUSIONI: LIMITI E SVILUPPI FUTURI

Per concludere la presentazione di questa ricerca, è opportuno soffermarsi brevemente sui suoi limiti, in quanto essi possono essere considerati il punto di partenza per ulteriori sviluppi futuri.
Il primo limite è rappresentato dal fatto che *Philo-1-preview* è stato ottimizzato soltanto per la correzione di alcune categorie di errori filologici. Per espanderne il raggio di azione, si sta già lavorando allo sviluppo di un modello più ampio, basato su BART (*Bidirectional and Auto-Regressive Transformers*) (Lewis et al., 2019), che sia in grado di identificare e correggere tutte le tipologie di errori riscontrabili nella tradizione di

un'opera antica. La scelta di questo modello dipende dalla sua particolare architettura, che combina la capacità di comprensione profonda del testo, tipica di modelli bidirezionali come BERT (Devlin et al., 2019), e l'abilità di generazione sequenziale, caratteristica di modelli autoregressivi come GPT (*Generative Pre-trained Transformer*) (Radford & Narasimhan, 2018; Lewis et al., 2019: 1–2).
Il secondo limite è rappresentato dalle dimensioni del modello in termini di numero di parametri. Per quanto *Philo-1-preview* non sia, tecnicamente, tra i modelli più piccoli, aumentarne il numero complessivo di parametri potrebbe permettergli di raggiungere una comprensione più profonda delle relazioni sintattiche e delle sfumature semantiche presenti nei testi greci e latini e di produrre correzioni di qualità ancora superiore (Kaplan et al., 2020; Gazit, 2024: 174). Questo obiettivo potrà essere raggiunto apportando alcune modifiche all'architettura del modello, al processo di tokenizzazione e alla fase di *training*.
Il terzo limite risiede nel fatto che non si è sviluppato *ex novo* un modello specifico per il greco antico e per il latino, ma si è utilizzato un modello pre-addestrato, del quale non è stato possibile verificare, in via preliminare, l'effettiva qualità dell'intero processo di *pre-training*, fondamentale per garantire la capacità del modello di produrre inferenze corrette. Questa criticità verrà superata attraverso lo sviluppo, controllato in ogni sua fase, del modello BART cui si è fatto cenno sopra.
Il quarto limite della ricerca risiede nell'aver utilizzato, per il *fine-tuning* di *PhilTa*, un *corpus* ristretto di testi greci e latini, che non comprende tutti quelli effettivamente disponibili *online* sotto licenza *open-source*. Questa scelta, da un lato, potrebbe limitare la capacità dell'attuale versione del modello di fornire correzioni di testi di epoche, stili, generi letterari (e dialetti, nel caso del greco antico) sottorappresentati e di termini rari o tecnici poco frequenti o assenti nel *corpus* di *Perseus Digital Library*, dall'altro è stata dettata dalla necessità di validare l'approccio adottato nello sviluppo di *Philo-1-preview*, prima di estenderlo a *dataset* più ampi. Si tenga comunque conto del fatto che il numero totale di *record* ottenuti si è rivelato sufficiente per sviluppare un modello dotato di prestazioni elevate nel *task* di interesse.

BIBLIOGRAFIA

Abraham, A., Bajaj, A., & Hanne, T. (A c. Di). (2024). Machine learning solutions. International Conference on Intelligent Systems Design and Applications, Cham. Springer.

Assael, Y., Sommerschield, T., & Prag, J. (2019). Restoring ancient text using deep learning: A case study on Greek epigraphy (arXiv:1910.06262). arXiv. http://arxiv.org/abs/1910.06262.

Assael, Y., Sommerschield, T., Shillingford, B., Bordbar, M., Pavlopoulos, J., Chatzipanagiotou, M., Androutsopoulos, I., Prag, J., & de Freitas, N. (2022). Restoring and attributing ancient texts using deep neural networks. Nature, 603(7900), 280–283. https://doi.org/10.1038/s41586-022-04448-z.

Bamman, D., & Burns, P. J. (2020). Latin BERT: A Contextual Language Model for Classical Philology (arXiv:2009.10053). arXiv. http://arxiv.org/abs/2009.10053.

Cowen-Breen, C., Brooks, C., Haubold, J., & Graziosi, B. (2023). Logion: Machine Learning for Greek Philology (arXiv:2305.01099). arXiv. https://doi.org/10.48550/arXiv.2305.01099.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (arXiv:1810.04805). arXiv. http://arxiv.org/abs/1810.04805.

Ferrara, G. (2023). Ianus 2.0. Corruttele filologiche, congetture digitali. La memoria digitale: forme del testo e organizzazione della conoscenza. Atti del XII Convegno Annuale AIUCD, 85–91. https://amsacta.unibo.it/id/eprint/7721/1/atti_aiucd_2023.pdf.

Gazit, L. (con un contributo di Ghaffari, M., & Saxena, A.). (2024). Mastering NLP from Foundations to LLMs: Apply advanced rule-based techniques to LLMs and solve real-world business problems using Python (1a ed.). Packt Publishing Limited.

Graziosi, B., Haubold, J., Cowen-Breen, C., & Brooks, C. (2023). Machine Learning and the Future of Philology: A Case Study. TAPA, 153(1), 253–284. https://doi.org/10.1353/apa.2023.a901022.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). Scaling Laws for Neural Language Models (arXiv:2001.08361). arXiv. https://doi.org/10.48550/arXiv.2001.08361.

Levine, L., Li, C., BremerMcCollum, L., Wagner, N., & Zeldes, A. (2024). Lacuna Language Learning: Leveraging RNNs for Ranked Text Completion in Digitized Coptic Manuscripts. Proceedings of the 1st Workshop on Machine Learning for Ancient Languages (ML4AL 2024), 61–70. https://doi.org/10.18653/v1/2024.ml4al-1.8.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2019). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension (arXiv:1910.13461). arXiv. https://doi.org/10.48550/arXiv.1910.13461.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2001). BLEU: A method for automatic evaluation of machine translation. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02, 311. https://doi.org/10.3115/1073083.1073135.

Radford, A., & Narasimhan, K. (2018). Improving Language Understanding by Generative Pre-Training. https://api.semanticscholar.org/CorpusID:49313245.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2019). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer (Versione 4). arXiv. https://doi.org/10.48550/ARXIV.1910.10683.

Rao, D., & McMahan, B. (2019). Natural language processing with PyTorch: Build intelligent language applications using deep learning (First edition). O'Reilly.

Riemenschneider, F., & Frank, A. (2023). Exploring Large Language Models for Classical Philology (arXiv:2305.13698). arXiv. http://arxiv.org/abs/2305.13698.

Shmidman, A., Shmidman, O., Gershuni, H., & Koppel, M. (2024). MsBERT: A New Model for the Reconstruction of Lacunae in Hebrew Manuscripts. Proceedings of the 1st Workshop on Machine Learning for Ancient Languages (ML4AL 2024), 13–18. https://doi.org/10.18653/v1/2024.ml4al-1.2.

Singh, P., Rutten, G., & Lefever, E. (2021). A Pilot Study for BERT Language Modelling and Morphological Analysis for Ancient and Medieval Greek. Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, 128–137. https://doi.org/10.18653/v1/2021.latechclfl-1.15.

"Glottolab: A Linguistic Adventure": Lo sviluppo di un'attività gamificata incentrata sulla linguistica

Cecilia Cattaneo[1], Claudia Roberta Combei[2], Chiara Zanchi[3]
[1] Università di Pavia, Italia - cecilia.cattaneo99@gmail.com
[2] Università degli Studi di Roma "Tor Vergata", Italia - claudia.roberta.combei@uniroma2.it
[3] Università di Pavia, Italia - chiara.zanchi01@unipv.it

ABSTRACT (ITALIANO)

"Glottolab: A Linguistic Adventure" è un'attività di tipo *game-based learning*, sviluppata con RPG Maker MZ, allo scopo di avvicinare un pubblico di non specialisti alla linguistica in modo interattivo e coinvolgente. Il gioco introduce ai partecipanti alcuni concetti di linguistica generale, favorendo una riflessione sulla struttura interna del linguaggio e sugli usi concreti della lingua. L'obiettivo è promuovere una maggiore consapevolezza metalinguistica e superare l'approccio prescrittivo spesso adottato nell'insegnamento delle materie linguistiche nelle scuole secondarie italiane. Nel gioco, i partecipanti esplorano cinque regni, ciascuno rappresentativo di un livello di analisi linguistica. In ogni regno, devono risolvere *puzzle* basati su problemi linguistici, raccogliendo progressivamente gli elementi necessari per ricostruire la lingua perduta della città di Logos. Per valutare l'efficacia di Glottolab, si è ritenuto opportuno procedere con un'analisi quantitativa e qualitativa dell'esperienza di gioco. Nello specifico, 16 partecipanti hanno preso parte a sessioni di gioco, al termine delle quali è stato somministrato un questionario finalizzato a misurare quattro aspetti: esperienza di apprendimento, usabilità, esperienza di gioco ed efficacia dell'integrazione di sistemi di intelligenza artificiale. Si è poi scelto di osservare in tempo reale l'esperienza di tre giocatori attraverso l'analisi qualitativa delle registrazioni delle loro giocate. Il *feedback* ricevuto indica che *Glottolab* ha il potenziale per costituire uno strumento efficace per la divulgazione scientifica della linguistica, sebbene alcuni elementi e meccaniche di gioco necessitino di ulteriori miglioramenti. Complessivamente, l'esperienza di gioco è stata giudicata coinvolgente e stimolante, mentre l'integrazione di sistemi di intelligenza artificiale non è risultata convincente.
Parole chiave: *Game-based learning*; linguistica; consapevolezza metalinguistica; esperienza di gioco; intelligenza artificiale.

ABSTRACT (ENGLISH)

"Glottolab: A Linguistic Adventure." Designing a game-based learning activity on linguistics.

"Glottolab: A Linguistic Adventure" is a game-based learning activity developed with RPG Maker MZ, designed to familiarize a non-specialized audience with linguistics in an interactive and engaging way. The game introduces participants to concepts of general linguistics, encouraging reflection on the internal structure of language and its practical uses. The aim is to foster greater metalinguistic awareness and to challenge the prescriptive approach often adopted in the language teaching in secondary school education in Italy. In the game, participants explore five realms, each representing a level of linguistic analysis. In each realm, they are asked to solve puzzles based on linguistic problems, gradually collecting the elements needed to reconstruct the lost language of the city of Logos. To evaluate the effectiveness of the activity, both quantitative and qualitative analyses have been conducted. In fact, 16 participants took part in gameplay sessions, after which they completed a questionnaire designed to assess four aspects: learning experience, usability, gameplay experience, and the integration of artificial intelligence systems. Additionally, the real-time gameplay of three participants was recorded and analyzed qualitatively. The *feedback* suggests that *Glottolab* has the potential to serve as an effective tool for scientific dissemination in linguistics, although certain elements and game mechanics require further refinement. Overall, the gameplay experience was considered engaging and stimulating, while the integration of artificial intelligence systems was found to be less relevant for this type of activity.
Keywords: Game-based learning, linguistics, metalinguistic awareness, user experience, artificial intelligence.

1. INTRODUZIONE

Negli ultimi anni, gli studiosi del linguaggio hanno sottolineato la necessità di sviluppare strategie innovative per comunicare i risultati delle loro ricerche a un pubblico di non specialisti, con l'obiettivo di far conoscere la disciplina e/o renderla più accessibile al di fuori dei contesti accademici (Pannitto et al., 2021). Il presente studio va in questa direzione, proponendo una strategia innovativa, volta a favorire una

maggiore diffusione della linguistica e a incrementare la consapevolezza del pubblico rispetto ai temi trattati da questa disciplina.
Diversi metodi, strumenti e canali sono stati usati per comunicare la linguistica: siti web (es. Linguisticamente[1]), blog (es., Language Log[2]), podcast (es., Lingthusiasm[3]), programmi radiofonici e televisivi (es. Because Language[4]), account sui social media (es. Linguisticattiva[5]), e giochi educativi (es. Linguistics Puzzles[6]). Si è notato come le scienze del linguaggio sembrino trarre benefici dall'uso di esperienze ludiche, soprattutto per la raccolta di dati linguistici, l'apprendimento di una seconda lingua e l'esplorazione di applicazioni pratiche della linguistica (Bonetti & Tonelli, 2021). Tuttavia, nonostante la letteratura riporti diversi tentativi di gamificare *task* di annotazione linguistica e di sviluppare applicazioni ludiche per l'apprendimento linguistico, gli esempi di giochi esplicitamente mirati all'insegnamento diretto di nozioni di linguistica generale rimangono limitati.
Per questo motivo, è stato sviluppato *"Glottolab: A Linguistic Adventure",* un'attività di tipo *game-based learning* in lingua inglese dedicata alle scienze del linguaggio, con lo scopo di introdurre alcuni concetti di linguistica generale e di suscitare curiosità e interesse verso il mondo del linguaggio tra i non esperti. In questo lavoro si descrivono le fasi di progettazione, sviluppo e valutazione di questa attività e, allo stesso tempo, si evidenzia il potenziale didattico di uno strumento come questo per avvicinare un pubblico più ampio alle scienze del linguaggio.

2. QUADRO TEORICO

La divulgazione scientifica della linguistica in contesti diversi da quello accademico richiede l'adozione di strategie efficaci nel rendere i contenuti specialistici accessibili e rilevanti per un pubblico più ampio. Price & McIntyre (2023) suggeriscono che sia necessario aumentare la consapevolezza sul valore della linguistica e di dimostrare il suo potenziale ruolo nella vita quotidiana di tutti noi. Alcuni esempi riguardano la promozione di scelte linguistiche più consapevoli, la valorizzazione della diversità linguistica, la necessità di contrastare i pregiudizi e i falsi miti sul funzionamento delle lingue e del linguaggio, il supporto all'elaborazione di politiche linguistiche efficaci e la possibilità di offrire prospettive informate sul trattamento dei disturbi del linguaggio (Grandi & Masini, 2020). Per raggiungere questi obiettivi, è importante che i linguisti capiscano gli interessi, le curiosità e i dubbi del grande pubblico e creino programmi di divulgazione più mirati ed efficaci (McCulloch & Gawne, 2024).
I giochi digitali possono essere una strategia efficace per avvicinare un pubblico non esperto alla linguistica. Negli ultimi 15 anni, essi sono stati impiegati sempre più frequentemente anche in contesti didattici perché permettono di migliorare l'apprendimento linguistico, consentendo facilità di ripetizione, richiamo informale di concetti e lo sviluppo di diverse competenze trasversali (Ishaq et al., 2022; Miller et al., 2019).
La letteratura in questo ambito mette in luce numerosi vantaggi, tra cui un incremento della motivazione, un miglioramento delle performance didattiche, un maggiore coinvolgimento e una più attiva partecipazione. Vengono inoltre segnalate opportunità di collaborazione, la possibilità di ricevere *feedback* immediati e di monitorare in modo continuo i progressi. Gli studi sul tema offrono anche alcuni consigli per implementare strategie di questo tipo: definire obiettivi didattici chiari, progettare meccaniche di gioco attraverso la costruzione di storytelling avvincenti e *task* interattivi, garantire personalizzazione, collaborazione e interazione sociale (Wuluntari et al., 2023).
I linguisti hanno così notato che questo tipo di gioco può essere utile per sviluppare risorse linguistiche, osservare il comportamento linguistico all'interno di una comunità, studiare teorie formali e favorire l'apprendimento di una lingua seconda (Bonetti & Tonelli, 2020; Shortt et al., 2021). Le *Mobile Language Learning Applications* (MALL), come *Duolingo* e *Babbel* (Castañeda & Cho, 2016), e i *Games With A Purpose* (GWAP), come *Phrase Detectives* (Poesio et al., 2013) e *Word Ladders* (Bolognesi et al., 2024),

[1] Il sito web è accessibile al seguente link: https://www.linguisticamente.org (cons. 25/01/2025)

[2] Il blog è accessibile al seguente link: https://languagelog.ldc.upenn.edu/nll/ (cons. 25/01/2025)

[3] Il podcast è accessibile al seguente link: https://open.spotify.com/show/4IfWLwqURo177w2i4Ecj7t (cons. 25/01/2025)

[4] Il programma è accessibile al seguente link: https://becauselanguage.com (cons. 25/01/2025)

[5] Il profilo social media è accessibile al seguente link: https://www.instagram.com/linguisticattiva/ (cons. 25/01/2025)

[6] I giochi sono accessibili al seguente link: https://sites.google.com/princeton.edu/nacloprinceton/linguistic-puzzles (cons. 25/01/2025)

sono ormai ampiamente diffusi. Ad oggi, però, sono disponibili solo pochi esempi di giochi progettati per avvicinare in modo implicito o esplicito il pubblico non esperto alla linguistica (Medina et al., 2019). Negli ultimi anni l'intelligenza artificiale si è affermata come un tipo di supporto promettente per migliorare l'esperienza di gioco in contesti didattici (Chen & Chang, 2024). I nuovi approcci didattici basati sull'intelligenza artificiale consentono di creare contenuti personalizzati basati su bisogni di apprendimento specifici, fornire *feedback* in tempo reale e rispondere ai dubbi del giocatore, oltre che monitorarne la progressione all'interno del gioco (Tay & Cheah, 2021). I dialoghi, le trame e gli scenari virtuali possono, quindi, essere generati automaticamente attraverso l'intelligenza artificiale.

3. METODOLOGIA

Glottolab: A Linguistic Adventure è un'attività di tipo *game-based learning* costruita per stimolare l'interesse e la curiosità di un pubblico non esperto verso il mondo della linguistica. È stata progettata con RPG Maker MZ, un software *low-code* che permette di creare videogiochi di ruolo in due dimensioni.[7]

3.1. DESIGN DEL GIOCO

Glottolab è un *adventure game* che può essere classificato come esempio di *game-based learning*, una strategia didattica che ha l'obiettivo di far apprendere diverse nozioni e concetti teorici attraverso attività ludiche (Prensky, 2001). L'idea è che, attraverso il gioco, sia possibile acquisire, rafforzare o ampliare conoscenze specifiche (Bhandari et al., 2019). Nel nostro gioco, l'esperienza di apprendimento è stata definita combinando elementi dell'apprendimento esperienziale e contestuale e decostruendo le attività didattiche proposte in *puzzle game*, mentre le componenti della gamificazione sono state selezionate seguendo il modello di *game design* proposto da Shi & Shih (2015).
Glottolab è stato pensato come un *adventure quest*. La trama del gioco ruota attorno alla scomparsa della lingua degli abitanti di Logos, una città immaginaria. I giocatori devono guidare il protagonista attraverso cinque regni linguistici, risolvere cinque *puzzle game* con domande di linguistica, raccogliere le diverse componenti del linguaggio e ricostruire la lingua di Logos. Ad oggi, il gioco è fruibile in lingua inglese. Tale scelta è giustificata dall'interesse di raggiungere un più ampio pubblico possibile.
Il gioco è strutturato come una mappa interattiva (vd. Fig.1) che il giocatore esplora visitando cinque regni, ognuno corrispondente a un livello di analisi linguistica: fonetica e fonologia (regno dei suoni), morfologia (regno delle parole), sintassi (regno delle frasi), semantica (regno del significato) e pragmatica (regno del discorso).



Figura 1. La mappa del gioco con la città di Logos e i cinque regni linguistici.

Le meccaniche del gioco sono state definite per facilitare la progressione dei giocatori negli ambienti virtuali e ottimizzare l'esperienza didattica. Il gioco contiene tre tipi di meccaniche diverse: *task mechanics*, *feedback mechanics* e *interaction mechanics*. Per quanto riguarda le *task mechanics*, sono stati stabiliti un *task* principale (la ricostruzione del linguaggio) e una serie di *task* secondari che permettono di raggiungere l'obiettivo risolvendo sei quiz linguistici per ogni regno. I *puzzle game* sono basati su concetti di linguistica generale e presentano ciascuno un problema linguistico: identificazione delle coppie minime

[7] Il gioco è accessibile al seguente link: https://glottolab.z6.web.core.windows.net/ (cons. 22/01/2025)

per la fonetica e la fonologia nella *Echoing Cave*, meccanismi di formazione di parola per la morfologia nel *Words Weaver Lab* (vd. Fig.2), risoluzione di test di costituenza per la sintassi nella *Enchanted Forest*, riconoscimento di relazioni di significato per la semantica nel *Mirage Desert* e individuazione di fenomeni pragmatici nello *Speech Bazar*. Il sistema di ricompensa consiste in una serie di forzieri che il giocatore può sbloccare una volta completato ogni livello. Sbloccando tutti i forzieri, il giocatore raccoglie diversi elementi che permettono di far riacquisire la facoltà del linguaggio agli abitanti di Logos. Le *feedback mechanics* forniscono all'utente un *juicy feedback*, attraverso l'uso di suoni e componenti grafiche, oppure un *feedback* cognitivo, segnalando quanto positivamente l'utente stia procedendo nel gioco, offrendo commenti positivi ("Correct!") o negativi ("Not correct! Try again"). Infine, le *interaction mechanics* sono svolte da NPC (Non-Playable-Characters) che dialogano con il giocatore e forniscono informazioni utili per la progressione del gioco.



Figura 2. Un esempio di quiz nel *puzzle* game del regno delle parole.

3.2. SISTEMA DI DIALOGHI DEL GIOCO

Il sistema di dialoghi del gioco è creato utilizzando un approccio ibrido, combinando *dialogue trees* e dialoghi generati con l'intelligenza artificiale. I *dialogue trees* sono costruiti attraverso un sistema di dialoghi a turni, in cui i giocatori devono selezionare una tra le diverse opzioni fornite dagli sviluppatori, al fine di progredire nella conversazione e nel gioco, seguendo però un percorso predeterminato (Brabra et al., 2021).

I dialoghi generati con l'intelligenza artificiale hanno l'obiettivo di evitare interazioni ripetitive e aggiungere diversità e dinamicità all'interno della giocata. Una serie di NPC interagisce con i giocatori generando un testo simile a quello prodotto da esseri umani e impersonando vari ruoli come quelli di servitori del tempio, abitanti di Logos, le cui modalità di espressione presentano dei disturbi del linguaggio, e famosi linguisti, il cui compito è fornire spiegazioni su alcuni concetti di linguistica generale (vd. Fig.3).

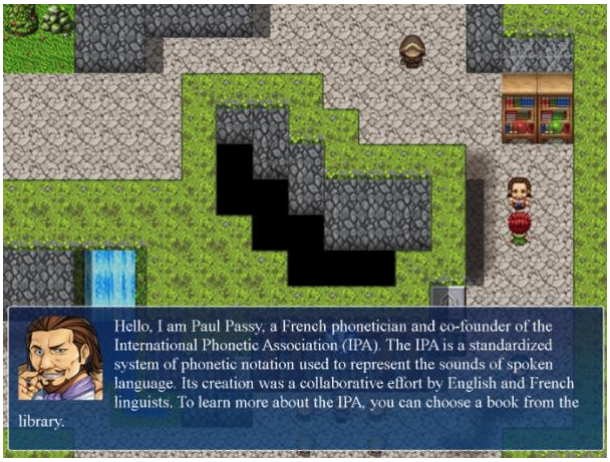


Figura 3. Un esempio di un dialogo generato con ChatGPT API.

Per ottenere questo risultato, sono stati preparati alcuni *prompt* come input per permettere a un *Large Language Model* di generare un dialogo personalizzato in tempo reale. Le interazioni sono generate automaticamente tramite l'integrazione di *ChatGPT API*, incorporato nel gioco utilizzando il plugin *ChatGPT_APIMZ*.[8] Per quanto riguarda le tecniche di *prompt engineering*, per la generazione è stato usato principalmente lo *zero-shot prompting*, mentre solo occasionalmente il *few-shot prompting*. Inoltre, si è prestata particolare attenzione a garantire l'inclusività adottando un linguaggio neutrale e rispettoso per la scrittura dei *prompt*.

3.3. PROCEDURA SPERIMENTALE

Al termine dello sviluppo del gioco, è stata condotta una procedura di valutazione sperimentale articolata in due fasi. Sono state organizzate alcune sessioni di gioco alle quali hanno preso parte 16 giocatori. Si è scelto di prendere in considerazione un campione variegato per abitudini di gioco, età, livello di istruzione, conoscenza della lingua inglese e familiarità con la linguistica generale. L'età media dei partecipanti è stata di 29 anni e 10 mesi; nove partecipanti hanno conseguito una laurea triennale, cinque una laurea magistrale e due un diploma di scuola superiore; la maggior parte dei partecipanti (sette) dichiara un livello di conoscenza della lingua inglese pari a B2, sei un livello pari a C1, tre un livello pari ad A2; sei partecipanti sono studenti di linguistica, mentre dieci non hanno intrapreso un percorso di studi in cui fosse previsto l'insegnamento di questa disciplina; infine, dieci partecipanti dichiarano di avere familiarità con i videogiochi. Dopo la sessione di gioco, è stato chiesto ai partecipanti di compilare un questionario finalizzato a valutare l'efficacia didattica dell'attività proposta.[9] Il questionario è stato strutturato in sezioni basate su quattro domande di ricerca: Q1: Il gioco supporta l'apprendimento della linguistica con attività interattive? Q2: Gli elementi e le meccaniche di gioco sono progettati in modo efficace? Q3: L'esperienza dell'utente favorisce coinvolgimento e motivazione a completare il gioco e a imparare nozioni nuove? Q4: L'integrazione di sistemi di intelligenza artificiale offre un'esperienza migliore in termini di funzionalità e opportunità per i giocatori?

4. RISULTATI E DISCUSSIONE

Di seguito verrà presentato e discusso il *feedback* ricevuto dai giocatori, attraverso l'analisi dei questionari e delle giocate.

4.1. ANALISI DEI RISULTATI DEL QUESTIONARIO

L'analisi dei risultati del questionario fornisce una risposta alle quattro domande di ricerca. Complessivamente, il *feedback* sull'acquisizione di nuove conoscenze è stato molto positivo. Nello specifico, tutti i partecipanti senza una preparazione specifica nell'ambito della linguistica hanno dichiarato di aver imparato concetti nuovi durante le giocate e hanno manifestato un incremento della loro curiosità verso la linguistica al termine della sessione. In generale, i risultati indicano quindi che l'esperienza di apprendimento è stata efficace soprattutto per coloro che si approcciavano per la prima volta alla disciplina, e l'attività proposta rappresenta un'introduzione accessibile ai concetti di base della linguistica generale. La percezione dell'usabilità del gioco è risultata inferiore alle aspettative, evidenziando la necessità di apportare alcuni miglioramenti. In primo luogo, nove partecipanti (56,25%) hanno considerato il gioco poco intuitivo, undici (68,75%) hanno giudicato poco chiare le istruzioni e i comandi, mentre otto (50%) hanno segnalato che i *task* proposti e la progressione dei livelli non erano progettati adeguatamente. L'analisi dei risultati ha rilevato che nonostante il gioco non abbia superato il test di usabilità, i partecipanti erano in generale soddisfatti dell'esperienza di gioco. In particolare, quasi tutti i partecipanti (93,75%) hanno trovato il gioco coinvolgente, hanno apprezzato le modalità di esplorazione dei diversi regni linguistici e hanno ritenuto che la trama fosse avvincente.
Osservando i risultati del test di usabilità e dell'esperienza di gioco si è deciso di operare un'analisi tematica dei *feedback* dei partecipanti al fine di comprendere meglio le loro opinioni ed esperienze. Sono stati adottati due metodi: estrazione delle parole chiave con la libreria NLTK[10] e *topic modeling* usando il metodo *Latent Dirichlet Allocation* (LDA) proposto in Ugorji et al. (2024). Di seguito, è stata compilata una lista di 10 macrotemi. T1: i comandi e le istruzioni non sono sempre chiari. T2: il percorso da seguire è

[8] Il plugin di ChatGPTAPI è accessibile al seguente link: https://bit.ly/plugins-ChatGPT-APIMZ (cons. 22/01/2025)

[9] Il questionario è disponibile al seguente link: https://bit.ly/glottolab-survey (cons. 22/01/2025)

[10] Il link per la libreria NLTK: https://www.nltk.org/ (cons. 25/01/2025)

incerto. T3: il gioco è troppo lungo. T4: i testi a volte sono troppo prolissi. T5: la difficoltà dei vari livelli non è uniforme. T6: alcune parti sono difficili da comprendere. T7: alcuni concetti linguistici necessitano di spiegazioni più approfondite. T8: le interazioni con i linguisti famosi sono percepite come scollegate dal resto del gioco. T9: dovrebbe essere incluso un sistema di *switch*. T10: i sistemi di progressione e di punteggio non sono chiari.
Il questionario includeva anche domande volte a valutare la capacità dei giocatori di riconoscere l'utilizzo di sistemi di intelligenza artificiale all'interno del gioco e la loro percezione su questo tipo di implementazione. I risultati indicano che solo il 50% dei partecipanti ha capito che alcuni dialoghi, voci e *task* erano stati generati tramite l'intelligenza artificiale. Alcuni giocatori hanno riferito che quei dialoghi sono stati percepiti come generici, ripetitivi, talvolta poco accurati dal punto di vista concettuale, poco chiari, inespressivi e non sempre coerenti con il contesto. Le voci ottenute attraverso la sintesi vocale sono state descritte come innaturali, robotiche, eccessivamente perfette, con un'intonazione piatta.

4.2. ANALISI QUALITATIVA DELLE REGISTRAZIONI DELLE GIOCATE

Tre partecipanti hanno dato il loro consenso a registrare l'intera sessione di gioco, rendendo possibili ulteriori valutazioni qualitative. I tre giocatori sono stati selezionati per massimizzare la diversità di esperienze formative. Per questo esperimento sono stati scelti una studentessa di linguistica, una giocatrice senza una formazione specifica nella disciplina e un grande appassionato di videogiochi.
I giocatori hanno completato il gioco in tempi diversi, suggerendo che il livello di competenza di inglese e la familiarità con i videogiochi potrebbero avere un effetto sul tempo di gioco e sull'efficacia didattica dell'attività. Tutti i giocatori hanno esplorato i cinque regni, interagito con tutti gli NPC e completato tutti i *puzzle game* proposti. Tuttavia, i loro percorsi rivelano differenze significative. La studentessa di linguistica ha ottenuto risultati migliori rispetto agli altri due partecipanti in tutti i regni, tranne nel regno delle parole. Il giocatore appassionato di videogiochi ha incontrato difficoltà nel regno dei suoni a causa della complessità nell'identificare la parola corretta all'interno di una coppia minima. La giocatrice senza una formazione specifica in materia, invece, ha ottenuto risultati sorprendentemente positivi nel regno delle parole, ma ha riscontrato problemi nel regno del significato e nel regno del discorso.

5. CONCLUSIONE

Questo contributo ha presentato la progettazione, lo sviluppo e la valutazione di "Glottolab: A Linguistic Adventure", un gioco volto ad avvicinare il largo pubblico alle scienze del linguaggio e a promuovere la divulgazione scientifica della linguistica. Il lavoro ha dimostrato la validità dei metodi di tipo *game-based learning* per scopi divulgativi e didattici, evidenziando, però, l'importanza di valutare l'efficacia didattica sulla base della formazione dei giocatori. Complessivamente, l'attività è stata valutata positivamente dai partecipanti. Tuttavia, questi hanno segnalato anche alcuni limiti rispetto all'organizzazione di alcune meccaniche ed elementi di gioco e all'integrazione di sistemi di intelligenza artificiale che risultano poco funzionali e i cui output, spesso imprevedibili, possono compromettere la logica e la coerenza narrativa del gioco. Le prospettive future di questo lavoro prevedono quanto segue: il miglioramento di alcune dinamiche di gioco per favorire una maggiore usabilità e coinvolgimento dei giocatori, l'ampliamento del numero di partecipanti alla fase di valutazione, al fine di ottenere risultati più robusti e l'utilizzo di metriche oggettive per la valutazione dell'esperienza di apprendimento e dell'integrazione di sistemi di intelligenza artificiale, con particolare attenzione all'analisi della qualità degli output prodotti dai Large Language Model inseriti nel gioco.
Al termine delle fasi di valutazione e revisione, si ritiene che il gioco possa diventare uno strumento utile per organizzare programmi interattivi di divulgazione scientifica, rivolti soprattutto a studenti della scuola secondaria di secondo grado con un livello di inglese intermedio o avanzato. L'obiettivo è quello di suscitare il loro interesse per il mondo della linguistica e delle scienze del linguaggio, stimolando la curiosità verso le questioni teoriche e pratiche poste dalla disciplina e offrendo così un primo contatto con essa prima dell'eventuale ingresso all'università. Inoltre, il gioco può rappresentare un punto di partenza per arricchire la ricerca sperimentale nell'ambito della linguistica applicata e computazionale.

BIBLIOGRAFIA

Bhandari, S., Hallowell, M. R., Correll, J. (2019) Making construction safety training interesting: A field-based quasi-experiment to test the relationship between emotional arousal and situational interest among adult learners, Safety Science, 117, 58-70, https://doi.org/10.1016/j.ssci.2019.03.028.

Bolognesi, M. M., Collacciani, C., Ferrari, A., Genovese, F., Lamarra, T., Loia, A., Rambelli, G., Ravelli, A. A., Villani, C. (2024). Word Ladders: A mobile application for semantic data collection. arXiv. https://doi.org/10.48550/arxiv.2404.00184.

Bonetti, F., & Tonelli, S. (2020). A 3D role-playing game for abusive language annotation. Workshop on Games and Natural Language Processing, 39–43. European Language Resources Association.

Bonetti, F., & Tonelli, S. (2021). Challenges in designing games with a purpose for abusive language annotation. Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing, 60–65. Association for Computational Linguistics.

Brabra, H., Baez, M., Benatallah, B., Gaaloul, W., Bouguelia, S., & Zamanirad, S. (2021). Dialogue management in conversational systems: A review of approaches, challenges, and opportunities. IEEE Transactions on Cognitive and Developmental Systems, 1(1), 1–1.

Castañeda, D. A., & Cho, M. H. (2016). Use of a game-like application on a mobile device to improve accuracy in conjugating Spanish verbs. Computer Assisted Language Learning, 29(7), 1195–1204. https://doi.org/10.1080/09588221.2016.1197950.

Chen, C. H., & Chang, C. L. (2024). Effectiveness of AI-assisted game-based learning on science learning outcomes, intrinsic motivation, cognitive load, and learning behavior. Education and Information Technologies, 29, 18621–18642. https://doi.org/10.1007/s10639-024-12553-x.

Grandi, N., & Masini, F. (2020). La linguistica della divulgazione, la divulgazione della linguistica. In Atti del IV Convegno Interannuale SLI nuova serie (Bologna, 14-15 giugno 2018).

Ishaq, K., Rosdi, F., Zin, N. A. M., & Abid, A. (2022). Serious game design model for language learning in the cultural context. Educational Information Technology, 27, 9317–9355. https://doi.org/10.1007/s10639-022-10999-5.

McCulloch, G., & Gawne, L. (2024). Towards a theory of linguistic curiosity: Applying linguistic frameworks to lingcomm and scicomm. Linguistics Vanguard. https://doi.org/10.1515/lingvan-2024-0073.

Medina, D., Maturana Gonzalez, G., & Villa Garzon, F. (2019). A prototype for a serious digital game to teach linguistic ontologies. arXiv. https://doi.org/10.48550/arXiv.1909.07371.

Miller, J., Narayan, U., Hantsbarger, M., Cooper, S., & El-Nasr, M. (2019). Expertise and engagement: Re-designing citizen science games with players' minds in mind. FDG: Proceedings of the International Conference on Foundations of Digital Games, 1–11. https://doi.org/10.1145/3337722.3337735.

Pannitto, L., Busso, L., Combei, C. R., Messina, L., Miaschi, A., Sarti, G., & Nissim, M. (2021). Teaching NLP with Bracelets and Restaurant Menus: An Interactive Workshop for Italian Students. Proceedings of the Fifth Workshop on Teaching Natural Language Processing (NLP) at NAACL-HLT, Mexico City 10–11 June 2021. Jurgens, D. et al. (Eds.). 160-170. https://aclanthology.org/2021.teachingnlp-1.26/.

Poesio, M., Chamberlain, J., Kruschwitz, U., Robaldo, L., & Ducceschi, L. (2013). Phrase detectives: Utilizing collective intelligence for internet-scale language resource creation. ACM Transactions on Interactive Intelligent Systems, 3(1), Article 3. https://doi.org/10.1145/2448116.2448119.

Prensky, M. (2001). Digital game-based learning. McGraw-Hill.

Price, H., & McIntyre, D. (Eds.). (2023). Communicating linguistics: Language, community and public engagement (1st ed.). Routledge. https://doi.org/10.4324/9781003096078.

Shi, Y., & Shih, J. (2015). Game factors and game-based learning design model. International Journal of Computer Games Technology, 2015.

Shortt, M., Tilak, S., Kuznetcova, I., Martens, B., & Akinkuolie, B. (2021). Gamification in mobile-assisted language learning: A systematic review of Duolingo literature from public release of 2012 to early 2020. Computer Assisted Language Learning, 36. https://doi.org/10.1080/09588221.2021.1933540.

Tan, D. Y., & Cheah, C. W. (2021). Developing a gamified AI-enabled online learning application to improve students' perception of university physics. Computers and Education: Artificial Intelligence, 2, 100032. https://doi.org/10.1016/j.caeai.2021.100032.

Ugorji, C. C., Onyesolu, M., Asogwa, D., & Egwu, C. (2024). Exploring Latent Dirichlet Allocation (LDA) in topic modeling: Theory, applications, and future directions. Newport International Journal of Engineering and Physical Sciences, 4, 9–16. https://doi.org/10.59298/NIJEP/2024/41916.1.1100.

Wulantari, N., Rachman, A., Sari, M., Uktolseja, L., Ikip, A., Tabanan, S., Pahlawan, J., Peken, D., Tabanan, K., Tabanan, K., Oleo, U., Hijau, K., Tridharma, B., Anduonohu, K., Kambu, K., Kendari, S.

T., Muhammadiyah, S., Penuh, S., & Majalengka, J. (2024). The Role of Gamification in English Language Teaching: A Literature Review. *Journal on Education*, 6, 2847–2856. https://doi.org/10.31004/joe.v6i1.3328.

Concordanze e NLP: idee, metodi e regole per l'applicazione alla lingua italiana

Pietro Sichera[1], Christian D'Agata[2], Giuseppe Palazzolo[3]
[1]CNR ILIESI, Italia pietro.sichera@cnr.it
[2]Università degli Studi di Catania, christian.dagata@unict.it
[3]Università degli Studi di Catania, giuseppe.palazzolo@unict.it

ABSTRACT (ITALIANO)

La lemmatizzazione automatica dei testi rappresenta uno strumento fondamentale nell'elaborazione del linguaggio naturale (NLP), in quanto consente di associare le occorrenze testuali ai loro lemmi, ovvero alle voci di vocabolario da cui derivano. Questo processo, essenziale per l'analisi linguistica e semantica, presenta tuttavia sfide significative quando applicato alla lingua italiana. In particolare, le peculiarità morfologiche della lingua, come la presenza di enclitiche e omografi, rendono necessaria l'adozione di approcci che coniughino tecnologie avanzate e interventi di disambiguazione personalizzati.
In questo lavoro si analizza l'uso combinato di spaCy e UDPipe, per la lemmatizzazione automatica di testi italiani, evidenziandone i punti di forza e le limitazioni. Tali strumenti sono integrati con il software LiotroConcord_v2, sviluppato su piattaforma 4D, implementando regole personalizzate per la gestione di casi complessi. Questo studio si propone di discutere l'importanza di un approccio ibrido alla lemmatizzazione, che integri metodi automatici e correttivi manuali, al fine di migliorare la qualità dell'analisi linguistica nel contesto delle edizioni digitali e della ricerca filologica. In particolare, vengono approfonditi i contributi metodologici e tecnologici del sistema sviluppato, con uno sguardo alle potenziali applicazioni future e alla possibilità di estendere tali soluzioni a ulteriori ambiti di ricerca. Nella seconda parte viene presentato un caso di studio specifico relativo alla codifica e alla lemmatizzazione de I Viceré.
Parole chiave: lemmatization; concordance; NLP; model; digital scientific editions.

ABSTRACT (ENGLISH)

Concordances and NLP: ideas, methods and rules for application to the Italian language
Automatic text lemmatization represents a fundamental tool in natural language processing (NLP), as it allows textual occurrences to be associated with their lemmas, i.e., the vocabulary items from which they are derived. However, this process, which is essential for linguistic and semantic analysis, presents significant challenges when applied to the Italian language. In particular, the morphological peculiarities of the language, such as the presence of enclitics and homographs, make it necessary to adopt approaches that combine advanced technologies and customized disambiguation interventions. This paper analyzes the combined use of spaCy and UDPipe, for automatic lemmatization of Italian texts, highlighting their strengths and limitations. These tools are integrated with LiotroConcord_v2 software, developed on a 4D platform, implementing custom rules for handling complex cases. This study aims to discuss the importance of a hybrid approach to lemmatization, integrating automatic methods and manual corrections, in order to improve the quality of linguistic analysis in the context of digital editions and philological research. In particular, the methodological and technological contributions of the developed system are discussed in depth, with a look at potential future applications and the possibility of extending such solutions to further areas of research. In the second part, a specific case study related to the encoding and lemmatization of I Viceré is presented.
Keywords: lemmatization; concordance; NLP; model; digital scientific editions

1. INTRODUZIONE

La lemmatizzazione automatica dei testi rappresenta uno dei pilastri fondamentali nell'elaborazione del linguaggio naturale (Natural Language Processing, NLP), in quanto consente di collegare ogni occorrenza testuale al suo lemma, ovvero alla forma di base da cui deriva. Questo processo risulta particolarmente importante nel contesto dell'umanistica digitale, dove l'analisi dei testi richiede una comprensione accurata delle strutture linguistiche, al fine di creare strumenti affidabili per la ricerca filologica, la costruzione di edizioni digitali e l'interrogazione semantica dei corpora.
La lingua italiana, tuttavia, presenta una serie di peculiarità che rendono la lemmatizzazione un compito particolarmente complesso. Tra queste si annoverano la presenza di enclitiche, che si legano a forme verbali generando token composti, e gli omografi, parole con la stessa forma grafica ma significati e categorie grammaticali differenti. Questi fenomeni richiedono soluzioni avanzate che vadano oltre la

semplice associazione tra forma e lemma, includendo anche la disambiguazione contestuale e la segmentazione delle forme complesse.
La difficoltà di questi problemi è ulteriormente amplificata quando le variazioni linguistiche, le forme arcaiche e la mancanza di standardizzazione rendono inapplicabili molti strumenti di NLP progettati per il linguaggio contemporaneo. Pertanto, l'adozione di sistemi combinati, che uniscano le capacità di diverse tecnologie NLP, può rappresentare una soluzione efficace per affrontare tali complessità.
Questo articolo si concentra sull'analisi e l'integrazione di due tra i più avanzati strumenti di lemmatizzazione automatica: spaCy[1] e UDPipe.[2] Entrambi i sistemi sono stati scelti per la loro capacità di gestire testi italiani con un alto grado di accuratezza e per la possibilità di personalizzare le pipeline di analisi linguistica. Tuttavia, per affrontare le limitazioni intrinseche di ciascun sistema, è stato sviluppato un approccio ibrido che combina i risultati di spaCy e UDPipe con interventi di disambiguazione e segmentazione realizzati tramite il software LiotroConcord_v2, una piattaforma progettata per integrare l'automazione e il controllo manuale nell'elaborazione dei testi.
L'obiettivo del presente lavoro è duplice: da un lato, evidenziare le sfide specifiche della lemmatizzazione automatica per la lingua italiana e, dall'altro, presentare un metodo pratico che sfrutti l'integrazione di tecnologie avanzate per migliorare l'affidabilità dell'analisi linguistica. In particolare, si pone l'accento sull'importanza di approcci personalizzabili per il trattamento delle enclitiche, degli omografi e delle forme linguistiche non standard, contribuendo a una migliore comprensione e accessibilità dei testi nel contesto dell'umanistica digitale.

2. STATO DELL'ARTE

Storicamente, la lemmatizzazione è stata affrontata tramite due approcci principali: basati su regole e basati su modelli statistici. Mentre i sistemi tradizionali si affidavano a dizionari e regole morfologiche, le tecnologie contemporanee sfruttano modelli di apprendimento supervisionato e reti neurali per elaborare testi con maggiore accuratezza e flessibilità.
La lingua italiana presenta peculiarità che mettono alla prova le capacità dei lemmatizzatori. Tra queste, si evidenziano le enclitiche e le forme composte, come nel caso di parole quali "trovandotelo", che contengono forme verbali con pronomi oggetto e indiretti annessi, richiedendo una segmentazione corretta per individuare i lemmi "trovare", "te" e "lo". Un ulteriore ostacolo è costituito dagli omografi, ovvero termini come "calcio", che possono avere significati e categorie grammaticali diverse a seconda del contesto, come sostantivo o verbo. Le variazioni diacroniche e arcaiche presenti nei testi letterari e storici, ad esempio "vo" invece di "vado", complicano ulteriormente l'analisi automatica. Sebbene i modelli moderni come quelli di spaCy e UDPipe riescano a gestire alcune di queste problematiche attraverso reti neurali e addestramenti su corpora diversificati, non sono privi di limitazioni. Ad esempio, i modelli pre-addestrati per l'italiano spesso non distinguono correttamente tra parole contenenti enclitiche o forme non standard, richiedendo interventi manuali o l'uso di regole personalizzate.
Uno dei contributi metodologici fondamentali nel trattamento dei testi italiani è il metodo concordanziale di Giuseppe Savoca (Savoca G., 2000), che ha gettato le basi per un'analisi testuale sistematica nell'umanistica digitale. Come descritto da Di Silvestro & al. (2022), questo approccio si concentra sull'analisi lessicografica delle parole, costruendo concordanze dove ogni occorrenza è messa in relazione con il suo lemma attraverso il contesto d'uso. Questo metodo attribuisce particolare importanza alla parola come unità fondamentale dell'analisi, tralasciando informazioni grammaticali come tempo e numero nei verbi. Tale impostazione nasce dall'esigenza di mantenere i dati gestibili nel contesto di un processo quasi completamente manuale, evitando una granularità eccessiva nella catalogazione. Sebbene il metodo non sia automatizzato, ha influenzato lo sviluppo di strumenti di umanistica digitale, evidenziando l'importanza di approcci manuali per affrontare casi complessi.
Nel presente lavoro, il metodo concordanziale trova un'applicazione moderna nell'uso del software LiotroConcord_v2, che combina strumenti di lemmatizzazione automatica (spaCy e UDPipe) con regole

1 spaCy è una libreria Python open-source, progettata per essere altamente scalabile e performante, utilizzando modelli pre-addestrati basati su reti neurali per l'analisi morfosintattica e la lemmatizzazione https://spacy.io/ (cons. 26/01/2025)

2 UDPipe, sviluppato dall'Institute of Formal and Applied Linguistics, sfrutta i dati Universal Dependencies (UD) per addestrare modelli di analisi sintattica e morfologica su corpora annotati. La combinazione di analisi sintattica e morfologica rende UDPipe particolarmente adatto per lingue ricche di morfologia come l'italiano. https://ufal.mff.cuni.cz/udpipe (cons. 26/01/2025)

personalizzate implementate manualmente. Questo approccio consente di colmare le lacune dei modelli NLP, estendendo l'efficacia della lemmatizzazione anche ai testi storici e letterari.
Nonostante i progressi significativi, gli strumenti attuali presentano alcune limitazioni. La mancanza di personalizzazione rimane un problema centrale, poiché i modelli pre-addestrati spesso non tengono conto delle specificità di corpora storici o letterari, come la ridotta efficacia nella risoluzione delle ambiguità semantiche, legate a fenomeni come l'omografia e la polisemia: la dipendenza dai corpora annotati limita l'applicabilità dei modelli a domini linguistici poco rappresentati, evidenziando la necessità di soluzioni più flessibili e adattabili (Ciula & al., 2023).

3. NLP E LEMMATIZZAZIONE AUTOMATICA

Il processo di lemmatizzazione automatica per la lingua italiana può presentare diverse problematiche. I sistemi NLP offrono certamente un valido supporto al processo, ma alcuni passi devono essere trattati con un automatismo limitato, come la gestione delle enclitiche, degli omografi e delle forme linguistiche non standard. Per affrontare tali complessità, il flusso di lavoro integrato sviluppato nel software LiotroConcord_v2 utilizza un approccio ibrido basato su spaCy e UDPipe, combinato con regole personalizzate e interventi manuali. Questo sistema si basa su un'analisi comparativa degli output generati dai due strumenti di NLP, con ulteriori correzioni specifiche implementate all'interno della piattaforma 4D.
Il processo inizia con la preparazione del testo, che può essere fornito in formato TXT o XML-TEI. Nel caso di file XML, LiotroConcord_v2 include un modulo dedicato all'estrazione del contenuto dal tag <body>, escludendo automaticamente le sezioni ritenute non rilevanti, grazie a un sistema di esclusione basato su tag e attributi. Una volta estratto, il testo viene normalizzato attraverso una serie di trasformazioni che includono l'uniformazione dei caratteri di newline, la rimozione degli spazi ridondanti o superflui e la sostituzione di eventuali apici o virgolette non standard con caratteri uniformi. La normalizzazione garantisce un input coerente e privo di elementi che potrebbero interferire con l'analisi linguistica.
La tokenizzazione viene realizzata all'interno della piattaforma 4D mediante espressioni regolari altamente personalizzate. Questo approccio permette di identificare con precisione specificità linguistiche dell'italiano, come parole contenenti enclitiche o apostrofi sia iniziali che finali, che spesso non vengono trattate adeguatamente dai sistemi NLP standard. Un esempio è dato dalla parola 'Ntoni, presente nel testo I Malavoglia. Entrambi i sistemi, spaCy e UDPipe, tendono a ignorare l'apostrofo iniziale, che segnala l'aferesi, riconoscendo il token semplicemente come Ntoni. Questa interpretazione non è corretta, poiché l'apostrofo è parte integrante del nome e ne modifica la struttura morfologica.
Dopo la tokenizzazione, il testo viene analizzato separatamente sia da spaCy[3] sia da UDPipe.[4] Ciascun sistema elabora i token, assegnando loro un lemma e una categoria grammaticale (POS), e gli output generati vengono salvati in file distinti per consentire un confronto diretto.

```python
if __name__ == "__main__":
    import argparse

    parser = argparse.ArgumentParser(description="Analizza un file di testo con SpaCy e UDPipe.")
    parser.add_argument("file_path", help="Percorso al file di testo da analizzare.")
    parser.add_argument("udpipe_model_path", help="Percorso al file modello UDPipe.")
    parser.add_argument("language_code", help="Codice lingua (es. 'it' o 'en').")
    args = parser.parse_args()

    analyze_text(args.file_path, args.udpipe_model_path, args.language_code)
```

Figura 1. Entry point dello script Python: esegue l'analisi di un file di testo utilizzando SpaCy e UDPipe, leggendo i parametri dalla riga di comando

Il software confronta quindi i risultati dei due sistemi. Quando entrambi concordano sul lemma e sulla categoria grammaticale, LiotroConcord_v2 accetta automaticamente l'output come valido. Tuttavia, nei casi in cui spaCy e UDPipe producono risultati discordanti, oppure quando il lemma non è presente nel vocabolario di riferimento interno a LiotroConcord_v2, il software, come vedremo nel capitolo successivo, interviene applicando regole personalizzate per risolvere le ambiguità o adattare i risultati al modello di riferimento.

3 Il modello utilizzato è it_core_news_sm

4 Il modello utilizzato è italian-isdt-ud-2.5

4. ERRORI DEI SISTEMI E METODI DI CORREZIONE

Un esempio di adattamento è la forma "del" che in questo modello viene ricondotta unicamente al lemma "di", senza esplicitare la combinazione articolata. Questo approccio differisce da quello adottato da spaCy e UDPipe: per spaCy, "del" è lemmatizzato come "di il ADP", mentre UDPipe separa ulteriormente il token in "di ADP"[5] e "il DET".[6] Per rispettare il modello di riferimento, LiotroConcord_v2 traduce automaticamente l'analisi di spaCy e UDPipe nella rappresentazione consolidata del lemma "di".
Il software effettua anche controlli specifici su differenze semantiche sottili ma sostanziali: se un verbo è preceduto da un articolo, il sistema interpreta la combinazione come un infinito sostantivato.
Come dicevamo, il sistema considera non plausibili i risultati che includono lemmi assenti dal vocabolario di riferimento, come nel caso delle enclitiche o, più in generale, delle forme o dei lemmi inesistenti.
Nel caso del token "tramutandotela", entrambi i sistemi restituiscono un lemma errato: "tramutandotelare". Questo errore dimostra la difficoltà di segmentare correttamente i componenti morfologici, in quanto né spaCy né UDPipe riescono a riconoscere i singoli lemmi "tramutare", "te" e "la". Al contrario, LiotroConcord_v2, grazie all'implementazione di una regola personalizzata che riconosce le enclitiche come le particelle pronominali o avverbiali, scompone correttamente il token nei suoi elementi, associando ciascuno di essi al lemma e alla categoria grammaticale corrette.

```
ARRAY TEXT($arrEnclitiche;17)
$arrEnclitiche{1}:="glie"
$arrEnclitiche{2}:="me"
$arrEnclitiche{3}:="mi"
$arrEnclitiche{4}:="te"
$arrEnclitiche{5}:="ti"
$arrEnclitiche{6}:="le"
$arrEnclitiche{7}:="lo"
$arrEnclitiche{8}:="li"
$arrEnclitiche{9}:="la"
$arrEnclitiche{10}:="ci"
$arrEnclitiche{11}:="ce"
$arrEnclitiche{12}:="vi"
$arrEnclitiche{13}:="ve"
$arrEnclitiche{14}:="si"
$arrEnclitiche{15}:="ne"
$arrEnclitiche{16}:="se"
$arrEnclitiche{17}:="gli"
```

```
: ((([Lemmi]Categoria="sm") | ([Lemmi]Categoria="sf")\
    | ([Lemmi]Categoria="ag") | ([Lemmi]Categoria="di")\
    | ([Lemmi]Categoria="ie") | ([Lemmi]Categoria="in")\
    | ([Lemmi]Categoria="po") | ([Lemmi]Categoria="re"))
  $categoriaGiusta:="ar"
  $lemmagiusto:=""
  If ($arrSplitted2{$i-1}#"l'")
    $lemmagiusto:=Lowercase($arrSplitted2{$i-1})
  Else
    If ([Lemmi]Categoria="sm")
      $lemmagiusto:="lo"
    End if
```

Figura 2. Frammenti di codice 4D per la gestione personalizzata delle enclitiche e degli articoli

Un caso di discordanza tra spaCy e UDPipe è rappresentato dal termine "bramosìa" (in "... la vaga bramosìa dell'ignoto..." da I Malavoglia). UDPipe lo analizza erroneamente come aggettivo (bramosio), mentre spaCy lo classifica correttamente come sostantivo.
Gli esempi di output permettono di comprendere l'efficacia del sistema integrato e le difficoltà che spaCy e UDPipe incontrano nell'analisi linguistica italiana. Tuttavia, grazie al confronto con il dizionario di riferimento integrato in LiotroConcord_v2, il software conferma l'interpretazione corretta, avendo già catalogato il lemma "bramosia" e le relative forme "bramosia" e "bramosie", eliminando così ogni ambiguità.

5. UN CASO DI STUDIO: *I VICERÉ* DALL'OCR ALL'OUTPUT DELLA LEMMATIZZAZIONE IN XML-TEI

Un caso di studio utile a comprendere il *workflow* completo (cfr. *Figura 3*) del sistema di lemmatizzazione è quello dell'edizione scientifica digitale commentata de *I viceré* di Federico De Roberto, promossa all'interno del progetto [PROJECT].

[5] "ADP" sta per *Adposition*; in italiano corrisponde alle preposizioni "di", "a", "da" ...

[6] "DET" sta per *Determiner*; in italiano corrisponde a varie categorie grammaticali: articoli determinativi (es. "il", "la", "i", "le"), articoli indeterminativi (es. "un", "una"), pronomi dimostrativi usati come determinanti (es. "questo", "quello"), determinanti indefiniti (es. "alcuni", "ogni").

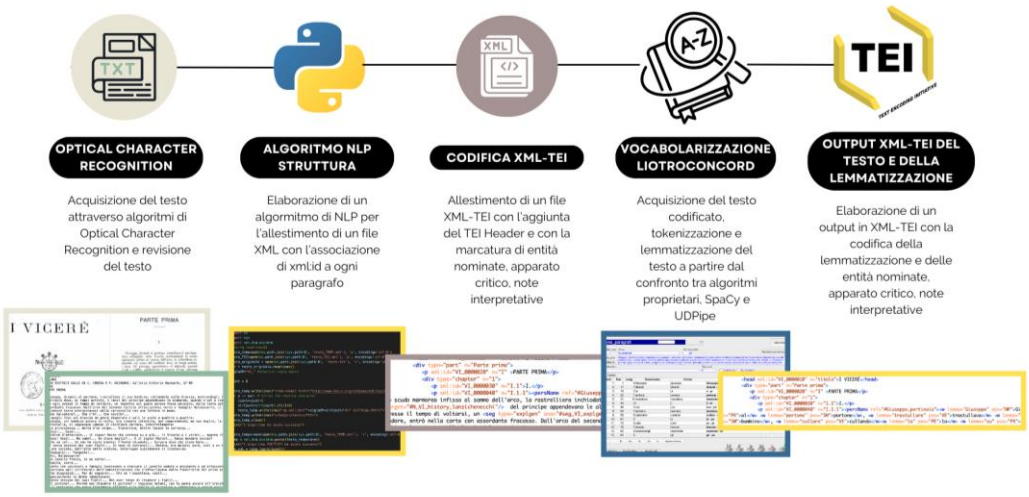


Figura 3. Workflow per la codifica e vocabolarizzazione dei Viceré

A partire dall'acquisizione delle immagini ad alta definizione dell'edizione Galli del 1894 e del conseguente riconoscimento automatico dei caratteri (OCR) è stato prodotto un documento in formato solo testo a cui è stato applicato un algoritmo di *Natural Language Processing* in Python. In particolare, sono state utilizzate le *regular expressions* e il metodo *readline* per costruire la struttura del documento XML associando un @xml:id univoco (fondamentale anche per associare le righe nel sistema di lemmatizzazione). Infine, attraverso l'API del *Document Object Model* è stato *parsato* il testo producendo il documento xml, poi marcato manualmente in linea con il modello di codifica promosso dalla *Text Encoding Initiative* (TEI). In particolare, sono state marcate le entità nominate (<persName>, <placeName>, <orgName>, <eventName>), i riferimenti a personaggi, luoghi, organizzazioni ed eventi (<rs>), alcune parole in disuso o di difficile comprensione (<distinct>), il discorso diretto (<q>), alcuni segmenti testuali a cui associare delle note di esegesi (<seg>) e infine alcune note di carattere storico-bibliografico o interpretativo (con dei *pointer* <ptr> e delle <note>). Si veda come esempio la codifica del primo paragrafo del primo capitolo:

```
<p xml:id="VI_0000040" n="I.1.1"><persName ref="#Giuseppe_portinaio">Giuseppe</persName>, dinanzi al portone, trastullava il suo bambino, cullandolo sulle braccia, mostrandogli lo scudo marmoreo infisso al sommo dell'arco, la rastrelliera inchiodata sul muro del vestibolo dove, ai tempi antichi, i <distinct ana="lanzo">lanzi</distinct><ptr type="history" target="#N_VI_history_lanzichenecchi"/>  del principe appendevano le alabarde, quando s'udì e crebbe rapidamente il rumore d'una carrozza arrivante a tutta carriera; e prima ancora che egli avesse il tempo di voltarsi, un <seg type="explgen" ana="#seg_VI_explgen_legnetto">legnetto</seg> sul quale pareva fosse nevicato, dalla tanta polvere, e il cui cavallo era tutto spumante di sudore, entrò nella corte con assordante fracasso. Dall'arco del secondo cortile affacciaronsi servi e famigli: <persName ref="#BaldassarreCrimi">Baldassarre</persName>, il maestro di casa, schiuse la vetrata della loggia del secondo piano, intanto che <persName ref="#SalvatoreCerra">Salvatore Cerra</persName> precipitavasi dalla carrozzella con una lettera in mano.</p>
```

Il documento così marcato è stato acquisito in *input* da LiotroConcord che, dopo la *tokenizzazione,* ha associato il lemma e il POS corrispondente a ogni forma in base al confronto tra il modello utilizzato e quelli promossi da SpaCy e UDPipe, come detto nei paragrafi precedenti. La *Graphical User Interface* (GUI) di LiotroConcord (cfr. Figura 4) permette dunque al concordatore di intervenire correggendo gli omografi, le associazioni o *tokenizzazioni* errate, i neologismi, ecc. Un importante aspetto riguarda anche la decisione del concordatore su alcuni aspetti che necessitano di una decisione interpretativa (come la scelta tra aggettivo e participio nei participi passati, tra articolo e numerale negli indeterminativi, ecc.).

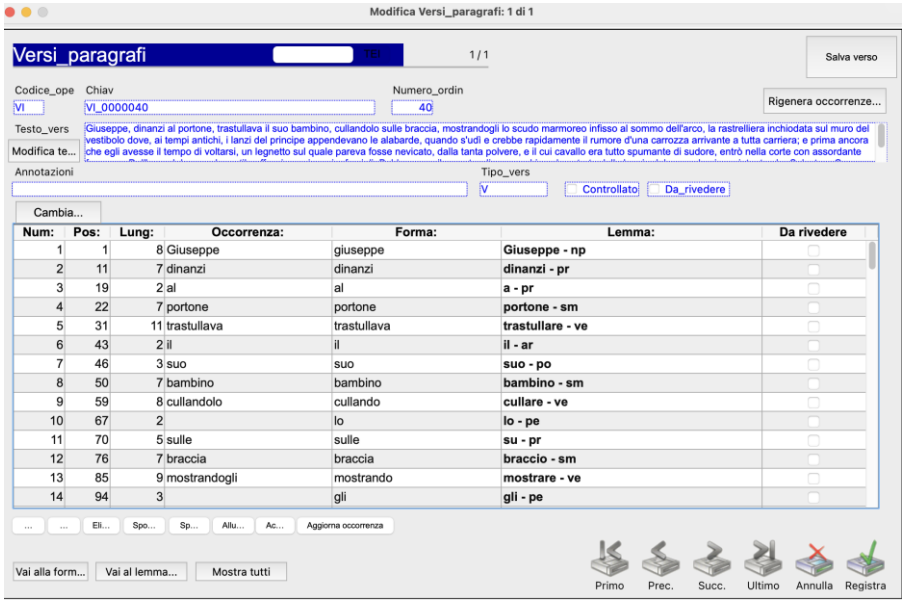


Figura 4. Schermata di LiotroConcord con la lemmatizzazione del primo paragrafo de I Viceré

Infine, la lemmatizzazione viene integrata nella codifica attraverso il tag <w> e gli attributi @lemma e @pos, così da rendere i dati aperti e interoperabili. Si veda dunque l'*output* arricchito del primo paragrafo:

```
<p xml:id="VI_0000040" n="I.1.1"><persName ref="#Giuseppe_portinaio"><w lemma="Giuseppe"
pos="NP">Giuseppe</w></persName>, <w lemma="dinanzi" pos="PR">dinanzi</w> <w lemma="a"
pos="PR">al</w> <w lemma="portone" pos="SM">portone</w>, <w lemma="trastullare"
pos="VE">trastullava</w> <w lemma="il" pos="AR">il</w> <w lemma="suo" pos="PO">suo</w> <w
lemma="bambino" pos="SM">bambino</w>, <w lemma="cullare" pos="VE">cullando</w><w
lemma="lo" pos="PE">lo</w> <w lemma="su" pos="PR">sulle</w> <w lemma="braccio"
pos="sm">braccia</w>, <w lemma="mostrare" pos="VE">mostrando</w><w lemma="gli"
pos="PE">gli</w> <w lemma="lo" pos="AR">lo</w> <w lemma="scudo" pos="SM">scudo</w> <w
lemma="marmoreo" pos="AG">marmoreo</w></p>
```

La possibilità di far dialogare realmente i dati della codifica (come quello sul discorso diretto o sulle entità nominate) e la lemmatizzazione diventa potenzialmente non soltanto un arricchimento della codifica, ma una vera e propria occasione per effettuare interrogazioni complesse con lo sviluppo di concordanze selettive (per capitoli, parti, personaggi) aprendo a un nuovo modo di intendere la vocabolarizzazione, la codifica, *distant* e *close reading* e l'applicazione ermeneutica di tali teorie e modelli.

6. CONCLUSIONI

Questo lavoro ha dimostrato come un approccio ibrido, che integra spaCy, UDPipe e il software personalizzato LiotroConcord_v2, possa affrontare efficacemente le complessità linguistiche della lingua italiana, garantendo una lemmatizzazione accurata grazie all'uso di tecnologie avanzate, combinato con regole manuali mirate.

I prossimi sviluppi prevedono l'addestramento di modelli NLP basati sulle concordanze già gestite da LiotroConcord_v2. Questo processo potrebbe migliorare ulteriormente l'accuratezza degli strumenti esistenti, affinando la capacità di spaCy e UDPipe di gestire le specificità della lingua italiana. I nuovi modelli, addestrati su un corpus così ricco e curato, potrebbero ridurre gli errori di lemmatizzazione e disambiguazione, offrendo risultati più precisi.

BIBLIOGRAFIA

Alishahi, A., Chrupała, G., & Linzen, T. (2019). Analyzing and interpreting neural networks for NLP: A report on the first BlackboxNLP workshop. Natural Language Engineering, 25(4), 543–557. https://doi.org/10.1017/S135132491900024X.

Blaise, A. (2022). Do Large Language Models Understand Us? Daedalus, 151(2), 183–197. https://doi.org/10.1162/daed_a_01909.

Boleda, G. (2020). Distributional semantics and linguistic theory. Annual Review of Linguistics, 6, 213–234. https://doi.org/10.1146/annurev-linguistics-011619-030303.

Borges, J. L. (1998). The library of Babel. In Collected Fictions.

Ciotti, F. (2023). "La codifica del testo, XML e la TEI", in Digital Humanities. Metodi, strumenti, saperi. Carocci.

Ciula, A., Eide, Ø., Marras, C., & Sahle, P. (2023). Modelling between digital and humanities: Thinking in practice. Open Book Publishers. https://doi.org/10.11647/OBP.0369.

Clarke, A. C. (1967). The nine billion names of God. Harcourt.

Di Silvestro, A., D'Agata, C., Palazzolo, G., & Sichera, P. (2022). Conservazione e fruizione di banche dati letterarie: L'archivio della poesia italiana dell'Otto/Novecento di Giuseppe Savoca. In F. Ciracì, G. Miglietta, & C. Gatto (Eds.), Quaderni di Umanistica Digitale. AIUCD 2022 - Proceedings (pp. 1–12). AIUCD. https://doi.org/10.6092/unibo/amsacta/6848.

Floridi, L. (2023). AI as agency without intelligence: On ChatGPT, large language models and other generative models. Philosophical Technology, 36, 15. https://doi.org/10.1007/s13347-023-00621-y.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. Cambridge: MIT Press.

Hassija, V., Chamola, V., Mahapatra, A., et al. (2024). Interpreting Black-Box Models: A review on explainable artificial intelligence. Cognitive Computation, 16, 45–74. https://doi.org/10.1007/s12559-023-10179-8.

Hewitt, J., & Manning, C. D. (2019). A structural probe for finding syntax in word representations. In Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 4129–4138). New York: Association for Computational Linguistics.

Kumar, S., Sumers, T. R., Yamakoshi, T., Goldstein, A., Hasson, U., Norman, K. A., et al. (2023). Reconstructing the cascade of language processing in the brain using the internal computations of a transformer-based language model. bioRxiv. https://doi.org/10.1101/2022.06.08.495348.

Lodder, J. (2009). Binary Arithmetic: From Leibniz to von Neumann. https://doi.org/10.5948/UPO9780883859742.023.

Mickus, T., Paperno, D., & Constant, M. (2022). How to dissect a muppet: The structure of transformer embedding spaces. Transactions of the Association for Computational Linguistics, 10, 981–996. https://doi.org/10.1162/tacl_a_00501.

Min, B., Ross, H., Sulem, E., Veyseh, A. P. B., Nguyen, T. H., Sainz, O., et al. (2023). Recent advances in natural language processing via large pre-trained language models: A survey. ACM Computing Surveys, 56(1), 1–40. https://doi.org/10.1145/3605943.

Palazzolo, G. (2023). Lessicografia letteraria e metodo concordanziale: Storia, modelli e sfide per le Digital Humanities. Linguistica e Letteratura, 48(1–2), 23–45. https://digital.casalini.it/10.19272/202301602006.

Primiero, G. (2020). On the Foundations of Computing. Oxford: Oxford University Press.

Savoca, G. (2000). Lessicografia letteraria e metodo concordanziale. Firenze: Olschki.

Tenney, I., Das, D., & Pavlick, E. (2019). BERT rediscovers the classical NLP pipeline. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 4593–4601). Somerset, NJ: Association for Computational Linguistics.

Turing, A. (1950). Computing machinery and intelligence. Mind, 59(236), 433–460. https://doi.org/10.1093/mind/LIX.236.433.

Reverse Engineering Critical Apparatuses for HTR Ground Truth Creation: The Case of Kennicott's Collation of the Hebrew Bible

Luigi Bambaci[1], Nachum Dershowitz[2], Daniel Stökl Ben Ezra[3]
[1] École Pratique des Hautes Études, Paris Sciences & Lettres, Paris, France, luigi.bambaci@ephe.psl.eu
[2] School of Computer Science and AI, Tel Aviv University, Ramat Aviv, Israel, nachum@tau.ac.il
[3] École Pratique des Hautes Études, Paris Sciences & Lettres, Paris, France, Daniel.Stoekl@ephe.psl.eu

ABSTRACT (ITALIANO)

Gli apparati critici sono strumenti fondamentali per lo studio della storia della tradizione e critica del testo. Grazie alla registrazione delle varianti testuali emerse durante il processo di trasmissione, essi offrono dati imprescindibili per l'analisi filologica, linguistica, paleografica, e non solo. Tuttavia, la loro complessità strutturale e l'elevata mole di dati rendono difficile l'analisi su larga scala con approcci manuali. In questo contributo, presentiamo un piano di lavoro per la digitalizzazione e la codifica automatica degli apparati critici, basato su tecnologie all'avanguardia di OCR e trattamento automatico del linguaggio. Oltre alla digitalizzazione e alla codifica, perseguiamo un obiettivo innovativo: la ricostruzione automatica di testi manoscritti a partire dalle varianti registrate in apparato, con l'intento di generare nuova ground truth per migliorare l'addestramento di modelli di Handwritten Text Recognition (HTR). Il nostro caso studio si concentra sull'edizione di Benjamin Kennicott (1776-1780), la più ampia raccolta di varianti testuali mai realizzata per la Bibbia Ebraica, e oggetto del progetto *Reverse Engineering Kennicott* (REK). Dimostriamo come sia possibile segmentare, trascrivere e codificare automaticamente grandi volumi di dati testuali, generando testi manoscritti completi con intervento manuale contenuto. Affrontiamo le principali sfide tecniche e metodologiche emerse durante tutto il processo, facendo infine qualche previsione per il lavoro futuro.

Parole chiave: apparato critico; riconoscimento del testo manoscritto; trattamento del linguaggio naturale; manoscritti della Bibbia Ebraica; collazione di Kennicott

ABSTRACT (ENGLISH)

Ingegneria Inversa degli Apparati Critici per la Creazione di Ground Truth per l'HTR: Il Caso della Collazione di Kennicott della Bibbia Ebraica

Critical apparatuses are essential tools for the study of textual tradition and textual criticism. By recording the variants that emerge during the transmission process, they provide indispensable data for philological, linguistic, paleographic, and other analysis. However, their structural complexity and the sheer volume of data they contain make large-scale analysis through manual methods particularly challenging. In this paper, we present a workflow for the digitization and automatic encoding of critical apparatuses, based on state-of-the-art technologies in OCR and natural language processing. Beyond digitization and encoding, we pursue an innovative objective: the automatic reconstruction of manuscript texts from the variants recorded in the apparatus, with the goal of generating new ground truth data to improve the training of Handwritten Text Recognition (HTR) models.

Our case study focuses on Benjamin Kennicott's edition (1776–1780), the most extensive collection of textual variants ever compiled for the Hebrew Bible, and the object of the *Reverse Engineering Kennicott* (REK) project.

We demonstrate how large volumes of textual data can be segmented, transcribed, and encoded automatically, generating complete manuscript texts with limited manual intervention. Finally, we discuss the main technical and methodological challenges encountered throughout the process and outline directions for future work.

Keywords: critical apparatus; Handwritten Text Recognition; Natural Language Processing; Hebrew Bible manuscripts; Kennicott's collation

1. INTRODUCTION

Critical apparatuses are an indispensable resource for textual and philological research. Initially conceived as interfaces to justify editorial choices and provide readers with a synthetic view of textual traditions, they have progressively evolved into (semi-)structured data repositories serving purposes far beyond traditional textual philology—including the study of scribal practices, paleography, historical linguistics, and the computational analysis of textual transmission.

Despite their value, critical apparatuses remain challenging to manage. Their structural complexity and the volume of data they contain often limit accessibility, requiring analytical efforts that are not always scalable for large or data-driven projects. In recent years, digital technologies have opened new possibilities for utilizing critical apparatuses. Advances in Optical Character Recognition (OCR) have enabled automated transcription of multiformat and multilingual documents with impressive results. Natural Language Processing (NLP) techniques facilitate the organization and interpretation of complex information, while encoding systems like the Text Encoding Initiative (TEI) provide robust standards for preserving data in interoperable formats, ensuring long-term usability and exchangeability.

Among the studies on the digital processing of critical apparatuses relevant to our purposes, we highlight (Boschetti and Zemanek, 2007), which addresses the challenge of aligning textual variants present in critical apparatuses with their reference texts. By employing edit distance algorithms, the study proposes methodologies to accurately identify the position of each variant within the reference text and link it to additional annotations, such as morphological or metrical data. (Boschetti, 2007) examines approaches to enrich classical digital corpora by integrating textual variants and conjectures from critical apparatuses, using XML markup and parsing algorithms.

On the OCR/HTR front, (Boschetti et al., 2009) propose a workflow that combines multiple alignment techniques with customized OCR engines to improve the transcription of critical Greek and Latin editions, with particular attention to complex layouts including critical apparatuses. (Romanello et al., 2021) compare the performance of different OCR pipelines on 19th-century classical commentaries, highlighting the importance of ground truth datasets and advanced post-processing for optimizing character recognition in multilingual texts.

While the digitization and computational processing of critical apparatuses has been—as these studies testify—the focus of considerable attention and progress, the use of textual variants to automatically reconstruct manuscript texts for HTR training remains a largely unexplored field.

In a previous study (Bambaci and Stoekl Ben Ezra 2023), we took our first steps in this direction, using the critical apparatus from Kennicott's collation of the Hebrew Bible (Kennicott 1776–1780). Through the transcription and analysis of the first of Kennicott's two volumes, we successfully reconstructed complete texts for over 100 witnesses of the book of Genesis, generating data corresponding approximately to 5,800 manuscript pages.

In this paper, we aim to expand and refine those results, providing a detailed report on the progress made so far and the prospects for future development.

The approach presented here has been developed as part of the "Reverse Engineering Kennicott" (REK) project, dedicated to digitizing and encoding Kennicott's work and reconstructing the texts of his witnesses.

2. KENNICOTT'S COLLATION OF THE HEBREW BIBLE

The Hebrew Bible is among the most influential texts in human history, forming the foundation of Jewish and Christian religious and cultural heritage. Originally written in Hebrew and Aramaic, and translated into numerous languages since antiquity, it comprises approximately 300,000 tokens across 24 books.

One of the most ambitious scholarly efforts ever undertaken on its textual tradition is Benjamin Kennicott's *Vetus Testamentum Hebraicum cum variis lectionibus*, published in Oxford between 1776 and 1780. Drawing on over 600 manuscripts and 70 printed editions, this two-volume work documents an estimated 1.5 million textual variants, meticulously organized in a critical apparatus spanning roughly 1,400 pages—making it the most comprehensive and systematic collation ever assembled for the Hebrew witnesses of the Bible.

Besides its scope, three features make Kennicott's edition stand out from both earlier and later works: the clear distinction between fully and partially collated witnesses; the high level of granularity of the collation; and the highly formalized language of the apparatus.

The first two features have clear philological relevance, as they legitimise interpreting the absence of a variant as agreement with the reference text—effectively compensating for the negative structure of the apparatus, which records only deviations. The third, instead, plays a key role from a computational perspective: the use of a controlled vocabulary and consistent syntactic structures, combined with minimal reliance on natural language, facilitates parsing and structured data extraction.

Taken together, these characteristics make Kennicott's collation particularly well-suited for digital processing. We demonstrate this in the following section, where we outline the key operational phases of

our workflow, namely: image acquisition, segmentation and transcription, parsing and encoding, and finally manuscript text reconstruction followed by alignment with the original manuscripts.

3. WORKFLOW

The image acquisition phase involved two complementary datasets. First, we collected high-resolution images of both volumes of Kennicott's *Vetus Testamentum*, primarily sourced from Archive.org. In parallel, we obtained approximately 15,000 images from 20 Hebrew Bible manuscripts fully collated by Kennicott, retrieved from Ktiv[1] and selected for their historical and textual significance, as well as for the regularity of their layout and script.

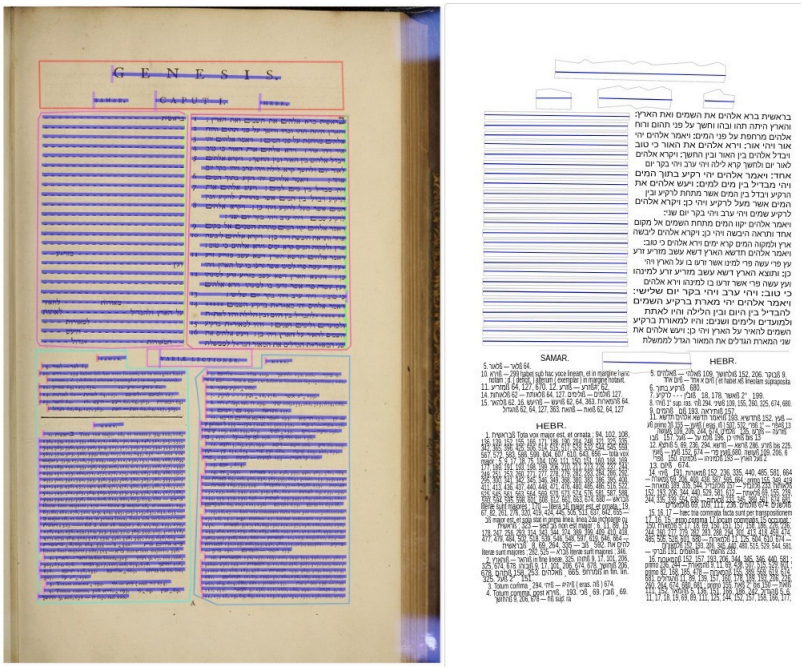


Figure 1: Segmentation and transcription of the *Vetus Testamentum*

The next step was segmentation and transcription. Kennicott's edition presents a particularly complex layout (Fig. 1), with two distinct textual zones: at the top, one or two columns of Hebrew containing the reference text used as the basis for his collation; at the bottom, two columns dedicated to the critical apparatus, where textual variants are recorded. The reference text is primarily in Hebrew with RTL (right-to-left) directionality, while the apparatus includes Hebrew and Latin text with LTR (left-to-right) directionality. Using eScriptorium cum Kraken (Stokes et al. 2021, Kiessling et al. 2019), we segmented the regions of interest (reference text and apparatus), excluding irrelevant paratextual elements such as headers, page numbers, and catchwords. We trained our segmentation models through iterative cycles of manual correction and retraining, using batches of 30–50 pages per round, until the desired level of segmentation accuracy was achieved.

Automatic transcription presented significant challenges, with notable differences between the treatment of the reference text and the critical apparatus. The transcription of the reference text was relatively straightforward, as it benefited from the biblical Hebrew HTR models developed within the BiblIA project (Stoekl Ben Ezra et al., 2021), which delivered highly accurate results. Transcribing the critical apparatus, by contrast, was significantly more complex, owing to its multilingual content, symbolic notation, and mixed text directionality (RTL Hebrew and LTR Latin). To address these challenges, we trained custom HTR models on targeted samples, following an iterative process similar to the one adopted for segmentation.

We then proceeded with parsing. For the reference text, we used simple Python scripts to align the automatic transcription with an already existing digital version of the biblical text, producing XML code structured into chapters, verses, and tokens. For the critical apparatus, a much more sophisticated approach was required. Using context-free grammars (CFGs) and the ANTLR4 software (Parr 2010, 2012),

[1] An international digital initiative that provides centralized access to thousands of Hebrew manuscripts from collections around the world, developed by the National Library of Israel. Available at: https://www.nli.org.il/en/discover/manuscripts/hebrew-manuscripts

and following the methodology outlined in detail in (Bambaci and Boschetti 2020, Bambaci 2021), we developed a rule-based parser capable of identifying apparatus components—such as lemmas, variants, and witness sigla (Fig. 2)—and automatically producing XML code (Fig. 3).

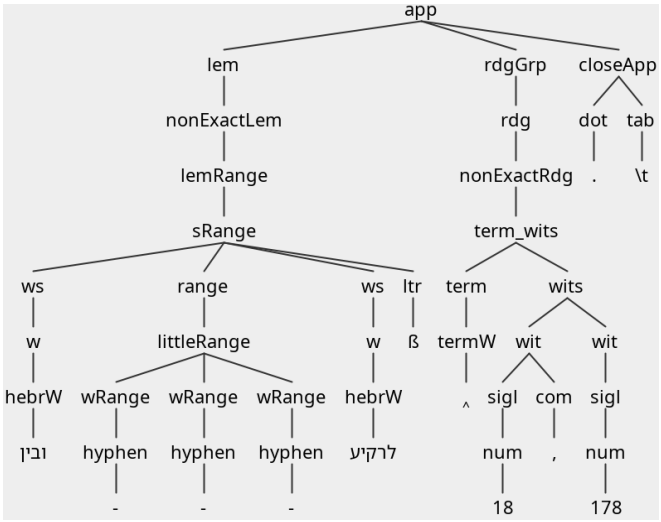


Figure 3: Parser's syntactic tree for Genesis 1:7

```
<app line="7" part="466569">
  <lem>
    <nonExactLem>
      <lemRange>
        <sRange>
          <ws>
            <w>
              <hebrW>וב׳ן</hebrW>
            </w>
          </ws>
          <range>
            <littleRange>
              <wRange>
                <hyphen>-</hyphen>
              </wRange>
              <wRange>
                <hyphen>-</hyphen>
              </wRange>
              <wRange>
                <hyphen>-</hyphen>
              </wRange>
            </littleRange>
          </range>
          <ws>
            <w>
              <hebrW>לרק׳יע</hebrW>
            </w>
          </ws>
          <ltr>ß</ltr>
        </sRange>
      </lemRange>
    </nonExactLem>
  </lem>
  <rdgGrp>
    <rdg>
      <nonExactRdg>
        <term_wits>
          <term>
            <termW>□</termW>
          </term>
          <wits>
            <wit>
              <sigl>
                <num>18</num>
              </sigl>
              <com>,</com>
            </wit>
            <wit>
              <sigl>
                <num>178</num>
              </sigl>
            </wit>
          </wits>
        </term_wits>
      </nonExactRdg>
    </rdg>
  </rdgGrp>
  <closeApp>
    <dot>.</dot>
    <tab>\t</tab>
  </closeApp>
</app>
```

Figure 2: XML output generated by the parser

The next phase focused on reconstructing complete manuscript texts from the variants recorded in the critical apparatus. This process involved three main steps, carried out using alignment techniques based primarily on Levenshtein distance: mapping each lemma from the apparatus onto its corresponding position in the reference text; reconstructing the Hebrew text of the variant readings when necessary; and inserting the variants into the reconstructed manuscript texts at the appropriate points, thus generating full versions of the witnesses with variants presented *in textu* (Figg. 4, 5).

```
<app line="7" part="466569">
  <lem>
    <exactLem>
      <recoLem>
        <s>
          <ws>
            <w>וב׳ן</w>
            <w n="2">המ׳ם</w>
            <w n="2">אשר</w>
            <w>מעל</w>
            <w n="2">לרק׳יע</w>
          </ws>
          <ltr>ß</ltr>
        </s>
        <recoLemSep>]</recoLemSep>
      </recoLem>
    </exactLem>
  </lem>
  <rdgGrp>
    <rdg>
      <exactRdg>
        <rdgContent>
          <w line="7"/>
          <w line="7"/>
          <w line="7"/>
          <w line="7"/>
          <w line="7"/>
        </rdgContent>
        <wits>
          <wit>
            <sigl>
              <num>18</num>
            </sigl>
            <com>,</com>
          </wit>
          <wit>
            <sigl>
              <num>178</num>
            </sigl>
          </wit>
        </wits>
      </exactRdg>
    </rdg>
  </rdgGrp>
  <closeApp>
    <dot>.</dot>
    <tab> </tab>
  </closeApp>
</app>
```

Figure 5: XML apparatus with reconstructed lemma

```
<ab chap="1" n="7">
  <w>ויעש</w>
  <w>אלהים</w>
  <w>את</w>
  <w>הרק׳יע</w>
  <w>ויבדל</w>
  <w>ב׳ן</w>
  <w>המ׳ם</w>
  <w>אשר</w>
  <w>מתחת</w>
  <w>לרק׳יע</w>
  <app>
    <lem>
      <w line="7"/>
      <w line="7"/>
      <w line="7"/>
      <w line="7"/>
      <w line="7"/>
    </lem>
    <rdg>
      <w>וב׳ן</w>
      <w n="2">המ׳ם</w>
      <w n="2">אשר</w>
      <w>מעל</w>
      <w n="2">לרק׳יע</w>
    </rdg>
  </app>
  <w>ויהי</w>
  <w>כן</w>
  <w>:</w>
</ab>
```

Figure 4: XML output of the manuscript text (ms. no. 18)

The final phase is textual alignment. We first automatically transcribe the manuscript images using the dedicated segmentation and transcription models mentioned above. We then align these automatic transcriptions with the reconstructed texts using Passim (Smith 2015, 2023), a text-reuse detection tool recently integrated into
eScriptorium (Fig. 6).

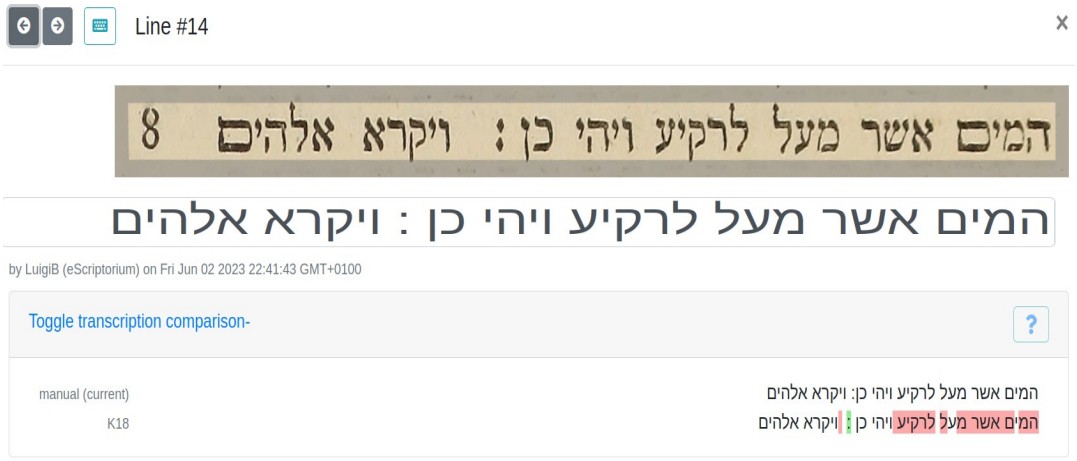


Figure 6: Passim alignment in eScriptorium: reference text vs. reconstructed text (ms. no. 18)

This alignment enables us to systematically compare the two outputs, identifying both errors—whether introduced during the automatic transcription phase or during the reconstruction of the manuscript texts from the apparatus—and actual variants. A portion of these aligned transcriptions will be manually corrected to produce a gold-standard dataset, which will be then used as ground truth to refine and train new Hebrew HTR models.

4. RESULTS

To date, we have completed the segmentation and transcription of both volumes of the *Vetus Testamentum*, amounting to approximately 90,000 lines for the reference text and 115,000 lines for the critical apparatus.

Ten of the 20 manuscripts fully collated by Kennicott have also been entirely segmented and automatically transcribed.

Parsing has been completed for the entire first volume, covering nine books of the Hebrew Bible, from Genesis to Kings (the Enneateuch). The automatic encoding of the critical apparatus produced approximately 50,000 lemmas (<lem>) and 100,000 variants (<rdg>).

In terms of reconstruction, we generated complete texts for 153 manuscripts, amounting to around 250,000 lemmas, an equal number of variants, and approximately 14 million tokens of text.

We are now aligning the reconstructed texts with the automatic transcriptions of the 10 manuscripts using Passim. We are still in the early stages of this process: while preliminary tests have confirmed the technical feasibility of the procedure, most of the alignment work remains to be completed. The coming months will be crucial for assessing the overall reliability of the reconstruction process and, consequently, for the creation of ground truth data.

Once this phase is completed, we will move on to Kennicott's second volume and to the remaining ten manuscripts, through which we expect to approximately double the volume of extracted and reconstructed data units. All relevant textual data, currently encoded in custom XML, will be converted into TEI-compliant format to ensure broader reusability.

5. CONCLUSION

The workflow we have implemented has proven effective across the various phases of the project. Segmentation and transcription models achieved an overall accuracy of around 98%, significantly reducing the need for manual correction. The *Vetus Testamentum* required considerable training effort during the initial phases, but the results—refined through multiple revision cycles—are now highly satisfactory. As for the ten manuscripts, the results also appear promising, thanks both to the regularity of their layout and the quality of their script (which guided their selection), as well as to the performance of the pretrained segmentation and transcription models used.

The parser achieved approximately 97% accuracy—a notable result given the volume and complexity of the data. The highly formalized language of Kennicott's apparatus allowed for the automatic processing and encoding of tens of thousands of entries, with minimal manual intervention.

The automatic reconstruction of manuscript texts from variant data proved to be a viable solution, allowing for the systematic generation of large textual outputs.
Finally, the alignment system based on Passim is proving effective for matching the reconstructed texts with the original manuscript transcriptions. Its seamless integration into eScriptorium's user-friendly interface, in particular, is greatly facilitating the ongoing correction.
While the results are promising, the workflow has also revealed a number of challenges and unresolved issues that deserve careful consideration.
First, although Kennicott's critical apparatus is highly formalized, some residual cases remain—due either to occasional use of natural language, to rare patterns, or to a combination of both—which require specific adjustments to the original transcriptions for proper parsing. However, such adjustments are negligible in quantitative terms and not critical in qualitative ones, as the philological information is preserved.
Numerically more significant are certain systematic ambiguities (approximately 3% of cases), where variants are mistakenly parsed as lemmas. Fortunately, we are able to identify and resolve them automatically by cross-checking the apparatus data against the reference text. Spot checks on particularly complex apparatus entries have not revealed additional semantic errors, though their presence elsewhere cannot be entirely ruled out.
Substantial manual intervention is instead proving necessary in the phase of textual reconstruction. Among these are cases in which multiple lemmas are returned with identical Levenshtein distance scores. Although numerous, they are generally easy to disambiguate manually, as they merely require selecting the correct match. Far more complex to handle are the approximately 4% of lemmas and variants that fall outside our mapping rules, due to irregular or complex structures. While modest in relation to the total, this percentage still amounts to several hundred cases that have required—and will likely continue to require—not only manual intervention, but also a certain degree of subjective judgment.
A final consideration concerns the overall reliability of Kennicott's collation. While extraordinary for its time, the accuracy of his work—as many scholars have long noted—is not always consistent, and the reconstructed texts will inevitably reflect this.
Since the alignment phase is still underway, we are not yet able to provide the necessary statistics on such flaws, nor on errors introduced during transcription, parsing, or automatic reconstruction. However, we anticipate that these will be unevenly distributed: minimal in the case of transcription and parsing, where both the models and the source data have proven reliable, and more frequent in the reconstruction process, where manual intervention has been greater and where Kennicott's collation may at times prove less reliable.
While the alignment phase is still ongoing, the results obtained so far confirm the viability of reconstructing manuscript texts from structured variant data, provided a sufficient degree of formalization in the apparatus. As this phase advances, we expect it to consolidate the workflow, clarify its limitations, and support the creation of reliable ground truth datasets for future applications in HTR of Hebrew biblical manuscripts.
All the textual data relevant to the project as well as the tools implemented and used will be shared on our GitHub repository,[2] and possibly on Zenodo.

6. ACKNOWLEDGEMENT

Our research received generous funding from the Agence Nationale de Recherche as part of the Programme d'investissements d'avenir within the France 2030 framework, under the reference ANR-21-ESRE-0005. Additionally, we benefited from funding by the European Union through the MiDRASH project (ERC, project number 101071829). Please note that the views and opinions expressed in this paper are those of the author(s) alone and do not necessarily represent the views of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for the expressed views.

7. REFERENCES

Bambaci, L. (2021). Critical Apparatus as Domain Specific Languages. A Rule-based Parser for Encoding an Eighteenth-Century Collation of Hebrew Manuscripts. *International Journal of Information Science and Technology*, *5*(1), 22–33.

Bambaci, L., & Boschetti, F. (2020). Encoding the Critical Apparatus by Domain Specific Languages – The Case of the Hebrew Book of Qohelet. In C. Marras, M. Passarotti, G. Franzini, & E. Litta (Eds.), *Atti del*

[2] https://github.com/LuigiBambaci/

IX Convegno Annuale AIUCD. La svolta inevitabile: Sfide e prospettive per l'Informatica Umanistica (pp. 7–13). Università Cattolica del Sacro Cuore. http://amsacta.unibo.it/6316/

Bambaci, L., & Stoekl Ben Ezra, D. (2023). Enhancing HTR of Historical Texts through Scholarly Editions: A Case Study from an Ancient Collation of the Hebrew Bible. In A. Šeļa, F. Jannidis, & I. Romanowska (Eds.), *Proceedings of the Computational Humanities Research Conference 2023* (Vol. 3558, pp. 554–576). CEUR. https://ceur-ws.org/Vol-3558/#paper6310

Boschetti, F. (2007). Methods to Extend Greek and Latin Corpora with Variants and Conjectures: Mapping Critical Apparatuses Onto Reference Text. *Proceedings of the Corpus Linguistics Conference CL, 14-17 Jul. 2007*. https://www.birmingham.ac.uk/documents/college-artslaw/corpus/conference-archives/2007/150Paper.pdf

Boschetti, F., Romanello, M., Babeu, A., Bamman, D., & Crane, G. (2009). Improving OCR Accuracy for Classical Critical Editions. In M. Agosti, J. Borbinha, S. Kapidakis, C. Papatheodorou, & G. Tsakonas (Eds.), *Research and Advanced Technology for Digital Libraries* (pp. 156–167). Springer Berlin Heidelberg.

Boschetti, F., & Zemanek, P. (2007). Alignment of variant readings for linkage of multiple annotations. *Proceedings of the ECAL 2007 Electronic Corpora of Ancient Languages, Prague 16-17 November 2007*, 11–24.

Kennicott, B. (1776-1780). *Vetus Testamentum Hebraicum cum variis lectionibus*, 2 Voll. Clarendon.

Kiessling, B., Tissot, R., Stokes, P., & Stoekl Ben Ezra, D. (2019). eScriptorium: An Open Source Platform for Historical Document Analysis. *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, *2*, 19–24. https://doi.org/10.1109/ICDARW.2019.10032

Parr, T. (2010). *Language Implementation Patterns: Create Your Own Domain-specific and General Programming Languages*. Pragmatic Bookshelf.

Parr, T. (2012). *The Definitive ANTLR 4 Reference*. Pragmatic Bookshelf.

Romanello, M., Najem-Meyer, S., & Robertson, B. (2021). Optical character recognition of 19th century classical commentaries: The current state of affairs. *The 6th International Workshop on Historical Document Imaging and Processing*, 1–6.

Smith, D. A., Cordell, R., & Mullen, A. (2015). Computational Methods for Uncovering Reprinted Texts in Antebellum Newspapers. *American Literary History*, *27*(3), E1–E15. https://doi.org/10.1093/alh/ajv029

Smith, D. A., Murel, J., Allen, J. P., & Miller, M. T. (2023). Automatic Collation for Diversifying Corpora: Commonly Copied Texts as Distant Supervision for Handwritten Text Recognition. *Proceedings of the Computational Humanities Research Conference 2023, Paris, France, December 6-8, 2023.*, 206–221. https://ceur-ws.org/Vol-3558/paper1708.pdf

Stoekl Ben Ezra, D., Brown-DeVost, B., Jablonski, P., Lapin, H., Kiessling, B., & Lolli, E. (2021). BiblIA - A General Model for Medieval Hebrew Manuscripts and an Open Annotated Dataset. *The 6th International Workshop on Historical Document Imaging and Processing*, 61–66. https://doi.org/10.1145/3476887.3476896

Stokes, P., Kiessling, B., Stoekl Ben Ezra, D., Tissot, R., & Gargem, E. (2021). The eScriptorium VRE for Manuscript Cultures. *Ancient Manuscripts and Virtual Research Environments*, *Special issue of Classics@ 18*. https://classics-at.chs.harvard.edu/classics18-stokes-kiessling-stokl-ben-ezra-tissot-gargem/

Data and Knowledge Representation

Prototyping an Atlas of Early Modern English Drama: An Experiment on DraCor Data

Luca Giovannini[1], Andreas Wagner[2]
[1] University of Potsdam, Germany – giovannini@uni-potsdam.de
[2] University of Erlangen-Nuremberg, Germany - andreas.w.wagner@fau.de

ABSTRACT (ENGLISH)
This short contribution leverages the recently released *English Drama Corpus* from the *DraCor* project (Fischer et al. 2019) to create an interactive map of geographical references in early modern English drama by means of Named Entity Recognition (NER) tools. The prototype allows to visualise interesting patterns in the playwrights' use of locations while underscoring the need of improved NLP tools to effectively tackle geospatial research questions about early modern literature.

Keywords: early modern drama; spatial humanities; geocoding; digital cartography

ABSTRACT (ITALIANO)
Un prototipo di atlante per il dramma inglese della prima modernità. Questo contributo presenta una mappa interattiva dei riferimenti geografici contenuti all'interno dell'*English Drama Corpus*, recentemente pubblicato all'interno della piattaforma *DraCor* (Fischer et al. 2019). Il prototipo mostra alcuni interessanti prospettive nell'uso dei luoghi da parte dei drammaturghi, ma sottolinea al contempo la necessità di sviluppare nuovi e più avanzati modelli per il trattamento automatico di testi della prima modernità.

Parole chiave: letteratura drammatica; geocodifica; cartografia digitale

1. BACKGROUND
Best exemplified by works like the *Atlas of Literature* (Bradbury 1996) or the *Atlas of the European Novel, 1800-1900* (Moretti 1999), the 'spatial turn' the humanities experienced in the last 20 years has long invested literary studies as well, resulting in multiple projects that utilise cartographic tools to explore texts in their relations with real or fictional places. Furthermore, the large availability of digitised corpora and the constant improvements in Named Entity Recognition (NER) techniques (cf. Keraghel, Morbieu and Nadif, 2024) have naturally given many of these projects a digital edge (cf. Perenič 2014, Mitchell 2017). While the geographical dimension of early modern English drama has often been investigated with respects to its performance locations, [1] quantitative approaches to the plays' contents from this angle are still less developed. This contribution aims to constitute a first exploratory step in that direction by prototyping and showcasing a map of places mentioned in a large number of early modern English texts. In doing so, this proof of concept also represents one of the first research applications of the recently released *English Drama Corpus* (*EngDraCor*). This collection situates itself within the wider *DraCor* ecosystem (Fischer et al. 2019), i.e., a large collection of TEI-encoded, machine-actionable corpora of dramatic texts from different literary traditions. *EngDracor* currently features 434 English plays from the 1550s to the 1650s, mostly onboarded from the *Early Print* initiative (Mueller 2012), and can serve as a valuable resource for computational literary scholars, enabling the exploration of diverse research questions related to the structure, themes, and stylistic elements of the works.

2. DATA MODELLING
In our experiment, we first utilised the *DraCor* API [2] to iteratively extract and store all spoken text from the plays. [3] Performing NER on early modern texts, however, proved to be quite challenging: the procedures surveyed in Humbel et al. (2021) were often tailored to specific use cases and rather difficult to generalise, even though we ended up implementing a pipeline somehow similar to Gregory and Hardie (2021). Directly deploying appropriate linguistic models was also hard: MacBERTh (Manjavacas and

[1] See e.g. the well-established *Records of Early English Drama* (*REED*) project (https://ereed.org).
[2] https://dracor.org/doc/api.
[3] This is the format of the API call used: https://dracor.org/api/v1/corpora/eng/plays/{slug}/spoken-text. The slug is the unique identifier for each DraCor play, usually made up by the author's surname and the shortened title (e.g. *marlowe-doctor-faustus*).

Fonteyn 2022), a BERT model fine-tuned on early modern texts, lacks a module devoted to NER, while MonadGPT (Langlais 2023) is a conversational agent quite unsuited to our task. [4] On the other hand, modernisation of spelling to improve NER performance, e.g. through software like VARD (Baron and Rayson 2008), seemed not a sustainable option in our case.

Given the lack of satisfactory out-of-the-box solutions, we opted for experimenting with modern standard named entity recognition pipelines, testing the performances of the packages *spaCy* and *stanza*. [5] Despite some preprocessing steps to improve general reliability, the first package performed quite poorly, with close reading of results revealing a very low precision rate (only about 20% of the items retrieved were actual geographical entities [6]). Results with *stanza* were more satisfactory, but still required extensive post-correction. To this aim, we experimented with rule-based solutions, involving dictionaries of geographical entities and gazetteers (e.g. the *World Historical Gazetteer*, cf. Jakacki 2024), but without satisfying results. Eventually, after an extended trial-and-error phase, we settled for the most time- and result-efficient pipeline, which involved fixing most recurrent errors through the 'clustering' and 'reconciliation' functions of the OpenRefine software [7] and then performing a thorough LLM-assisted clean-up (via GPT4, Achiam et al. 2023). In this phase, we tried to lay the groundwork for the later geocoding by normalising spellings (e.g. *Brytaine* to *Britain*), retracing back adjectives to nouns (e.g. *Tartarian* to *Tartary*), and removing common and proper names which slipped in.

We then geocoded the list of locations for each play via the Nominatim geocoder, based on OpenStreetMap data, [8] and performed an extensive manual clean-up of the results, correcting further errors and rectifying wrong coordinates. Ultimately, we assigned every item to one of these four categories: 'city' (including elements such as roads, squares, and neighbourhoods), 'country' or 'region', 'continent', and 'natural feature' (like mountains or rivers).

3. VISUALISATION AND DISCUSSION

To present our results, a React web application was been built upon the database and is now available at https://atlas-emed.github.io. The website displays two main components: a globe map view on the left and a content view on the right (Figure 1). The map view hosts locally loaded *geojson* files that are built whilst deployment from the database file, with one *geojson* file created and loaded for each layer. When a user clicks on one of the more than 500 locations featured in the map, the app displays a list of plays associated to it in the content view (Figure 2). This takes the form of a series of "DramaCard" components, which in turn fetch and display some metadata (author name, play genre, year) from the *DraCor* API. By selecting a card, the map is updated to show only locations tied to the that text, allowing a more granular exploration of geographical references.

On a methodological level, it should be noted that the extraction conducted here is based on the spoken text of the plays, and not on the <set> element from the TEI markup. Consequently, the map does not reflect the actual setting of the works, as done e.g. in the *Database of German-Language One-Act Plays 1740–1850* (Çakir and Fischer 2022), [9] but rather attempts to register all geographic mentions within them.

Even a quick glance at this map provides some initial insights into the "cartographic imagination" (Smith 2016) of early modern English playwrights. As expected, the British peninsula – and in particular the metropolitan area of London, with its iconic alleys, squares, and roads – remains central. Two other main areas of interest coincide with the Italian and Hellenic peninsulas: this can be easily explained by the constant influence of classical (ancient Greek and Roman) motifs and tropes, and by the Italian context being a net exporter of narrative materials towards England. Accordingly, cities most often mentioned – after European powerhouses such as London and Paris – are Renaissance centres like Venice, Naples, and Florence. Smaller clusters of often-mentioned locations can be recognised also in the main western states

[4] From the official documentation: "MonadGPT is a finetune of Mistral-Hermes 2 on 11,000 early modern texts in English, French and Latin, mostly coming from EEBO and Gallica".

[5] https://spacy.io; https://stanfordnlp.github.io/stanza. We also wanted to test the state-of-art *BookNLP* package (https://github.com/booknlp/booknlp), whose entity tagging model is trained on a large dataset of literary works (Bamman et al., 2019), but it is currently not working due to a upstream dependency issue which has not yet been fixed (see https://github.com/booknlp/booknlp/pull/25).

[6] We employed here the smallest English-language model, *en_core_web_sm* (https://github.com/explosion/spacy-models/releases/tag/en_core_web_sm-3.7.1).

[7] https://openrefine.org.

[8] https://nominatim.org.

[9] https://einakter.dracor.org/locations.

(France, Spain, Germany) and in the Middle East, due to the presence of Biblical-themed contents. Nonetheless, the boundaries of the world as imagined by the *EngDraCor* authors extend way beyond Europe, stretching from the recently encountered West Indies to the near-mythological steppes of Scythia and Tartary and as far as Japan.

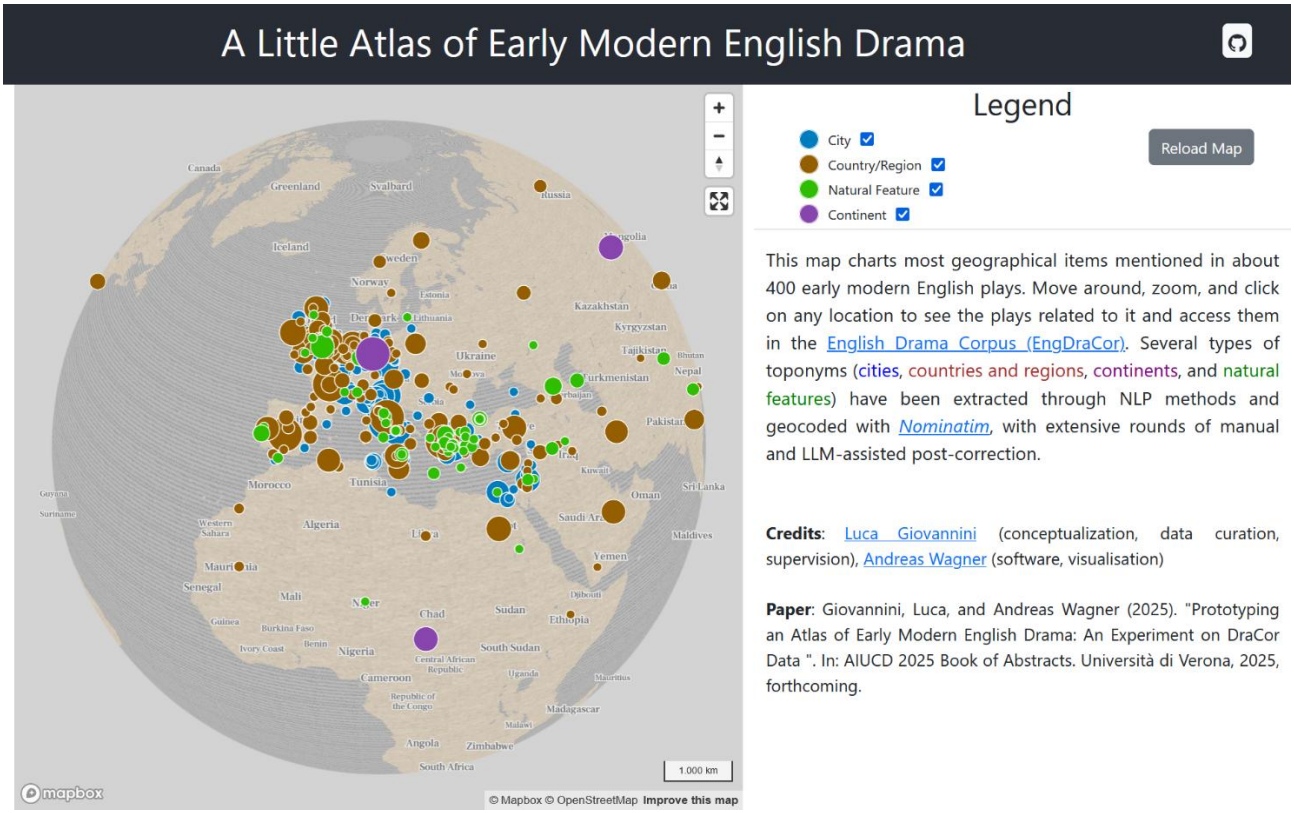


Figure 1. Homepage of the *Little Atlas of Early Modern English Drama* (atlas-emed.github.io).

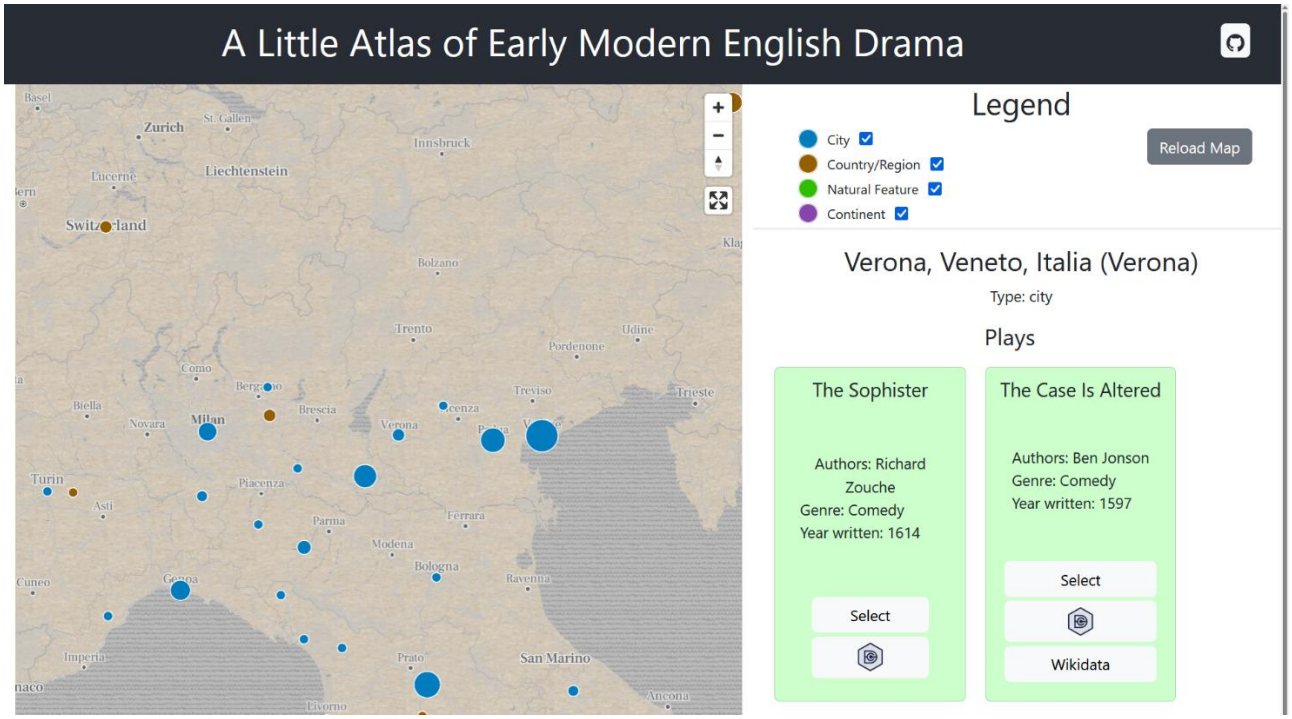


Figure 2. The *Little Atlas* displaying cards for *EngDraCor* plays where Verona is mentioned.

While the map might prove useful for digital English studies, and especially in teaching contexts, this contribution is however mostly meant to start a broader conversation about the need for tailored NLP tools in early modern literary studies. The pipeline followed here, while ultimately delivering acceptable results, was indeed too long and cumbersome to be easily redeployed on other early modern texts. The lack of more reliable and scalable solutions for NER tasks hindered a swift extraction of geographic items, and the results after the extensive manual and AI-powered post-correction were still not on par with the accuracy of entity taggers for modern languages. Given the importance of early modern sources within computational literary studies and digital humanities at large, it seems thus necessary to work towards improving NER tools for such texts. At the same time, one should also accept that, when dealing with historical geographical names, a certain degree of fuzziness is unavoidable, since one often works with entities whose boundaries have changed, evolved, or shifted over time.
From a research-oriented, *DraCor*-centric perspective, instead, the next step would naturally involve the reproduction of this experiment on the other *DraCor* corpora, which feature more recent texts that modern NLP tools may handle more effectively. In this way, it might be possible to begin investigating geographical discourse from a transnational perspective, e.g. by exploring the significance of locations which occur frequently in different dramatic traditions and at different times.

REFERENCES

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... & McGrew, B. (2023). GPT-4 Technical Report. arXiv Preprint. DOI: https://doi.org/arXiv:2303.08774.

Bamman, D., Popat, S., & Shen, S. (2019, June). An Annotated Dataset of Literary Entities. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (pp. 2138–2144).

Baron, A., & Rayson, P. (2008). VARD 2: A Tool for Dealing with Spelling Variation in Historical Corpora. In Proceedings of the Postgraduate Conference in Corpus Linguistics. Aston University, Birmingham, UK. URL: https://ucrel.lancs.ac.uk/people/paul/publications/BaronRaysonAston2008.pdf.

Bradbury, M. (Ed.). (1996). The Atlas of Literature. De Agostini.

Fischer, F., Börner, I., Göbel, M., Hechtl, A., Kittel, C., Milling, C., & Trilcke, P. (2019). Programmable Corpora: Introducing DraCor, an Infrastructure for the Research on European Drama. In Proceedings of DH2019: "Complexities", Utrecht, July 9–12, 2019. Utrecht University. DOI: https://doi.org/10.5281/zenodo.4284002.

Gregory, I. N., & Hardie, A. (2011). Visual GISting: Bringing Together Corpus Linguistics and Geographical Information Systems. Literary and Linguistic Computing, 26(3), 297–314. DOI: https://doi.org/10.1093/llc/fqr022.

Humbel, M., Nyhan, J., Vlachidis, A., Sloan, K., & Ortolja-Baird, A. (2021). Named-Entity Recognition for Early Modern Textual Documents: A Review of Capabilities and Challenges with Strategies for the Future. Journal of Documentation, 77, 1223–1247. https://doi.org/10.1108/JD-02-2021-0032.

Jakacki, D. K. (2024). Review: World Historical Gazetteer. Reviews in Digital Humanities, 5(5). https://doi.org/10.21428/3e88f64f.6bceb2bf.

Keraghel, I., Morbieu, S., & Nadif, M. (2024). A survey on recent advances in named entity recognition. arXiv preprint. DOI: https://doi.org/10.48550/arXiv.2401.10825.

Langlais, P. C. (2023). MonadGPT. URL: https://huggingface.co/Pclanglais/MonadGPT.

Manjavacas, E., & Fonteyn, L. (2022). Adapting vs. Pre-Training Language Models for Historical Languages. Journal of Data Mining & Digital Humanities NLP4DH. DOI: https://doi.org/10.46298/jdmdh.9152.

Mitchell, P. (2017). Literary Geography and the Digital: The Emergence of Neogeography. In The Routledge Handbook of Literature and Space (pp. 85–94). Routledge. https://doi.org/10.4324/9781315745978-8.

Moretti, F. (1999). Atlas of the European Novel: 1800–1900. Verso.

Mueller, M., and the Early Print team. (2012). EarlyPrint. URL: https://earlyprint.org.

Perenič, U. (2014). An Overview of Literary Mapping Projects on Cities: Literary Spaces, Literary Maps and Sociological (Re)Conceptualisations of Space. Neohelicon, 41(1), 13–25. DOI: https://doi.org/10.1007/s11059-013-0226-5.

Smith, D. K. (2016). The Cartographic Imagination in Early Modern England: Re-Writing the World in Marlowe, Spenser, Raleigh and Marvell. Routledge.

ATLAS: A data model for describing FAIR Digital Humanities research outcomes

Chiara Martignano[1], Giorgia Rubin[2], Sebastiano Giacomini[3], Alessia Bardi[4], Marina Buzzoni[5], Marilena Daquino[6], Riccardo Del Gratta[7], Angelo Mario Del Grosso[8], Franz Fischer[9], Roberto Rosselli Del Turco[10], Francesca Tomasi[11]

[1] Università Ca' Foscari Venezia, Italia – chiara.martignano@unive.it
[2] CNR-ILC, Italia – giorgia.rubin@ilc.cnr.it
[3] Università di Bologna, Italia – sebastiano.giacomin2@unibo.it
[4] CNR-ISTI, Italia – alessia.bardi@isti.cnr.it
[5] Università Ca' Foscari Venezia, Italia – mbuzzoni@unive.it
[6] Università di Bologna, Italia – marilena.daquino2@unibo.it
[7] CNR-ILC, Italia – riccardo.delgratta@ilc.cnr.it
[8] CNR-ILC, Italia - angelo.delgrosso@ilc.cnr.it
[9] Università Ca' Foscari Venezia, Italia - franz.fischer@unive.it
[10] Università di Torino, Italia - roberto.rossellidelturco@unito.it
[11] Università di Bologna, Italia – francesca.tomasi@unibo.it

ABSTRACT (ENGLISH)

This paper addresses the challenges of cataloguing and representing Digital Humanities (DH) research outputs within the framework of FAIR principles. Despite advancements in Semantic Web technologies and data aggregators, the scholarly community still lacks unified frameworks and domain-specific models to describe heterogeneous outputs such as digital editions, textual collections, and other scholarly resources published as Linked Open Data. The ATLAS project proposes an ontology and a knowledge graph to bridge these gaps. The methodology includes metadata modeling based on existing frameworks and novel extensions, supported by pilot studies on the Italian cultural heritage. The project, currently in its initial version, aims to enhance metadata interoperability and data accessibility, contributing to the optimization of cataloguing practices and the development of guidelines for the discovery and reuse of DH resources while providing a model applicable beyond the Italian context.

Keywords: metadata modeling; research infrastructure; scholarly outcome; linked open data; FAIR principles

ABSTRACT (ITALIANO)

ATLAS: un modello di dati per descrivere i risultati della ricerca nelle Digital Humanities secondo i principi FAIR. L'articolo discute le sfide poste dalla catalogazione e rappresentazione degli output di ricerca nel campo delle Digital Humanities (DH) in conformità ai principi FAIR. Nonostante i progressi nelle tecnologie del Web Semantico e nei sistemi di aggregazione dei dati, la comunità accademica continua a mancare di framework unificati e modelli specifici per descrivere output eterogenei come edizioni digitali, collezioni testuali e altre risorse scientifiche pubblicate come Linked Open Data. Il progetto ATLAS propone un'ontologia e un knowledge graph per colmare queste lacune. La metodologia integra modellazione dei metadati basata su framework esistenti ed estensioni innovative, supportata da studi pilota sul patrimonio culturale italiano. Il progetto, che al momento si presenta nella sua prima versione, punta a migliorare l'interoperabilità dei metadati e l'accessibilità dei dati, contribuendo a ottimizzare le pratiche di catalogazione e a sviluppare linee guida per la scoperta e il riutilizzo delle risorse DH offrendo un modello applicabile oltre il contesto italiano.

Parole chiave: modellazione di metadati; infrastruttura di ricerca; prodotto scientifico; linked open data; principi FAIR

1. INTRODUCTION[1]

The increasing adoption of World Wide Web technologies in Humanities research has, in recent years, profoundly influenced the way scholarly research is performed, particularly within the field of Digital Humanities (DH) (Tomasi, 2022). Such advancements have created new opportunities for preserving,

[1] Authors responsibility: Chiara Martignano is responsible for section 3 and 4; Giorgia Rubin is responsible for section 2 and 5; all authors contributed to section 1.

sharing, and reusing research outputs, enabling unprecedented levels of collaboration and dissemination. However, the rapid increase of available scholarly data highlights the critical need for standardized models and guidelines to effectively manage, aggregate, and explore this wealth of meta-information.
In the scholarly landscape, several platforms play a pivotal role in ensuring the persistent identification, preservation, and enhanced accessibility of research data (PARTHENOS et al., 2018). Examples of key initiatives include Zenodo, OpenAIRE (Baglioni et al. 2019), and Research Infrastructures (RIs) such as CLARIN and DARIAH. Zenodo serves as a repository for documents and various research materials, while OpenAIRE offers a comprehensive range of services, including dedicated web portals for specific communities, such as the Digital Humanities and Cultural Heritage Gateway. In contrast, RIs combine both the storage of documents and research materials with the capability to access services that utilize and generate these resources. These initiatives pave the way for enhancing the sharing of research outcomes and promoting the advancement of open science practices.
Likewise, various catalogues document research in the field of DH, encompassing digital scholarly editions and projects from both national[2] and international[3] associations and research centers. However, there is still an absence of comprehensive catalogues dedicated specifically to DH projects concerning Italian Cultural Heritage, as well as a lack of structured collections for DH research outputs employing Semantic Web technologies (Tomasi, 2013).
The ATLAS[4] project aims to create a knowledge graph of DH research related to Italian Digital Cultural Heritage. By introducing the ATLAS Ontology and its knowledge graph, the project aims to create a semantic framework capable of representing the varied outputs of DH research, also beyond the Italian borders. ATLAS tackles the challenges of describing and interlinking scholarly data—including the raw text to perform full text analysis which is of primary importance for—ensuring enriched and accessible metadata to enhance both discoverability and reusability of these cultural assets. By thoroughly analyzing and mapping existing models and vocabularies, as well as surveying representative projects in the Italian DH landscape, the project introduces a reliable strategy for incorporating Italian DH resources into the global knowledge ecosystem, thereby promoting enhanced interdisciplinary collaboration and resource discovery across institutions.

2. STATE OF THE ART

In Italy, considerable advancements have been made in the digitization and aggregation of cultural heritage using Linked Open Data (LOD) collections. Prominent examples include the *dati.culturaitalia* (Di Giorgio, 2015) platform developed by the Italian Ministry of Culture and the ArCO project (Carriero et al., 2019), which has built a Knowledge Graph derived from the General Catalog of Italian Cultural Heritage. These initiatives are in line with European digitization projects such as ARIADNE (Meghini et al. 2017) and Europeana (Di Giorgio, 2016). Although these programs offer interoperable Linked Open Data, a gap still remains in 1) the construction of a catalogue of DH projects leveraging Italian Cultural Heritage, and 2) a research framework to support best practices and the enhancement of the discoverability and reusability of Italian heritage-related DH data (Carriero et al., 2019).
Existing ontologies and models fail to adequately capture the complexities of the contemporary Digital Humanities landscape. Most models address a limited range of research products types, mostly focusing on articles and publications. OpenAIRE's data model, for example, includes four categories of research products: publication, data, software and other. DH projects generate a wide range of outputs (e.g., textual archives, digital scholarly editions, digital collections, etc.), each requiring tailored descriptive strategies. Key factors, such as textual typologies and editorial criteria, are not sufficiently addressed. Moreover, current models lack effective mechanisms for linking research activities to their related Cultural Heritage objects, despite the potential provided by Linked Open Data (Daquino et al., 2024b).

3. RESEARCH QUESTIONS AND METHODOLOGY

Considering the current landscape of research product catalogues, we assessed strategies for the optimisation of cataloguing practices for DH projects based on the Italian digital cultural heritage. Our research focused on three key questions: What types of research products exist? How can we represent

[2] AIUCD's list of DH projects: https://www.aiucd.it/progetti/.

[3] For example, EADH's list of DH projects: https://eadh.org/projects.

[4] ATLAS is a project funded by the Next Generation program of the European Commission for 24 months (October 2023 - October 2025).

different types of research products in a way that highlights their distinctive features? Which metadata should we employ to ensure long-term preservation and improve the findability of research products?
In the first phase, we identified pilot research products related to Italian cultural heritage or developed by Italian scholars to determine the most suitable metadata for the catalogue. These pilots were selected as key references in the Italian DH landscape and span five categories, namely:

1. Text collections: ALIM (Archive of the Italian Latinity of the Middle Ages, D'Angelo & Monella, 2019); Biblioteca Italiana (Quondam, 2021); BUP - Digital Humanities (Amendola, 2021); Musisque Deoque (Venuti et al., 2023).
2. Digital Scholarly Editions: VaSto (VArchi STOria fiorentina, Brancato et al., 2021); Codice Pelavicino Digitale (Salvatori et al., 2017); Leges Langobardorum (Buzzoni & Del Turco, 2015); Digital Edition of Aldo Moro's works (Moro, 2021).
3. Linked Open Data: Zeri & LODE (Daquino et al., 2017); DanteSources (Bartalesi et al., 2015); LiLa - Linking Latin (Passarotti, 2022); Biflow - Toscana Bilingue Catalogue (Montefusco & Mancinelli, 2020).
4. Ontologies: CIDOC-CRM; SPAR (Peroni & Shotton, 2018); HiCO (Daquino & Tomasi, 2015).
5. Software tools: EVT (Edition Visualisation Technology, Del Turco et al., 2019); Voyant Tools (Sinclair & Rockwell, 2015).

For each pilot, at least one ATLAS project member was directly involved in its creation, ensuring complete and accurate information about the pilots' current state and development process.
Our analysis of the pilot research products yielded several key findings. First, we identified both common and category-specific metadata for use in the catalogue. Second, we uncovered critical issues affecting data usability and long-term preservation. Common issues across research products included: lack of data storage in "trustworthy repositories" (PARTHENOS et al., 2018) such as Zenodo, unclear dataset access points and methods, missing information about dataset status (e.g., completed, under development), unavailable data models and references to existing standards, and insufficient documentation about usage, applied methodologies, and used technologies.
Based on the identified issues and on existing guidelines for data FAIRness (Wilkinson et al., 2016; PARTHENOS et al., 2018), we developed a set of recommendations and best practices specific to each research product type.[5]
We then refined the identified metadata fields through mapping (Daquino et al., 2024a) with major existing models for describing research products, specifically: RO-Crate,[6] KNOT,[7] OpenAIRE Graph,[8] OpenAIRE Application Profile,[9] SKG-IF,[10] IRIS.[11] To translate the metadata into RDF properties, we primarily used Schema.org.[12]

[5] Recommendations and best practices are available in the ATLAS whitebook: https://doi.org/10.5281/zenodo.14925266.

[6] RO-Crate (Research Object Crate) is a method of aggregating and describing research data with associated metadata expressed through JSON-LD using linked data. https://w3id.org/ro/crate/1.1.

[7] KNOT Ontology (KNOT-O) describes digital scholarly activity and objects as examples of the digital cultural heritage of Italian universities. http://purl.org/knot/ontology.

[8] OpenAIRE Graph is one of the largest open scholarly record collections worldwide, key in fostering Open Science and establishing its practices in the daily research activities. https://graph.openaire.eu/docs/.

[9] OpenAIRE Application Profile describes how OpenAIRE enforces specific encoding schemes for the values of some DataCite properties. https://guidelines.openaire.eu/en/latest/data/application_profile.html.

[10] SKG-IF (Scientific Knowledge Graphs Interoperability Framework) enables the exchange of data about six core entities and their relationships (Research product, Agent, Grant, Venue, Topic, Data source). https://skg-if.github.io/.

[11] IRIS (Institutional Research Information System) is a Java-based platform for managing and enhancing research outputs adopted by numerous Italian universities (Bollini et al., 2016).

[12] Schema.org describes a data model to create, maintain, and promote schemas for structured data on the Internet. https://schema.org/docs/datamodel.html.

These initial phases produced two main outcomes: a data model formalized as an OWL 2 DL ontology, i.e., the ATLAS ontology (Tomasi et al., 2024), and the first version of the knowledge graph[13] (Daquino et al., 2024a), accessible online through an extended version of the CLEF[14] software (Daquino et al., 2023). In the data model (Figure 1), research products are modeled as `schema:Dataset`. Different types of research products are implemented as subclasses of `schema:Dataset` and aligned with subclasses of `frbr:Expression` from the FaBiO ontology—for example, `fabio:ComputerProgram` for software tools. Each research product can be linked to a research project, represented by the class `schema:ResearchProject`, along with representations of people, organizations, websites, and computer programs.

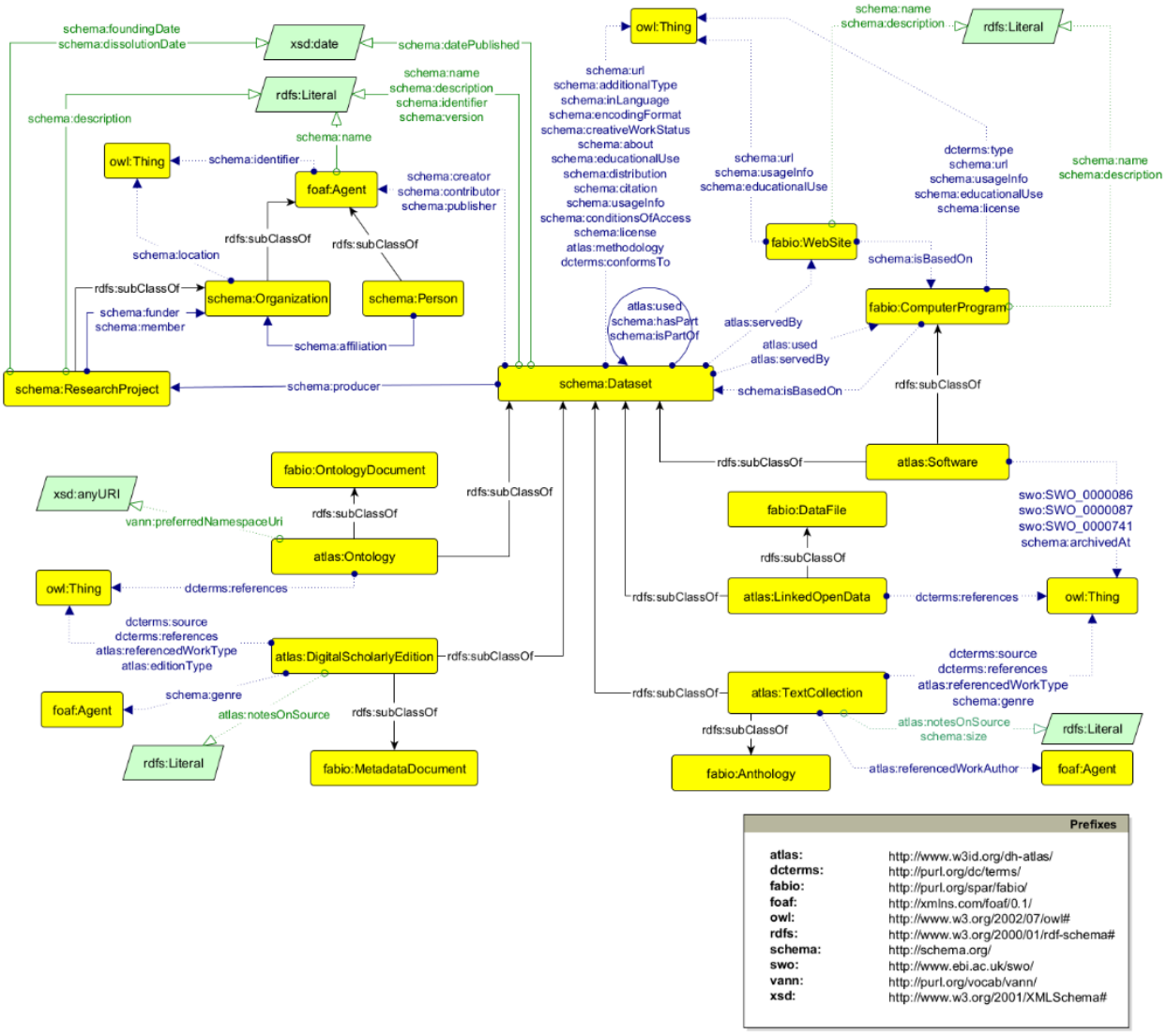


Figure 1. A visual diagram of the ATLAS ontology

The web application of ATLAS serves as a testing and evaluation platform for the ontology through an application-based validation approach. The graph currently contains the select group of research products mentioned above, together with their associated research projects, websites, people, and organizations.

4. METADATA ANALYSIS

Traditional catalogues present textual archives and (digital) scholarly editions primarily through content-focused editorial information—which remains essential and is included in our model. A novel aspect of the

[13] The first version of the ATLAS Knowledge Graph can be consulted in the ATLAS platform (https://projects.dharc.unibo.it/atlas/).

[14] The ATLAS catalogue is built on CLEF v3.0, https://polifonia-project.github.io/clef/.

ATLAS modeling approach is that it describes digital scholarly editions and text collections primarily as datasets, while also emphasizing features and methodologies specific to the "digital paradigm" (Sahle, 2016) that distinguish these research products from their corresponding print-oriented versions.

Some metadata selected for describing research products are basic and common across most existing models. The mandatory metadata in our model include: *title* (`schema:name`), *description* (`schema:description`), *creator* (`schema:creator`), *publisher* (`schema:publisher`), *release date* (`schema:datePublished`), *landing page* (`schema:url`), *access rights* (`schema:conditionsOfAccess`), and *license* (`schema:license`). We have added two additional mandatory properties: *type* (`schema:additionalType`) to specify the research product type, and *research activities* (`schema:educationalUse`), which uses TaDiRAH vocabulary terms (Borek et al, 2020) to describe the activities enabled by the research product, such as analysis, visualization, and modeling.

Our model introduces several new properties compared to the existing models mentioned above. These include "research activities," *status* (`schema:creativeWorkStatus`) which describes the research product's current lifecycle state, and *documentation* (`schema:usageInfo`) containing the URL to the output's documentation. While many existing models include data encoding format—represented in our model as *format* (`schema:encodingFormat`)—we have added *metadata standards* (`dcterms:conformsTo`) to indicate the models and standards used for metadata modeling. The *access point* property (`atlas:servedBy`) complements the "landing page" concept by specifically presenting web addresses for dataset access, distinct from the presentation pages. We have also added *academic field* (`schema:about`) to indicate disciplinary areas, and *methodology* (`atlas:methodology`) alongside *software reuse* (`atlas:used`) to describe development processes, including specific activities and tools used.

For research products containing sub-products, such as collections within a digital archive, collections can be described separately as independent text collections and linked to the main output using the `has part` property and its inverse `is part of`.

While mandatory properties ensure basic identification and usage information, optional properties highlight specific information for research product reuse that is typically hard to find or absent in the products' documentation.

Each research product type provides additional specific metadata beyond the properties common across different types. For text collections and digital scholarly editions we have included traditional cataloguing metadata to describe content: *work* (`dcterms:source`), *author* (`atlas:referencedAuthor`), and *genre* (`schema:genre`). Beyond these "work" level properties (IFLA's Library Reference Model, Riva et al., 2020), we have added properties describing "Items" or "Manifestations"—the documents, attestations, and witnesses used by editors: *reference to the edited text* (`dcterms:references`), and *bibliographic reference of edited text*. This enables future catalogue users to filter search results to view different editions of the same textual resource. We have also created properties to specify the type of digital scholarly editions (`atlas:editionType`). Since no comprehensive vocabularies existed for edition types, we developed new values based on the *Parvum Lexicon Stemmatologicum* (Roelli & Macé, 2015) and the *Lexicon of Scholarly Editing* (Dillen, 2020). It is possible to indicate the quantity of individual texts or tokens (`schema:size`) for text collections.

To showcase model and standard reuse within the ATLAS catalogue, ontologies and linked open data can use `schema:references` to specify imported models and RDF ontologies used in data modeling. Ontologies can also indicate their *namespaces* (`vann:preferredNamespacePrefix`, Davis, 2005).

Software tools can be catalogued both as research products and as components of other research products' development and access methods. The *based on* property (`schema:isBasedOn`) specifies libraries, extensions, and components used in development. To facilitate workflow creation across tools, we include properties for *input* and *output formats* (`swo:hasSpecifiedDataInput` and `swo:hasSpecifiedDataOutput`, Lister et al., 2023). We have also added a code repository *URL* property (`schema:archivedAt`) to encourage code reuse.

5. NEXT STEPS AND FINAL REMARKS

The ATLAS ontology leverages and builds upon established models for describing digital cultural heritage, providing a comprehensive framework with carefully selected terminology and granular detail levels. This approach enables precise descriptions of the diverse and unique characteristics found across different types of research outputs within the Digital Humanities field. Additionally, the ontology facilitates detailed analysis of the methodologies employed in creating these research outputs, offering valuable insights into the research process itself.

The initial version of the ontology and the platform were recently tested through a datathon[15] during the ATLAS Workshop.[16] A group of scholars, researchers, and PhD students participated in the cataloguing process, providing valuable feedback and insights. Each participant catalogued a Research Product from international scholarly research on Italian Digital Cultural Heritage. This phase served the dual purpose of populating the database and evaluating the ATLAS platform's usability. The datathon successfully added 60 new records to the knowledge graph.

The ontology needs additional refinement in key areas to ensure its comprehensiveness, namely, 1) expanding the research product categories to address the variety of data sources, 2) enhancing the ATLAS controlled vocabularies to address identified gaps in the current state of the art, and 3) representing extracted data—such as people, places, and organizations—from source files to enable full-text searches. Once the model is consolidated we plan to test the ontology and expand the knowledge graph by including additional research projects and their diverse outputs, thus enriching the ontology in a bottom-up approach. A second datathon will help us to validate the model's final version.

To ensure long-term accessibility and preservation, the model, alongside regular snapshots of the knowledge graph and of the web application, will be deposited in the CLARIN national repository hosted at ILC-CNR. Furthermore, the knowledge graph will be integrated into the OpenAIRE Research Graph and, therefore into the European Open Science Cloud.

The model outlined above will therefore prove effective in representing not only the Italian Cultural Heritage but also research outcomes in the global Digital Humanities landscape.

ACKNOWLEDGEMENTS

PRIN 2022 project funded by the European Union - Next Generation EU, Mission 4 Component 1 CUP J53D23013000006.

REFERENCES

Amendola, C. (2021). Editoria universitaria, open access e nuove frontiere del lavoro umanistico: la Basilicata University Press e la collana "Digital Humanities". DigItalia, 16(2), 135–142. https://doi.org/10.36181/digitalia-00042.

Baglioni, M., Bardi, A., Kokogiannaki, A., Manghi, P., Iatropoulou, K., Principe, P., Vieira, A., Nielsen, L. H., Dimitropoulos, H., Foufoulas, I., Manola, N., Atzori, C., La Bruzzo, S., Lazzeri, E., Artini, M., De Bonis, M., & Dell'Amico, A. (2019). The OpenAIRE Research Community Dashboard: On Blending Scientific Workflows and Scientific Publishing. Zenodo. https://doi.org/10.5281/zenodo.3467104.

Bartalesi, V., Meghini, C., Andriani, P., & Tavoni, M. (2015). Towards a semantic network of Dante's works and their contextual knowledge. Digital Scholarship in the Humanities, 30(suppl_1), i28–i35. https://doi.org/10.1093/llc/fqv044.

Brancato, D., Corbellini, M., Italia, P., Pasqual, V., Priore, R. (2021). "VaSto: un'edizione digitale interdisciplinare". magazén, 2(1), 139-170 http://doi.org/10.30687/mag/2724-3923/2021/03/006.

Bollini, A., Mennielli, M., Mornati, S., & Palmer, D. T. (2016). IRIS: Supporting & Managing the Research Life-cycle. https://doi.org/10.13189/ujer.2016.040410.

Borek, L., Hastik, C., Khramova, V., & Geiger, J. (2020). TaDiRAH: Taxonomy of Digital Research Activities in the Humanities (Version 2.0.1) [Turtle]. https://vocabs.dariah.eu/tadirah/.

Buzzoni, M., Del Turco, R. Rosselli, (2015). Verso un'edizione digitale dell'Editto di Rotari, in *I Longobardi in Italia: lingua e cultura*, a cura di C. Falluomini, Alessandria, 2015, pp. 37-85. https://hdl.handle.net/10278/3662705.

[15] A datathon is a data science or data-centric hackathon, i.e., an event where participants work together intensively on data-related tasks.

[16] https://dh-atlas.github.io/workshop.html.

Carriero, V. A., Gangemi, A., Mancinelli, M. L., Marinucci, L., Nuzzolese, A. G., Presutti, V., & Veninata, C. (2019). ArCo: The Italian Cultural Heritage Knowledge Graph. In *The Semantic Web – ISWC 2019* (pp. 36–52). Springer International Publishing. https://doi.org/10.1007/978-3-030-30796-7_3.

CIDOC-Conceptual Reference Model. https://cidoc-crm.org.

D'Angelo, E., & Monella, P. (2019). ALIM (Archivio della Latinità Italiana del Medioevo): Storia, attualità, prospettive di una banca-dati di testi mediolatini. In Filologia medievale: comparatistica, critica del testo e attualità: Atti del convegno (Viterbo, 26-28 settembre 2018) (pp. 203–223). Roma: L'Erma di Bretschneider. http://digital.casalini.it/46342.

Davis, I. (2005). VANN: A vocabulary for annotating vocabulary descriptions [Dataset]. https://vocab.org/vann/

Daquino, M., Giacomini, S., Bardi, A., Buzzoni, M., Del Gratta, R., Del Grosso, A. M., Del Turco, R. R., Fischer, F., Martignano, C., Rubin, G., & Tomasi, F. (2024a). DH ATLAS (Version v1.0.1) [Dataset]. Zenodo. https://doi.org/10.5281/zenodo.13993057

Daquino, M., Bardi, A., Buzzoni, M., Del Gratta, R., Del Grosso, A. M., Fischer, F., Tomasi, F., & Rosselli Del Turco, R. (2024b). The ATLAS: A Knowledge Graph of Digital Scholarly Research on Italian Cultural Heritage. In A. Di Silvestro & D. Spampinato (Eds.), *Me.Te. Digitali. Mediterraneo in rete tra testi e contesti, Proceedings del XIII Convegno Annuale AIUCD2024* (pp. 588–592). https://doi.org/10.6092/unibo/amsacta/7927.

Daquino, M., Wigham, M., Daga, E., Giagnolini, L., & Tomasi, F. (2023). CLEF. A Linked Open Data Native System for Crowdsourcing. *J. Comput. Cult. Herit.*, 16(3), 41:1-41:17. https://doi.org/10.1145/3594721

Daquino, M., Mambelli, F., Peroni, S., Tomasi, F., Vitali, F. (2017). Enhancing Semantic Expressivity in the Cultural Heritage Domain: Exposing the Zeri Photo Archive as Linked Open Data. Association for Computing Machinery. https://doi.org/10.1145/3051487.

Daquino, M., & Tomasi, F. (2015). "Historical Context Ontology (HiCO): A Conceptual Model for Describing Context Information of Cultural Heritage Objects." Communications in Computer and Information Science. Springer International Publishing. https://doi.org/10.1007/978-3-319-24129-6_37.

Del Turco, R. Rosselli, Di Pietro, C., & Martignano, C. (2019). Progettazione e implementazione di nuove funzionalità per EVT 2: lo stato attuale dello sviluppo. Umanistica Digitale, 3(7). https://doi.org/10.6092/issn.2532-8816/9322.

Dillen, W. (2020). *Lexicon of Scholarly Editing*. https://lexiconse.uantwerpen.be/.

Di Giorgio, S. (2016). Gli archivi del MiBACT. L'integrazione dei dati archeologici digital. In P. Ronzino (Ed.), *Proceedings del Workshop L'integrazione dei dati archeologici digitali - Esperienze e prospettive in Italia (Lecce, Italia, 1-2 ottobre, 2015)* (pp. 47–55).

Di Giorgio, S. (2015). DATI.CULTURAITALIA.IT. Un progetto pilota dedicato ai dati aperti e ai Linked Open Data. Archeologia e Calcolatori, Supplemento 7, 103-106 http://www.archcalc.cnr.it/indice/Suppl_7/13_Di%20Giorgio.pdf.

European Organization For Nuclear Research, OpenAIRE (2013). Zenodo. CERN. https://doi.org/10.25495/7GXK-RD71.

Frosini, L., Bardi, A., Manghi, P., & Pagano, P. (2018). An Aggregation Framework for Digital Humanities Infrastructures: The PARTHENOS Experience. *SCIRES-IT - SCIentific RESearch and Information Technology*, 8, 33–45. https://doi.org/10.2423/i22394303v8n1p33.

Lister, A., Brown, A., Hull, D., Parkinson, H., Malone, J., Ison, J., Badarinarayan, N., & Stevens, R. (2023). Software Ontology (SWO) (http://www.ebi.ac.uk/swo/swo.owl; Version 2023-03-05) [Dataset].

Meghini, C., Scopigno, R., & Richards, J. (2017). ARIADNE: a research infrastructure for archaeology. https://doi.org/10.5281/zenodo.2552101.

Montefusco, A., & Mancinelli, T. (2020). The BIFLOW-Toscana Bilingue Catalogue: A Digital Representation of the Socio-cultural History of Translation in the Tuscan Middle Ages (1260–1430). https://doi.org/10.14434/textual.v13i2.31597.

Moro, A., (2021). Edizione Nazionale delle Opere di Aldo Moro, voll., Bologna, Università di Bologna. ISBN: 9788854970496; DOI: https://doi.org/10.6092/unibo/aldomoro.

OpenAIRE CONNECT Gateway on Digital Humanities and Cultural Heritage: https://dh-ch.openaire.eu.

OWL 2 Web Ontology Language Document Overview (Second Edition). https://www.w3.org/TR/2012/REC-owl2-overview-20121211/

PARTHENOS, Hollander, H., Morselli, F., Uiterwaal, F., Admiraal, F., Trippel, T., & Di Giorgio, S. (2018). PARTHENOS Guidelines to FAIRify data management and make data reusable. https://doi.org/10.5281/zenodo.2668479.

Passarotti, M. (2022). Il latino sbarca nel Semantic Web: La Knowledge Base LiLa. Culture digitali, 4. https://doi.org/10.5281/zenodo.6479137.

Peroni, S., & Shotton, D. (2018) "The SPAR Ontologies." Lecture Notes in Computer Science. Springer International Publishing. https://doi.org/10.1007/978-3-030-00668-6_8.

Quondam, A. (2021). Memorie per una storia dell'italianistica digitale: «Biblioteca italiana». Griseldaonline, 20(2), 137–147. https://doi.org/10.6092/issn.1721-4777/12360.

Riva, P., Le Boeuf, P., & Žumer, M. (2020). *IFLA Library Reference Model: Un modello concettuale per le informazioni bibliografiche*. https://repository.ifla.org/handle/123456789/44.

Roelli, P., & Macé, C. (2015, November 13). *Parvum Lexicon Stemmatologicum*. https://wiki.helsinki.fi/xwiki/bin/view/stemmatology/Parvum%20lexicon%20stemmatologicum/.

Sahle, P. (2016). What is a Scholarly Digital Edition? In M. J. Driscoll & E. Pierazzo (Eds.), *Digital Scholarly Editing* (1st ed., Vol. 4, pp. 19–40). Open Book Publishers. https://www.jstor.org/stable/j.ctt1fzhh6v.6.

Salvatori, E., Rosselli Del Turco, R., Alzetta, C., Di Pietro, C., Mannari, C., & Miaschi, A. (2017). Il Codice Pelavicino tra edizione digitale e Public History. Umanistica Digitale, 1(1). https://doi.org/10.6092/issn.2532-8816/7232.

Sinclair, S., & Rockwell, G. (2015). Text Analysis and Visualization. In A New Companion to Digital Humanities (pp. 274–290). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118680605.ch19.

Tomasi, F., Daquino, M., & Giacomini, S. (2024). The ATLAS Ontology (Version 1.0) [Dataset]. http://www.w3id.org/dh-atlas/.

Tomasi, F. (2022). *Organizzare la conoscenza: Digital Humanities e Web semantico*. Editrice Bibliografica. https://doi.org/10.53134/9788893573573.

Tomasi, F. (2013). *Vespasiano Da Bisticci Letters*. http://vespasianodabisticciletters.unibo.it. https://doi.org/10.6092/unibo/vespasianodabisticciletters.

Venuti, M., Del Grosso, A. M., Boschetti, F., Tessarolo, L., Prontera, A., Bovet, D., Cattaneo, G., & Melis, V. (2023). La 'Galassia MQDQ:' un concetto di filologia tradizionale, digitale, sostenibile. Magazèn, 4(1), 50. https://dx.doi.org/10.30687/mag/2724-3923/2023/07/003.

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1), Article 1. https://doi.org/10.1038/sdata.2016.18.

Cantautorato e *Digital Humanities*. Per una valorizzazione dell'opera di Fabrizio De André, Lucio Dalla, Gianmaria Testa

Marcello Ranieri[1]
[1] Alma Mater Studiorum - Università di Bologna, Italia – marcello.ranieri2@unibo.it

ABSTRACT (ITALIANO)
La produzione artistica di famosi cantautori italiani è ampiamente rappresentata nel web. Anche nel semantic web essi sono presenti, ma per dare un'immagine adeguata della loro creatività sarebbe necessario fornire informazioni più complete. Normalmente, i dati disponibili nelle piattaforme della conoscenza per ciascuno di questi artisti costituiscono un semplice segnaposto. Il numero di record recuperabile negli OPAC testimonia di una quantità di pubblicazioni discografiche che supera di gran lunga quello che riportano i più noti aggregatori. Questo contributo presenta un metodo, applicato ai casi di Fabrizio De André, Lucio Dalla e Gianmaria Testa, per convertire in Linked Open Data pronti per la pubblicazione i record bibliografici in formato MARC tratti dall'OPAC del Servizio Bibliotecario Nazionale, ponendo particolare attenzione alla riconciliazione dei dati.
Parole chiave: cantautori; discografia; digital library; Linked Open Data; web semantico

ABSTRACT (ENGLISH)
The "cantautorato" and the Digital Humanities. For a valorization of the work by Fabrizio De André, Lucio Dalla, Gianmaria Testa
The artistic production of well-known Italian songwriters is widely represented on the Web. They are on the Semantic Web as well, but to adequately acknowledge their creativity it would be necessary to provide more complete information. Normally, the data available for each of those artists on knowledge platforms serves merely as a placeholder. The number of records available in OPACs bears witness to a quantity of published discography that far exceed what most famous aggregators register. This contribution presents a method, applied to the cases of Fabrizio De André, Lucio Dalla e Gianmaria Testa, for converting to Linked Open Data ready for publication the MARC formatted bibliographic records downloaded from the OPAC of the National Library Service, with particular attention to data reconciliation.
Keywords: songwriters; discography; digital library; Linked Open Data; Semantic Web

1. INTRODUZIONE
Le opere[1] del cantautorato italiano possono a pieno titolo essere considerate parte del patrimonio culturale per originalità di contenuti, profondità di ricerca musicale, attenzione ai temi sociali, popolarità trasversale.



Figura 1. Immagine dal sito del Ministero della Cultura: la presentazione della "Casa dei Cantautori"

[1] Va qui chiarito che la parola *opera* viene usata nel senso di *produzione artistica* che ha comunemente nell'italiano standard e non in quello tecnico assunto nello schema concettuale IFLA FRBR.

Un riconoscimento ufficiale dell'importanza di questa manifestazione creativa nella musica italiana si è avuto con la presentazione a Genova nel 2021 della "Casa dei Cantautori" (vedi Fig. 1), un luogo della cultura dedicato al cantautorato e collocato in una delle città che ne hanno visto le origini.[2] La canzone d'autore è fiorita nel Secondo Dopoguerra, ha raggiunto un altissimo livello negli anni Sessanta del Novecento e lo ha mantenuto costantemente evolvendosi parallelamente e in un certo senso in opposizione (Porciani, 2020: 379) al gusto corrente della musica leggera, rispecchiando allo stesso tempo i mutamenti del costume (Ciofalo, 2022: 335). Molti nomi rientrerebbero nel lungo elenco degli autori di tali testi e musiche da allora ai giorni nostri, permettendo la classificazione delle più o meno celebri canzoni prodotte in svariati sottogeneri. I testi dei cantautori sono stati e sono un mezzo inclusivo di circolazione delle idee e hanno sempre prestato una particolare attenzione alle categorie socialmente marginali e svantaggiate. Al di là di ogni considerazione storico-critica, valutando se l'opera dei cantautori abbia rappresentanza anche nella conoscenza trasposta in modelli semantici, non si può non notare come a una pervasiva presenza della produzione artistica cantautoriale nel web non corrisponda una altrettanto completa rappresentazione di essa sotto forma di Linked Open Data. Si prenda a titolo di mero esempio il caso di tre dei più noti e amati tra i cantautori italiani degli ultimi decenni: Fabrizio De André (1940-1999), Lucio Dalla (1943-2012), Gianmaria Testa (1958-2016). Se per ciascuno di essi si verifica la presenza delle rispettive opere nei due noti e importanti luoghi di diffusione di contenuti culturali Europeana e Wikidata,[3] si troveranno i risultati seguenti. Su Europeana, per De andré i documenti disponibili risultano 235, per Dalla 139, per Testa 2, come è evidente nella Fig. 2. Su Wikidata, a parte dati di carattere generale per ciascuno dei tre autori considerati e i loro identificatori nei principali social network, la sezione discografica è piuttosto esigua.

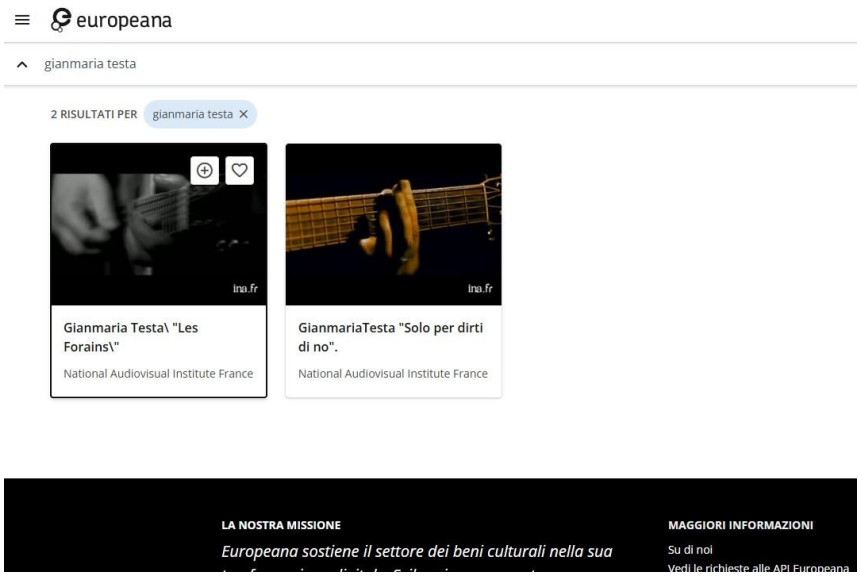


Figura 2. Risultati di una ricerca su Europeana

Alla voce "Fabrizio De André discography" (contrassegnata come item Q3709717) l'informazione viene suddivisa nei quattro tronconi "Fabrizio De André's albums in chronological order" (item Q76901267), "Fabrizio De André live albums discography" (item Q76901439), "Fabrizio De André singles discography" (item Q80600819), "Fabrizio De André tribute albums" (item Q106409719). Considerando anche questa ultima voce, il totale dei documenti segnalati attraverso la proprietà has part(s) (property P527) è di 27. Per Lucio Dalla i tronconi sono soltanto tre: "Lucio Dalla's albums in chronological order" (item Q64567874), "Lucio Dalla singles discography" (item Q64567991), "Lucio Dalla compilation albums discography" (item Q64620677). Il totale dei documenti in questo caso è di 40. Per Gianmaria Testa, infine, non compaiono dati sulla discografia. Se si verifica invece la presenza di documenti relativi a questi

[2] https://cultura.gov.it/casadeicantautori (cons. 25/01/2025)
[3] In particolare le pagine https://www.europeana.eu/it/search?page=1&view=grid&query=fabrizio%20de%20andr%C3%A9, https://www.europeana.eu/it/search?page=1&view=grid&query=lucio%20dalla e https://www.europeana.eu/it/search?page=1&view=grid&query=gianmaria%20testa per Europeana e https://www.wikidata.org/wiki/Q25106 , https://www.wikidata.org/wiki/Q167546 e https://www.wikidata.org/wiki/Q1522516 per Wikidata (cons. 25/01/2025)

tre cantautori con una ricerca sull'OPAC del Servizio Bibliotecario Nazionale (SBN),[4] avendo cura di scegliere la ricerca avanzata e di inserire nome e cognome del cantautore nel campo "Autore" della maschera di ricerca, i documenti risultanti qualificati come "monografia" (in opposizione a "spoglio") e contemporaneamente come "registrazione sonora musicale" risultano essere 948 per Fabrizio De André, 1097 per Lucio Dalla, 53 per Gianmaria Testa. Per quanto differenziati secondo la rispettiva produzione artistica dei tre cantautori scelti, e dunque più o meno grandi a seconda dei casi, questi numeri evidenziano una netta sproporzione rispetto a quanto confluisce nelle due piattaforme della conoscenza sopra considerate. Anche accettando la possibilità di ridondanze dovute a duplicati presenti tra i record SBN, la differenza di visibilità raggiunta nei due diversi canali dal lavoro dei tre cantautori è evidente.

2. METODOLOGIA ADOTTATA

La differenza di risultati nei casi osservati al paragrafo precedente non avrebbe motivo di esistere. La sproporzione con il portale Alphabetica curato dall'Istituto Centrale per il Catalogo Unico delle Biblioteche Italiane e per le informazioni bibliografiche (ICCU), ad esempio, è netta.[5] Non si esaminano qui le possibili motivazioni della dotazione ancora scarna di molte piattaforme del web semantico, né si insiste sull'importanza della presenza di un dataset nello spazio globale dei dati, considerando l'assodata versatilità dei Linked Open Data nel creare relazioni tra i dati e nel permettere di "sfruttare tali dati in modo coerente, consapevole e arricchito dal punto di vista di uno o più mediatori culturali" (Daquino & Tomasi, 2016: 33). Semplicemente, si ipotizza un auspicabile riequilibrio della presenza della produzione cantautoriale in un'organizzazione della conoscenza che rispecchi il più possibile il mondo reale. Una condivisione proficua di dati in quest'ottica è possibile, come tra l'altro testimoniano diverse infrastrutture basate sulla semantica, come ad esempio il portale Linked Open Data del Coordinamento Biblioteche Speciali e Specialistiche di Torino (CoBiS).[6] Oppure, in un contesto internazionale e per un diverso genere musicale, il portale Linked Jazz (Thorsen & Pattuelli, 2016).[7] In assenza di una sistematica operazione di *harvesting* che consenta alle basi di conoscenza di aggregare dati pubblicati dai fornitori, si individua qui una via praticabile per coadiuvare un simile riequilibrio. La metodologia è quella di individuare le risorse presenti nell'OPAC SBN e confrontarle con i risultati delle medesime ricerche in Europeana e in Wikidata per inserirvi quelle mancanti. Le strategie possibili per compiere tale lavoro sarebbero naturalmente molte e diverse. Si presenta qui un modello di procedura semiautomatizzata, applicata al caso di Europeana, nella quale ci si avvale anche di software in linguaggio Python per la conversione e il trattamento dei dati. La strategia di lavoro ideata consiste di una prima fase in cui si recuperano dati dall'OPAC SBN. In un secondo momento i record vengono convertiti dal formato UNIMARC adottato da SBN a quello dei Linked Open Data dell'Europeana Data Model. A questo punto viene effettuato il confronto semiautomatizzato con i risultati prodotti da una query all'API fornita da Europeana. Tutti i casi di record già presenti in Europeana vengono registrati in un apposito file disponibile per un controllo ulteriore da parte dell'operatore umano che individua eventuali falsi positivi per aggiungerli nuovamente alla lista di quelli da inserire. Infine, è previsto l'inserimento vero e proprio in Europeana, anch'esso da effettuarsi massivamente per mezzo di apposito software.

3. RISULTATI

Per la fase di ricognizione iniziale, dato l'ambito musicale delle risorse analizzate, si è individuato un polo particolare tra quelli presenti in SBN: quello dell'Istituto Centrale Beni Sonori e Audiovisivi (ICBSA).[8] Si è ritenuto che per lo scopo dimostrativo della ricerca effettuata i risultati forniti dall'OPAC del polo SBN dell'ICBSA fossero assolutamente qualificati a rappresentare il patrimonio documentale posseduto dalle biblioteche italiane. Per ciascuna delle tre ricerche si seleziona "Esporta Unimarc" tra le funzioni di servizio e si ottiene il relativo file, in formato testuale. Questa funzione è particolarmente apprezzabile per osservare i dati completi che compongono il record bibliografico. Si riporta in Fig. 3 come esempio il record

[4] https://opac.sbn.it/web/opacsbn/risultati-ricerca-avanzata#1744293646260, https://opac.sbn.it/web/opacsbn/risultati-ricerca-avanzata#1744293719667 e https://opac.sbn.it/web/opacsbn/risultati-ricerca-avanzata#1744293767985 (cons. 25/01/2025)
[5] https://alphabetica.it/risultati/-/s/results?input=Fabrizio+De+Andr%C3%A9, https://alphabetica.it/risultati/-/s/results?input=Lucio+Dalla e https://alphabetica.it/risultati/-/s/results?input=Gianmaria+Testa (cons. 25/01/2025)
[6] Il portale aggrega per Fabrizio De André 59 risultati, per Lucio Dalla 32, per Gianmaria Testa 7: https://dati.cobis.to.it/ricerca?q=Fabrizio%20De%20Andr%C3%A9&start=1, https://dati.cobis.to.it/ricerca?q=Lucio%20Dalla&start=1 e https://dati.cobis.to.it/ricerca?q=Gianmaria%20Testa&start=1 (cons. 25/01/2025)
[7] https://linkedjazz.org (cons. 25/01/2025)
[8] https://cloud.sbn.it/opac2/DDS/ricercaSemplice (cons. 25/01/2025)

bibliografico del disco *Montgolfieres*, pubblicato da Gianmaria Testa nel 1995, come appare nel file testuale ricavato dalla codifica UNIMARC. Si noterà, tra l'altro, anche la presenza dell'URL della versione digitalizzata della risorsa, all'interno del campo 899.[9] Essa, sia pur presente anche su Internet Culturale,[10] l'aggregatore del Ministero della Cultura, non è citata da Europeana (vedi Fig. 2) e non compare in Wikidata (come del resto nessun'altra opera pubblicata dal cantautore). Si ripete dunque la ricerca per ciascuno dei tre cantautori prescelti in modo da procedere. La ricerca avanzata di opere che hanno come autore Fabrizio De André, dopo aver eliminato gli spogli e trattenuto le sole "monografie", cioè le opere complessive dalle quali essi siano espunti, e tra le monografie le sole "registrazioni sonore musicali", fornisce 233 risultati. Per Lucio Dalla sono invece 192 e per Gianmaria Testa soltanto 5. Per la fase di conversione dei dati ricavati dal formato UNIMARC e mappatura di questi al formato Linked Open Data

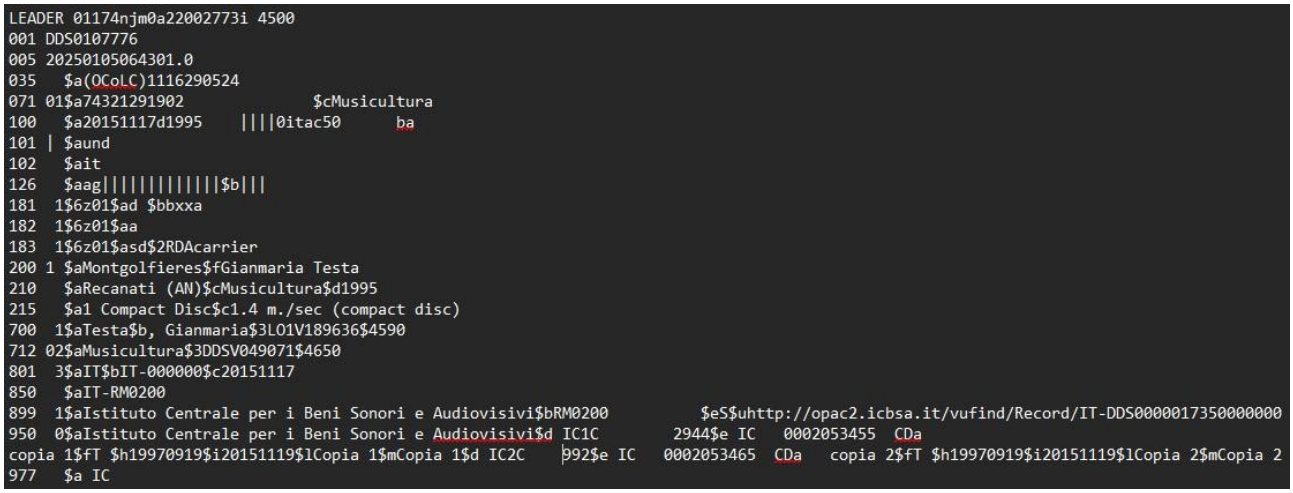


Figura 3. Un record UNIMARC tratto dall'OPAC SBN

di Europeana, serializzati in un file JSONLD, ci si è serviti di un apposito script in linguaggio Python. Tutto il software realizzato nell'ambito di questa ricerca è disponibile su Github.[11] Simili modalità operative sono documentate in diversi ambiti del contesto dei Beni Culturali (Lorenzini, 2016: 214-215). Per la fase successiva del confronto con i documenti presenti in Europeana, si interroga l'*endpoint* per ottenere la lista in formato JSONLD.[12] A questo punto, mediante l'esecuzione di un apposito script in linguaggio Python (presente nel *repository* Github) il confronto viene automatizzato: in due cicli annidati un record per volta del file convertito da UNIMARC viene messo a confronto con un record per volta del file dei record tratti da Europeana fino a esaurimento o fino al riscontro di una similarità. Il confronto avviene tra autore e titolo presenti nei due record. Se sono simili (l'algoritmo misura se la similarità tra le stringhe supera una certa soglia, per gestire anche i casi di grafie leggermente discordanti) il record di partenza viene accantonato in un apposito file che memorizza i record già presenti e il confronto riparte dal record SBN successivo. Tutti i record del file di partenza esportato dall'OPAC SBN e convertito in JSONLD che non trovano corrispondenza nei record Europeana vengono trascritti invece in un apposito file pronto per l'inserimento in Europeana. Un esame successivo effettuato direttamente da un operatore umano sui record accantonati verifica l'esattezza del riconoscimento per individuare eventuali falsi positivi che, quando presenti, vengono trasferiti nel file dei record pronti per l'inserimento. Un ultimo script Python, anch'esso caricato in Github, è stato predisposto per effettuare tramite l'API l'inserimento massivo dei record mancanti in Europeana.

4. LAVORI FUTURI

Stabilire un flusso di lavoro che permetta di andare oltre il *proof of concept* e ottenere risultati apprezzabili, considerando la mole di dati da elaborare in modo sistematico, richiederà un progetto collettivo che veda la collaborazione tra più soggetti coinvolti con un intento condiviso e sotto un coordinamento efficace. Il presente contributo vuole essere uno stimolo affinché questo avvenga, con

[9] http://opac2.icbsa.it/vufind/Record/IT-DDS0000017350000000 (cons. 25/01/2025)
[10] https://www.internetculturale.it/it/16/search/detail?id=oai%3A192.168.10.31%3A22%3ARM0200%3AIT-DDS0000017350000000&mode=all&teca=ICBSA (cons. 25/01/2025)
[11] https://github.com/arts-lod/songwriters (cons. 25/01/2025)
[12] https://api.europeana.eu/console/index.html?url=docs/v3/search.json (cons. 25/01/2025. Per l'interrogazione dell'API è necessario ottenere una chiave di accesso che si può richiedere solo registrandosi)

l'auspicio che l'attenzione futura riservata a una fase storica fondamentale della musica italiana porti a trasferire anche nel web semantico tutta la conoscenza archiviata di essa. L'inserimento sulle piattaforme della conoscenza di dati ottenuti attraverso le procedure descritte richiederà in molti casi che vengano stipulati specifici accordi tra istituzioni che dovranno essere sanciti in via canonica con tutte le approvazioni necessarie.[13]

RINGRAZIAMENTI

Il presente lavoro nasce spontaneamente dall'ammirazione per figure di artisti che con il loro talento e la loro applicazione hanno mutato il panorama sonoro italiano. Non è possibile citare tutti e tutte, ma è a loro che va una immensa gratitudine. Il software depositato su Github è stato ottimizzato utilizzando ChatGPT.

BIBLIOGRAFIA

Ciofalo, G. (2022). La televisione e il cantautorato: lo scenario italiano degli anni Ottanta tra mediazione e mediatizzazione. *Comunicazioni Sociali*, *44*(3), 1. https://doi.org/10.26350/001200_000166

Daquino, M., & Tomasi, F. (2017). Linked Cultural Objects: dagli standard di catalogazione ai modelli per il web of data. Spunti di riflessione dalla Fototeca Zeri. *Umanistica Digitale*, (1). https://doi.org/10.6092/issn.2532-8816/7195

Lorenzini, M. (2016). Le API di Europeana come esempio di integrazione e rappresentazione delle risorse culturali. *Archeologia e calcolatori*, 211-216. ISSN 2385-202X

Porciani, E. (2020). Bob Dylan e gli strimpellatori. La popular music nella narrativa del Settantasette e dintorni. *Between (Cagliari)*, *10*(19). https://doi.org/10.13125/2039-6597/4042

Thorsen, H. K., & Pattuelli, M. C. (2016). Linked Open Data and the Cultural Heritage landscape. In E. Jones & M. Seikel (A c. Di), *Linked Data for Cultural Heritage*. Facet Publishing. ISBN 9781783301621

[13] https://pro.europeana.eu/page/aggregators (cons. 25/01/2025)

OWL Ontology on the European Integration Process between 1949 and 1979

Lorenzo Galvagno[1]
[1] University of Catania, Italy – lorenzo.galvagno@phd.unict.it

ABSTRACT (ENGLISH)

This proposal concerns the representation of the European Integration Process between 1949 and 1979, by means of a computational ontology described using the Web Ontology Language 2 (OWL2). This work of formalization aims at combining the participatory objectives of Public History with a relevant current topic, such as the historical-political process that led to the European Union we live in today. This matter often constitutes a source of widespread misunderstanding by European citizens. At last, some individuals of particular interest are explored, as well as some possible future developments and applications of this case study are listed.
Keywords: Computational ontologies; European Integration; EU Law; Semantic Web; Public History

ABSTRACT (ITALIANO)

Ontologia OWL sul Processo di Integrazione Europeo tra il 1949 e il 1979.
Viene presentata una proposta di rappresentazione del Processo di Integrazione Europeo nel periodo compreso tra il 1949 e il 1979, attraverso un'ontologia computazionale descritta utilizzando il Web Ontology Language 2 (OWL2). Questa formalizzazione intende coniugare gli obiettivi partecipativi della Public History con un tema di attualità: il percorso storico e politico che ha portato alla nascita dell'Unione Europea come la conosciamo oggi. Si tratta di un argomento spesso oggetto di incomprensione da parte dei cittadini europei. Infine, vengono analizzati alcuni individui di particolare interesse all'interno dell'ontologia e si presentano i possibili sviluppi futuri e le potenziali applicazioni di questo caso studio.
Parole chiave: Ontologie computazionali; Integrazione Europea; Diritto UE; Web Semantico; Storia Pubblica

1. INTRODUCTION. PREMISES AND OBJECTIVES OF THE ONTOLOGY

The work here discussed is an attempt to describe a complex subject, both on the historical (Rapone, 2002) and juridical side (Pocar, 1979; Pocar, 1986), using a computer tool. It was, in fact, created following the objectives of Public History, promoting culture and historical memory to an audience beyond academia. The purpose was to increase the participation and awareness of the community on these issues (Cauvin, 2016). European citizens have indeed shown ambivalent positions in the understanding of the functioning (De Waele, 2019) and in the support towards the European institutions (Motti-Stefanidi, 2018). In order to achieve this end, a computational ontology of the European Integration Process was developed using the W3C standard language for the representation of ontologies, the Web Ontology Language 2 (OWL2).

In this research project, the focus has been placed between the years 1949 and 1979. This specific time range has been selected, because it constitutes the basis of the "tree", that is the basis of the temporal ramification of events that have brought forward the evolution of the historical and political process of integration of the European institutions. 1949, the year of the creation of NATO and the Council of Europe, constitutes a historic break for European citizens from the war events of the previous fifty years. The European hegemony on the world stage was, in fact, definitively over and two new superpowers, the USA and the USSR, had emerged (Rapone, 2002).

The pro-European sentiment, whose demands were also present in the agendas of various partisan movements (both anti-fascist and anti-Nazi) in the post-war period, was also driven by a sense of unity, aimed at countering the spectre of communism that came from Eastern Europe (Rapone, 2002).

In this historical context, the first step in the process of integration and creation of a united Europe was the Schuman Plan. The French minister of Foreign Affairs, from whom the plan took its name, proposed to pool coal and steel resources of European countries that had wanted to join the initiative. The Treaty of Paris was signed on 18 April, 1951 by the governments of France, Italy, Netherlands, Belgium, Luxembourg and the Federal Republic of Germany, thus creating the ECSC (European Coal and Steel Community). At the time, (initially it was only an international organization) it would have constituted the initial nucleus of the future Community Europe (Rapone, 2002).

The first elections of the European Parliament, in 1979, marked a turning point in the history of this initial process and the beginning of a new phase (Rapone, 2002). Until then, the members of the European Parliament had been appointed by national governments (Pocar, 1986). The introduction of direct elections, on the other hand, contributed to a growing feeling among European citizens of greater control and influence over the decision-making bodies of the European Communities (Rapone, 2002).

2. MODELLING CRITERIA

The construction of this ontology required a deep analysis of this specific domain of knowledge, in order to extract the main characteristics and relations to be represented.
In particular, as in regard to the historical aspect, some fundamental issues were examined: the Schuman Declaration; the creation of the ECSC (and also the cases of EDC[1] and EPC,[2] described as prospects of political integration and defence between 1950 and 1953); the signing of the Treaties of Rome with the subsequent birth of the EEC[3] and Euratom or EAEC[4] in 1957; the birth of the CAP[5] in 1962; the crisis of the "Empty Chair" and the beginning of some specific Intergovernmental Summits, which would in future become known as the "European Council" since 1965; the launch of the Treaty of Fusion of the Executive Bodies of the European Communities (Merger Treaty); the first enlargement of the European Communities to Denmark, Ireland and the United Kingdom in 1973; and finally, the official birth of the European Council in 1974 and the first elections by universal suffrage to the European Parliament in 1979 (Rapone, 2002).
Instead, as in regard to the people included and described in the ontology, who have contributed (even involuntarily) or were decisive in the process of community construction, it was decided to start the analysis from the list of "Pioneers of the European Union" on the official website.[6] It was then, however, chosen to integrate the list with other relevant personalities (absent in the website) of this historical-political process. The definition "Outstanding Personalities" was chosen, in the end, to describe (in a common category) all these people, since it is not possible to associate them all with the labels of "Pioneers" or "Founding Fathers" of the European Union. The following "Outstanding Personalities" have therefore been described: Alcide De Gasperi; Altiero Spinelli; Charles de Gaulle; Jean Monnet; Johan Willem Beyen; Joseph Bech; Konrad Adenauer; Paul-Henri Spaak and Walter Hallstein (Rapone, 2002).
Another fundamental aspect of the adaptation of the community theme to the ontological representation was the need to take into account the dichotomy between the ideas of federalism and unionism. These themes were, in fact, heated sources of debate among the pro-European movements of the 1950s, as they expressed different visions on how the integration process could be achieved. The first, looking at the institutional model of the United States, tended towards the creation of supranational powers. The latter, on the other hand, was sceptical about the hypothesis of limitations of states' sovereignty, leaning instead towards the use of an intergovernmental method of agreement on individual issues. In the end, the first conception prevailed and was implemented into political initiative through the use of the functionalist method. Political integration, in short, was not achieved through the creation of supranational entities, with limitations to the sovereignty of states "imposed" from above. Rather, it turned to sectoral integration for specific areas, as in the initial case of the coal and steel market, with the aim of building a united Europe, one sector at a time. This can be clearly seen stated in the Schuman Declaration.
The main elements which were represented in the ontology were: the institutions and executive bodies of the European Communities; the treaties establishing the international organizations regarding this field; Nation-states; cities mentioned and the values of the European institutions.
It is also important to highlight that the work here presented is not a totally faithful portrayal of reality, but rather a useful modelling of it, in order to translate and define it within the representation limits provided by a language for the representation of knowledge, such as OWL2.
Moreover, in the construction of the ontology, it was also necessary to take into account the discrepancies between: the handbooks on history of European integration (Rapone, 2002); the handbooks on

[1] European Defence Community
[2] European Political Community
[3] European Economic Community
[4] European Atomic Energy Community
[5] Common Agricultural Policy
[6] European Union official website. EU pioneers https://european-union.europa.eu/principles-countries-history/history-eu/eu-pioneers_en (cons. 11/01/2025)

Community Law (Pocar, 1979; Pocar, 1986); and the official websites of the European Union,[7] of its bodies and institutions operating close to its geo-political areas of interest,[8] as a way to develop a mediation between them. However, despite the attempts to respect the juridical reality, both in terms of terminology and time dates, it was chosen to give priority to the historical handbooks and official websites of the European institutions. This was due to discrepancies between sources, especially in the case of dates. As a result, a historical reading was provided.
More importantly, the historical framework has made it possible to highlight the elements of structural change in the Community system, even when they have not yet been fully formalised and have not yet been translated into juridical form, such as in the case of the European Council, which will be discussed in detail shortly. On the other hand, however, juridical language has proved essential in defining the objects of the ontology, given its univocity, resulting more dialoguing with the world of information technology. The juridical discourse has thus played a bridging role between history and information technology.
Finally, during the study phase to assess the criteria and characteristics to be implemented in the modelling process of the ontology, a research was also carried out on the official websites of the European institutions. This was accomplished to check whether ontologies exist concerning the subject discussed here, or similar. Following the research, on the website of the Publications Office of the European Union, it emerged that there are several Ontological Vocabularies[9] such as the "Common Meta-Data Model" and the "European Commission Conceptual Framework (ECCF)". Each of these ontologies offer useful insights for a modelling work, but many of them have been found to be meager in terms of the number of individuals or properties represented, sometimes possessing only the class structure. For these reasons they have not been used.
Some classes were instead used from the CIDOC (Conceptual Reference Model) ontology (Bekiari, 2023) and from DBpedia (Lehmann, 2015) as superclasses of classes created *ad hoc* for modelling objectives.

3. ONTOLOGY

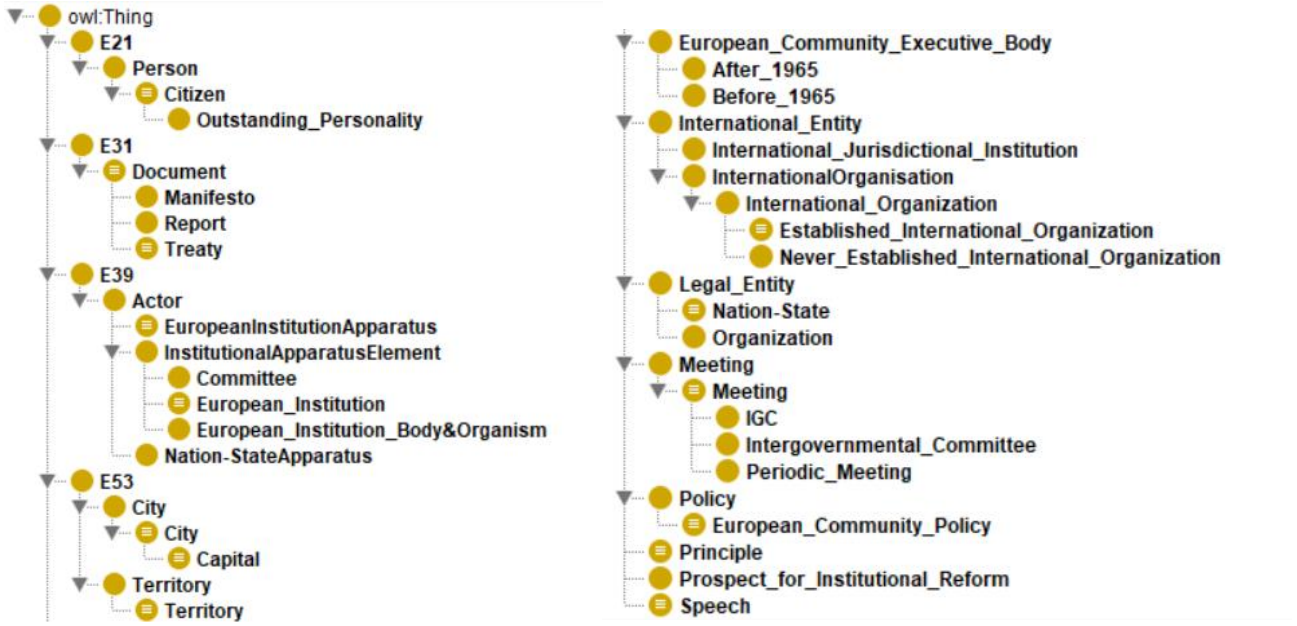


Figure 1. Classes structure

The ontology consists of 45 classes (see Fig. 1), 212 individuals, 87 object properties (39 of which are their inverse properties) and 16 data properties. It was created using the Protégé (Musen, 2015) editor. It

[7] European Union https://european-union.europa.eu/index_it (cons. 11/01/2025). Centre virtuel de la connaissance sur l'Europe (CVCE) https://www.cvce.eu/en (cons. 11/01/2025).
[8] Council of Europe https://coe.int/it/web/portal/ (cons. 11/01/2025).
NATO https://www.nato.int/ (cons. 11/01/2025).
[9] Publications Office of the European Union. EU Vocabularies https://op.europa.eu/en/web/eu-vocabularies/ontologies (cons. 11/01/2025).

was also possible to check the consistency of the axiomatic assertions, by using the Reasoner "Hermit 1.4.3.456", present by default in the editor. To increase the efficiency of the reasoner, inference rules have also been introduced (Antoniou, 2004).
Some individuals who have posed challenges in the modelling process will be discussed below.

4. EDC TREATY

The treaty (which would have established the European Defence Community) was described by (see Fig. 2): the Nation-states that signed the treaty; the city where it was signed; the Pleven Plan, which led to the drafting of the treaty; the respect for the values and principles that it would have enshrined and the fact that, if it had been ratified by all the signatory states, it would have established the EDC. The individual representing this institution was classified in the class "Never_Established_International_Organization". It was necessary to create this class and distinguish it from the "Established_International_Organization" class, since in this case the treaty had not been ratified (Rapone, 2002). In the case of the European Political Community (EPC), on the other hand, it was only drafted as a treaty by the ECSC assembly and never reached the national parliaments.[10] These two institutions (both of which are in the same class) that were never created, represented, in fact, a major setback for the process of integration of the origins, which had to return to the limited sectoral integrations of the functionalist method (Rapone, 2002). The particularity of the description of the EDC was expressed by the object property "IfRatifiedWasSupposedToEstablish", to indicate that if the treaty establishing the EDC had been ratified by all signatory states it would have been established, but this did not happen due to the non-ratification by the French Parliament.

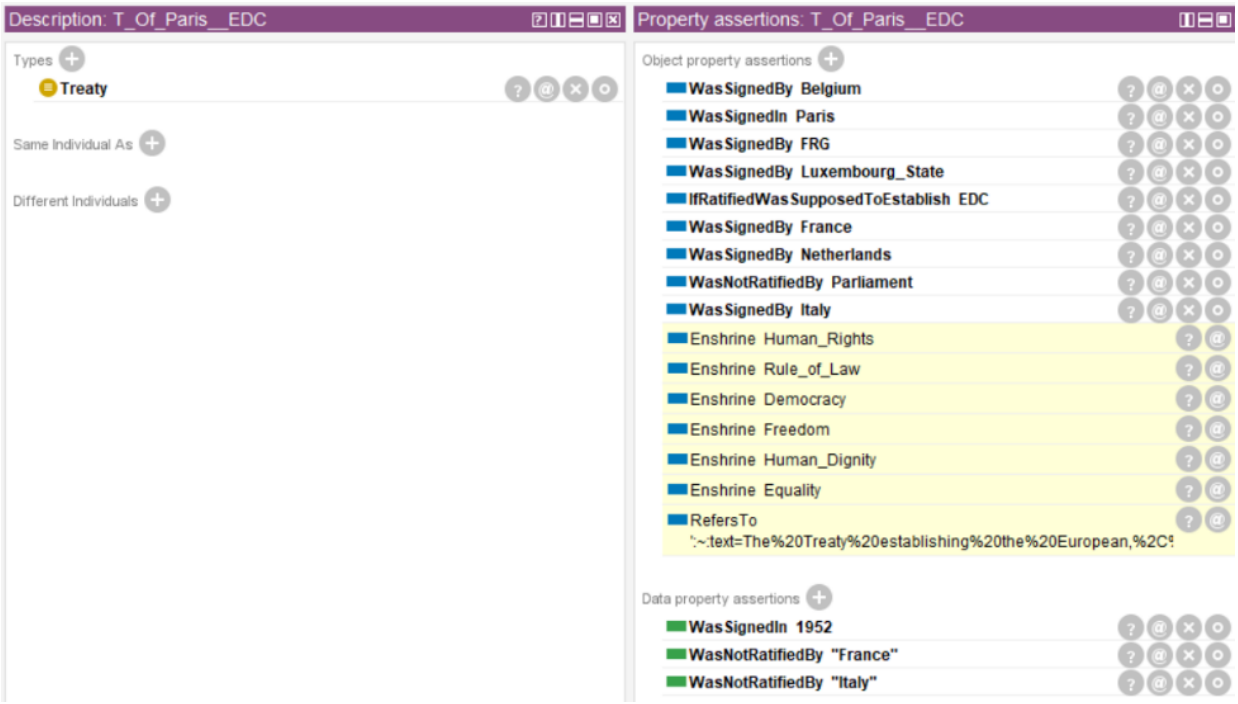


Figure 2. Individual EDC Treaty

5. EUROPEAN COUNCIL

In particular, representing the European Council (see Fig. 3) has been challenging while modelling this aspect of the ontology. First, there was a discrepancy between the sources, regarding the handbooks of history of European integration (Rapone, 2002), the official website of the European Union and the handbooks of Community Law (Pocar, 1979; Pocar, 1986). In the first case, the official date of birth of the European Council is 1975, but the decision was taken in 1974. In the second case, its creation is placed in 1974 as an informal forum, 1992 with official status (Maastricht Treaty) and 2009 as a real official

[10] Centre virtuel de la connaissance sur l'Europe (CVCE). EPC https://www.cvce.eu/collections/unit-content/-/unit/en/02bb76df-d066-4c08-a58a-d4686a3e68ff/6550430e-98c0-4441-8a60-ec7c001c357b/Resources#8b63810a-e5bd-4979-9d27-9a21c056fc8d_en&overlay (cons. 11/01/2025).

institution of the EU (Lisbon Treaty). Whereas, in the last case, the nature of institution of the European Council is mentioned, for the first time, only in the draft treaty establishing the European Union approved in 1984.
Unlike all the other institutions, however, the European Council was not established by a treaty but was conceived as a further form of consultation (not formally belonging to the institutional framework of the Community at that time) among the Heads of State and Government of the European Communities. The European Council was born as a specific intergovernmental summit in the mid-1960s (Pocar, 1979). Unlike the other summits, in fact, this became periodic and in 1974 was officially called, for the first time, "European Council". It will also be officially incorporated into the institutional framework, only with the Single European Act in 1986, while its functions and frequency of meetings will be defined, officially, only with the Treaty of the European Union (or Maastricht) of 1992.[11]

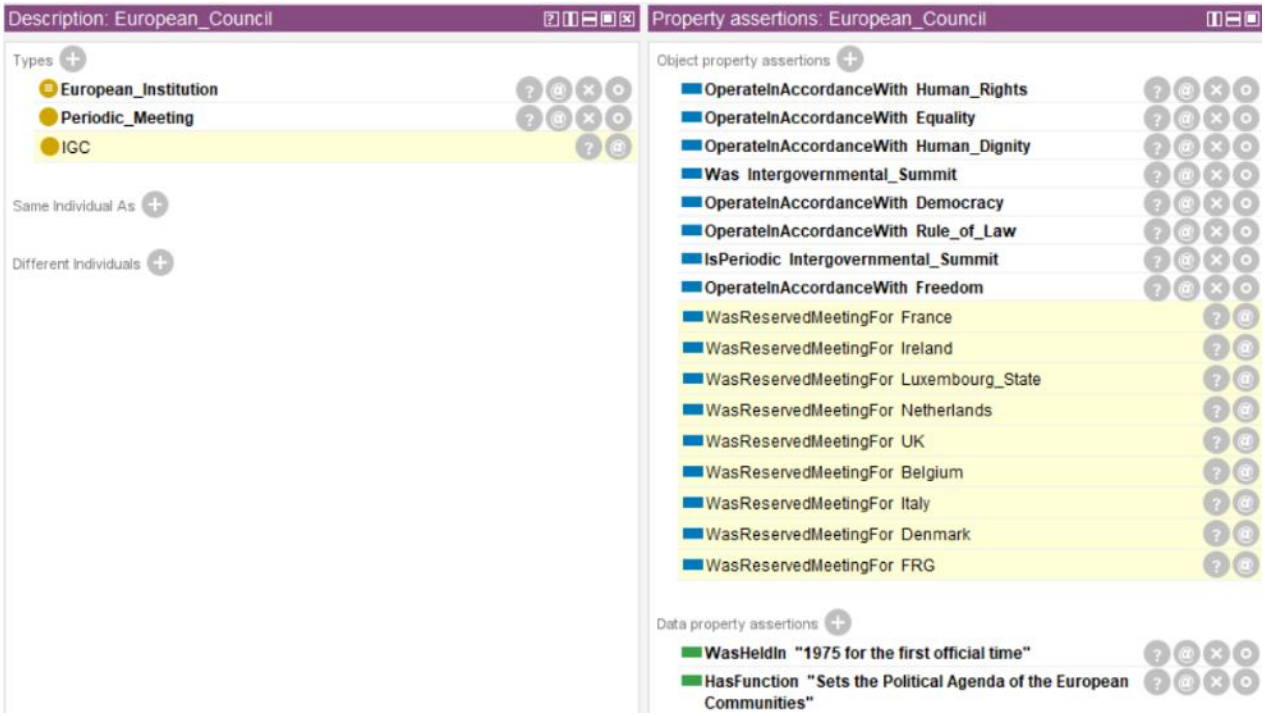


Figure 3. Individual European Council

For the purposes of research and modelling of this ontology, however, it was decided to follow the historical reading, as the date on which the first - official - European Council is placed, in Dublin in March 1975, although the decision was taken a year earlier (Rapone, 2002).
In the ontology, the European Council has been classified as: "European_Institution", since it operated in accordance with the values of the European institutions; as "IGC" or Intergovernmental Conference; and as "Periodic_Meeting". It was also expressed (through the object properties) the participation, at that time, by the Heads of State and Government of the nine member states in these meetings reserved only for them with the property "WasReservedMeetingFor" (inverse property of "HadAccessToReservedMeeting"). Other IGCs might not have the Heads of State or Government as representatives. The function of the European Council was to set the political agenda for the European Communities (Pocar, 1986).
The class "European_Institution" represents the instances of all the european institutions up to 1979. The European Community executive bodies before and after the Merger Treaty; the European Parliament; the Court of Justice of the European Communities; the European Council and the European Court of Auditors. The class "European_Institution" was placed as equivalent to the expression "OperateInAccordanceWith value Democracy, Equality, Freedom, Human_Dignity, Human_Rights, Rule_of_Law", which means that, European Institutions operate in accordance with those principles. The "IGC" class, on the other hand, represents the instances of intergovernmental meetings between Heads of State and Government. Its instances include: the Brussels conference; the European Council; the Hague summit; the Messina

[11] Centre virtuel de la connaissance sur l'Europe (CVCE). European Council https://www.cvce.eu/en/education/unit-content/-/unit/d5906df5-4f83-4603-85f7-0cabc24b9fe1/f53d9750-e539-4fe0-9ee8-2910f099eff1 (cons. 11/01/2025).

conference; the Paris summits of 1972 and 1974; the Venice conference and a generic instance of intergovernmental summit.
In addition, SWRL[12] (Antoniou, 2004) inference rules were defined to express some class affiliations and properties of the European Council.

Rule1: OperateInAccordanceWith(European_Council, Human_Rights), OperateInAccordanceWith(European_Council, Equality), OperateInAccordanceWith(European_Council, Human_Dignity), OperateInAccordanceWith(European_Council, Democracy), OperateInAccordanceWith(European_Council, Rule_of_Law), OperateInAccordanceWith(European_Council, Freedom), IsPeriodic(European_Council, Intergovernmental_Summit) -> IGC(European_Council)

Rule2: Periodic_Meeting(European_Council) -> WasReservedMeetingFor(European_Council, FRG), WasReservedMeetingFor(European_Council, Italy), WasReservedMeetingFor(European_Council, France), WasReservedMeetingFor(European_Council, Belgium), WasReservedMeetingFor(European_Council, Netherlands), WasReservedMeetingFor(European_Council, Luxembourg_State), WasReservedMeetingFor(European_Council, Denmark), WasReservedMeetingFor(European_Council, UK), WasReservedMeetingFor(European_Council, Ireland)

The first of these two rules can be read as follows: if the European Council operates in accordance with values such as human rights, equality, human dignity, democracy, rule of law and freedom and is also a periodic intergovernmental summit; then the individual "European Council" is definable as an instance of the IGC class. The second one states that: if the European Council is part of the "Periodic_Meeting" class, then the European Council is a meeting reserved for the Federal Republic of Germany, Italy, France, Belgium, Netherlands, Luxembourg, Denmark, the United Kingdom and Ireland.

6. CONCLUSIONS AND POSSIBLE DEVELOPMENTS

Querying the ontology presented so far, on Protégé, was possible by means of the Protégé plugin "Snap SPARQL Query", which produced in output the correct data requested by the provided QUERIES. In the future, it is planned to make the ontology and the data accessible online through the creation of a SPARQL Endpoint. One of the most used platforms, to achieve this goal, is Virtuoso OpenLink.[13] It has a multi-model architecture, which can handle data of various types: from relational databases (queryable with SQL) to RDF knowledge graphs (queryable with SPARQL), to XML data, documents and much more. Virtuoso is known for its high performance and scalability. Virtuoso is also used in the Linked Open Data (LOD) Cloud, the world's largest publicly accessible knowledge graph.[14]
The ontological dictionaries on the EU's Publications Office website also provide a further opportunity to reflect on possible developments. In a future, possible extension, it could be considered to integrate them into this ontology by reaching the second decade of the 21th century and increasing the level of detail of its representation by integrating those themes treated.
These plans of expansion and online usability would fully meet the participatory objectives of Public History (Cauvin, 2016) and would constitute a break from the "elite" academic practice of historical research.

[12] Semantic Web Rule Language
[13] Virtuoso OpenLink https://docs.openlinksw.com/virtuoso/ (cons. 11/01/2025)
[14] The Linked Open Data Cloud https://lod-cloud.net/ (cons. 11/01/2025)

REFERENCES

Antoniou, G., Groth, P., Harmelen., F., V, Rinke, H. (2004). *A semantic Web Primer*. MIT Press.

Bekiari, C., Bruseker, G., Canning, E., Doerr, M., Michon, P., Ore, C.-E., Stead, S., Velios, A. (2023). Definition of the CIDOC Conceptual Reference Model v7.2.3 (Version v7.2.3). The CIDOC Conceptual Reference Model Special Interest Group.

Cauvin, T. (2016). *Public History: A Textbook of Practice*. Routledge, Taylor & Francis Group.

De Waele, H. (2019). Union Citizens and the European Parliament. Perception, Accessibility, Visibility and Appreciation, in P. E. Minderhoud, S. A. Mantu, K. M. Zwaan, (Eds.), *Caught In Between Borders: Citizens, Migrants and Humans. Liber Amicorum in honour of prof. dr. Elspeth Guild* (pp. 175-182). Wolf Legal Publishers.

Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., Hellmann, S., Morsey, M., Van Kleef, P., Auer, S., Bizer, C. (2015). DBpedia – A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. *Semantic Web*, *6*(2), 167–195. DOI: 1570-0844/12/$27.50

Motti-Stefanidi, F., Cicognani, E. (2018). Bringing the European Union closer to its young citizens: Youth active citizenship in Europe and trust in EU institutions. *European Journal of Developmental Psychology*, *15*(3), 243-249.

Musen, M. A. (2015). The Protégé project: A look back and a look forward. *AI Matters. Association of Computing Machinery Specific Interest Group in Artificial Intelligence*, *1*, 4-12. DOI: 10.1145/2757001.2757003

Pocar, F. (1979). *Lezioni di diritto delle Comunità Europee*. Giuffrè Editore

Pocar, F. (1986). *Lezioni di diritto delle Comunità Europee*. Giuffrè Editore

Rapone, L. (2002). *Storia dell'Integrazione Europea*. Carocci Editore

LiITA, una Knowledge Base di risorse interconnesse per l'italiano

Eleonora Litta, Marco Passarotti, Paolo Brasolin[1]
Valerio Basile, Cristina Bosco, Andrea Di Fabio[2]
[1] Università Cattolica del Sacro Cuore, Milano, Italia – eleonora.litta@unicatt.it
[2] Università di Torino, Italia – valerio.basile@unito.it

ABSTRACT (ITALIANO)

Il progetto LiITA (https://www.liita.it/) si propone di rendere interoperabile e quindi sfruttare la ricchezza di risorse linguistiche disponibili per la lingua italiana (corpora, edizioni digitali, lessici, dizionari e glossari) attraverso l'applicazione dei principi Linked Open Data. Questo articolo descrive i primi passi per lo sviluppo di una Knowledge Base aperta, in cui le risorse linguistiche sono rese interoperabili attraverso il collegamento delle voci delle risorse lessicali e delle occorrenze delle risorse testuali alle forme di citazione registrate in una raccolta di lemmi (Lemma Bank) della lingua italiana. Seguendo i principi del paradigma Linked Open Data, l'interoperabilità concettuale tra le risorse distribuite connesse in LiITA è ottenuta applicando un vocabolario per la descrizione della conoscenza comunemente utilizzato nel mondo dei Linked Open Data linguistici, adottando modelli ampiamente utilizzati, tra cui il principale è OntoLex-Lemon.
Parole chiave: italiano; Linked Open Data; interoperabilità; RDF

ABSTRACT (ENGLISH)

The LiITA project, a Knowledge Base of interconnected resources for Italian. The LiITA project (https://www.liita.it/) aims to make interoperable and thus exploit the wealth of linguistic resources available for the Italian language (corpora, digital editions, lexicons, dictionaries and glossaries) through the application of Linked Open Data principles. This paper describes the first steps for the development of an open Knowledge Base, in which linguistic resources are made interoperable by linking the entries of lexical resources and the occurrences of textual resources to the citation forms recorded in a collection of lemmas (Lemma Bank) of the Italian language. Following the principles of the Linked Open Data paradigm, conceptual interoperability between the distributed resources connected in LiITA is achieved by applying a vocabulary for the description of knowledge commonly used in the world of linguistic Linked Open Data, adopting widely used models, among which the main one is OntoLex-Lemon.
Keywords: Italian, Linked Open Data, interoperability, RDF

1. INTRODUZIONE

L'italiano è tra le lingue più ricche per numero di risorse linguistiche digitali, sia lessicali che testuali. Una ricerca effettuata sul CLARIN Virtual Language Observatory,[1] un'interfaccia che permette di trovare risorse linguistiche in vari domini, restituisce più di 8.000 risultati per la lingua italiana. Come altre lingue ricche di risorse, per l'italiano esistono risorse fondamentali, tra cui WordNet (Pianta et al. 2002 e Roventini et al. 2016), treebank mantenute nel repository di Universal Dependencies,[2] corpora storici[3] e corpora di riferimento e monitoraggio della lingua scritta (*CORIS/CODIS*, Favretti et al. 2002) e parlata (*KIParla*, Mauri et al. 2019). Tuttavia, analogamente a quanto osservato per molte altre lingue, le risorse linguistiche disponibili per l'italiano presentano rilevanti eterogeneità in termini di formato dei dati, criteri di annotazione e tagset adottati. Questa variabilità ostacola l'integrazione e l'interoperabilità dei (meta)dati provenienti dalle diverse risorse, con effetti negativi sulla fruibilità delle stesse e sullo sviluppo di studi linguistici. Tali risorse, infatti, differiscono non solo per il tipo di informazioni fornite, ma anche per il livello di granularità applicato alla rappresentazione dello stesso oggetto linguistico, ovvero la parola, sia essa intesa come occorrenza in un corpus o come voce di una risorsa lessicale. Far interagire tale ricchezza di informazioni per sfruttare al meglio l'enorme patrimonio dei (meta)dati raccolti in decenni di lavoro, rappresenta una delle principali sfide odierne, non solo per quanto riguarda l'italiano, ma per tutte le lingue. Una linea di ricerca molto attiva si focalizza al momento sui cosiddetti Linguistic Linked Open Data (LLOD). L'obiettivo è quello di definire pratiche comuni per la rappresentazione e la pubblicazione di risorse

[1] https://vlo.clarin.eu (cons. 07/01/2025).

[2] https://universaldependencies.org (cons. 07/01/2025).

[3] https://www.corpusmidia.unito.it/, http://www.ovi.cnr.it/ (cons. 07/01/2025).

linguistiche secondo i principi del paradigma Linked Data, che sta alla base del Semantic Web. La Knowledge Base (KB) LiLa (Linking Latin) è probabilmente il principale caso d'uso LLOD attualmente disponibile. LiLa è una KB di risorse linguistiche per il latino rese interoperabili grazie alla loro rappresentazione e pubblicazione secondo i principi dei Linked Data. Sulla base dell'esperienza maturata con il progetto LiLa, e mediante il riutilizzo della relativa architettura, il progetto LiITA: Linking Italian[4] si propone di realizzare una Knowledge Base (KB) nella quale le risorse linguistiche per l'italiano, pubblicate secondo i principi dei Linked Data, risultino pienamente interoperabili. Il presente contributo descrive lo sviluppo della componente centrale della KB LiITA: la cosiddetta "Lemma Bank", ovvero un repertorio di lemmi della lingua italiana concepito come nodo di interconnessione tra le diverse risorse linguistiche incluse nella KB.

2. LINKED DATA

Il concetto di Semantic Web[5] si fonda sull'idea che i contenuti pubblicati sul World Wide Web debbano essere arricchiti con informazioni e metadati strutturati in modo da renderli accessibili, interrogabili e interpretabili non solo dagli esseri umani, ma anche da sistemi computazionali. Tale strutturazione è realizzata attraverso i Linked Data, che costituiscono l'infrastruttura portante del Semantic Web. A differenza di un Web fatto di ipertesti, in cui i link non sono interpretabili semanticamente, il Semantic Web è caratterizzato da collegamenti tra entità (o "oggetti") identificati univocamente mediante URI (Uniform Resource Identifier). I collegamenti tra oggetti nel contesto del Semantic Web acquisiscono significato semantico in quanto espressi mediante vocabolari controllati e formalizzati sotto forma di ontologie, ovvero modelli condivisi per la descrizione strutturata della conoscenza.
Il paradigma dei Linked Data si articola su quattro principi fondamentali:

1. Utilizzo degli URI come identificatori univoci e persistenti: ogni entità è identificata da un URI, che ne garantisce la disambiguazione e la tracciabilità nel tempo. Nel caso dei (meta)dati linguistici, le entità identificate includono oggetti (meta)linguistici quali occorrenze di parole in testi, lemmi nei dizionari, o parti del discorso;
2. Adozione di URI HTTP per l'accesso sul Web: ciò consente sia agli utenti umani sia ai sistemi automatici di interrogare direttamente le risorse identificate;
3. Utilizzo di standard semantici per la rappresentazione e il recupero delle informazioni: in particolare, RDF (Resource Description Framework) costituisce il modello dati alla base del Semantic Web. In RDF, l'informazione è strutturata in triple, composte da un soggetto, una proprietà e un oggetto, le cui relazioni semantiche sono definite da ontologie condivise. SPARQL[6] (SPARQL Protocol and RDF Query Language) è il linguaggio di interrogazione standard per dati rappresentati in RDF, e consente operazioni complesse di recupero e inferenza sui (meta)dati;
4. Creazione di collegamenti tra URI: la presenza di riferimenti ad altri URI all'interno delle risorse consente la navigazione semantica tra dati distribuiti.

L'applicazione dei principi del paradigma Linked Data ai (meta)dati derivati da risorse linguistiche e la loro pubblicazione sul Web offre diversi vantaggi (Chiarcos et al. 2013). In primo luogo, per quanto riguarda la rappresentazione e la modellazione dei (meta)dati, RDF è un modello particolarmente versatile, adatto a rappresentare metadati come quelli veicolati dai vari livelli di annotazione disponibili nelle risorse linguistiche (morfologia, sintassi, lemmatizzazione, ecc.). Inoltre, l'adozione di un modello di dati comune (RDF) consente due livelli di interoperabilità: l'interoperabilità strutturale (o sintattica), ovvero la capacità di sistemi diversi di elaborare dati scambiati utilizzando protocolli e formati condivisi (come HTTP e URI), e l'interoperabilità concettuale (o semantica), ovvero la capacità di un sistema di interpretare automaticamente e semanticamente le informazioni scambiate utilizzando un insieme comune di classi e categorie di dati definite in ontologie e vocabolari (Ide & Pustejovsky 2010).

[4] https://www.liita.it/.

[5] Introdotto da Tim Berners-Lee et alii 2001.

[6] https://www.w3.org/TR/rdf-sparql-query/ (cons. 08/04/2025).

Sebbene la lingua italiana non sia estranea ai Linked Data,[7] LiITA rappresenta il primo tentativo sistematico di realizzare una risorsa strutturata sotto forma di Lemma Bank, concepita attorno alla parola come unità centrale e punto di intersezione tra le diverse tipologie di risorse linguistiche.

3. LA KNOWLEDGE BASE LiITA: ARCHITETTURA E MODELLO

La componente principale di LiITA consiste in una raccolta di forme canoniche di citazione (lemmi) per la lingua italiana, accessibile dal sito https://www.liita.it/, e scaricabile dalla repository del progetto.[8]
L'architettura della KB di LiITA si rifà a quella della LiLa KB per il latino,[9] condividendone l'assunto centrale secondo cui la parola costituisce l'elemento cardine per l'interoperabilità tra risorse linguistiche: le risorse lessicali, come dizionari o lessici, che descrivono le proprietà delle parole, sono costituite da voci lessicali, e cioè parole, e le risorse testuali, come corpora e biblioteche digitali che consistono in raccolte di testi, sono costituite da token, e cioè occorrenze di parole.
Il collegamento ai rispettivi lemmi rende interoperabili in LiITA le voci lessicali e le occorrenze delle parole provenienti da risorse sparse e prive di interoperabilità, consentendo l'esecuzione di ricerche che spaziano tra le diverse risorse linguistiche. Ad esempio, è possibile cercare tutte le occorrenze del lemma *andare* in più corpora; oppure estrarre tutti quei token che hanno determinate proprietà fornite da una o più risorse lessicali.
Dato dunque il ruolo centrale delle parole nell'architettura di LiITA, la componente principale della KB è una raccolta di lemmi, forme di citazione lessicale adottate abitualmente nelle risorse linguistiche, chiamata Lemma Bank. Questi lemmi trovano corrispondenza, da un lato, nelle voci delle risorse lessicali e, dall'altro, nelle forme scelte per raccogliere tutte le occorrenze di una particolare parola nelle risorse testuali lemmatizzate.
Seguendo i principi del paradigma Linked Data, l'interoperabilità concettuale tra le risorse distribuite connesse in LiITA viene ottenuta utilizzando categorie per la descrizione della conoscenza comunemente utilizzate nel mondo dei Linguistic Linked Open Data. Nel caso specifico della Lemma Bank di LiITA, è stato adottato il modello specificato da OntoLex-Lemon (McCrae et al. 2017), uno dei modelli RDF più utilizzati per rappresentare e pubblicare le risorse lessicali Linked Data. Inoltre, i lemmi sono descritti in un'ontologia personalizzata basata sull'ontologia del progetto LiLa. Questa ontologia, da un lato, fornisce informazioni dettagliate su alcune caratteristiche morfologiche e linguistiche dei lemmi (ad esempio la parte del discorso e il genere grammaticale per i sostantivi) basandosi sul modello di annotazione OLiA [Cimiano et al., 151-155]. D'altro canto, l'ontologia LiLa definisce classi e proprietà per modellare il compito di lemmatizzazione, come la proprietà lila:hasLemma[10] che collega i lemmi ai token di un corpus. Le entrate lessicali (ontolex:LexicalEntry)[11] sono intese come unità di analisi del lessico che possono raccogliere una o più forme (ontolex:Form)[12] e uno o più sensi lessicali (ontolex:LexicalSense)[13], concetti lessicali (ontolex:LexicalConcept)[14] o entità da ontologie.[15] Una parola può inoltre essere scritta in forme diverse, chiamate varianti grafiche (written representations), rappresentate tramite la proprietà ontolex:writtenRep,[16] e avere una o più varianti fonetiche (proprietà ontolex:phoneticRep)[17]. Una di queste

[7] Esempi di risorse per l'italiano al momento disponibili come Linked Open Data sono il lessico CompL-it (http://hdl.handle.net/20.500.11752/ILC-1007, cons. 07/01/2025), ItalWordNet v.2 (http://hdl.handle.net/20.500.11752/ILC-66, cons. 07/01/2025) e una raccolta di nomi dal lessico PAROLE SIMPLE CLIPS (PSC) (http://hdl.handle.net/20.500.11752/ILC-558, cons. 07/01/2025).

[8] https://github.com/LiITA-LOD/LiITA_LemmaBank.

[9] https://lila-erc.eu/data-page/ (cons. 08/04/2025).

[10] http://lila-erc.eu/ontologies/lila/hasLemma.

[11] http://www.w3.org/ns/lemon/ontolex#LexicalEntry.

[12] http://www.w3.org/ns/lemon/ontolex#Form.

[13] http://www.w3.org/ns/lemon/ontolex#LexicalSense.

[14] http://www.w3.org/ns/lemon/ontolex#LexicalConcept.

[15] Per sensi lessicali si intendono i sensi lessicalizzati: un senso corrisponde a una singola voce lessicale. Gli aspetti semantici che possono essere espressi da più parole sono invece rappresentati tramite concetti lessicali (Lexical Concept), che possono quindi avere più di una lessicalizzazione.

[16] http://www.w3.org/ns/lemon/ontolex#writtenRep.

[17] http://www.w3.org/ns/lemon/ontolex#phoneticRep.

forme, l'oggetto della proprietà ontolex:canonicalForm,[18] è la forma che viene convenzionalmente scelta per rappresentare l'intero gruppo delle forme flesse di una voce lessicale. La Lemma Bank di LiITA consiste in una raccolta di queste forme, modellate nell'ontologia come "individui" della Classe lila:Lemma,[19] a sua volta una sottoclasse di ontolex:Form, creata originariamente per il progetto LiLa[20] e adottata di conseguenza nella LiITA Lemma Bank.
I lemmi della Lemma Bank di LiITA non sono tuttavia vincolati da alcuna relazione con una voce lessicale, poiché la Lemma Bank non è una raccolta di voci lessicali, ma un set di forme canoniche di citazione. Questo riflette il ruolo peculiare della Lemma Bank come strumento per rendere interoperabili le risorse. Utilizzando la proprietà lila:hasPos,[21] a ciascun lemma nella Lemma Bank viene assegnata una parte del discorso, utilizzando il tagset UPOS di Universal Dependencies. Nel caso in cui ci siano parole a cui sono assegnate più PoS nelle diverse risorse lessicali, vengono creati nella Lemma Bank più lemmi corrispondenti. Ad esempio, alla parola "rosso" vengono assegnate due PoS: aggettivo e sostantivo e nella Lemma Bank vengono creati due lemmi distinti con due diverse PoS rappresentate tramite la proprietà lila:hasPos.
Per armonizzare eventuali diversi criteri di lemmatizzazione applicati nelle molteplici risorse linguistiche disponibili per l'italiano, la Lemma Bank di LiITA implementa due proprietà specifiche aggiuntive. La prima è la proprietà simmetrica lila:lemmaVariant[22] che collega diverse forme del paradigma flessivo di una parola che possono essere utilizzati come lemmi indipendenti. Il caso tipico è quello dei nomi con doppio plurale (nomi in *-o* con plurale maschile in *-i* e femminile in *-a*, come per esempio *braccio*/*bracci*, *braccia* in cui la differenza nel plurale indica una decisa differenza semantica).[23] Queste parole possono essere lemmatizzate sia nella forma plurale che nella forma singolare: grazie alla proprietà lila:lemmaVariant i due lila:Lemma *braccio* e *braccia*, sono collegati tra loro nella Lemma Bank. La seconda proprietà, lila:hasHypolemma[24] (e la sua proprietà inversa lila:isHypolemma)[25] collega i lemmi che fanno parte del paradigma flessivo di una stessa parola ma hanno parti del discorso diverse. È il caso degli aggettivi usati come avverbi, ad esempio *veloce* che può essere interpretato (e lemmatizzato) sia come una forma di aggettivo (quindi modellato come un lila:Lemma) sia come un avverbio (quindi modellato come un lila:Hypolemma,[26] una sottoclasse di lila:Lemma).
I participi passati e presenti sono un altro tipo di ipolemma (si vedano ad esempio il participio passato *narcotizzato* e il participio presente *narcotizzante*), a cui nella Lemma Bank viene assegnata la PoS aggettivo. I participi sono modellati come individui della classe lila:Hypolemma e sono collegati al loro lemma verbale (*narcotizzare*) tramite la proprietà lila:isHypolemma. Prescindendo quindi dal fatto che due differenti risorse lemmatizzino i participi secondo criteri diversi (vale a dire, che sianno essi ricondotti ad una forma di citazione participiale o verbale), la Lemma Bank considera entrambe le casistiche e armonizza entrambe le possibilità in una.
I lemmi e l'informazione sulla parte del discorso che costituiscono la Lemma Bank si basano sulla base lessicale di una versione online del dizionario Nuovo De Mauro,[27] che ammonta complessivamente a circa 145.000 voci; di queste, 13.000 polirematiche sono state ritenute non rilevanti per gli scopi del progetto, in quanto i lemmatizzatori operano tipicamente su singoli token. Fino ad ora, circa 125.200 lemmi, e 16.330 ipolemmi sono stati estratti dalle restanti 131.000 voci. La logica di categorizzazione della parte del discorso della base lessicale originaria è stata adottata parzialmente e mappata sul tagset UPOS.[28] Poiché la categorizzazione grammaticale originaria del dizionario segue la tradizione linguistica italiana, è stato necessario armonizzare alcune parti del discorso con il sistema Universal Part-of-Speech (UPOS). In particolare, le congiunzioni, che nella fonte originale non distinguono tra forme coordinate e subordinate,

[18] http://www.w3.org/ns/lemon/ontolex#canonicalForm.

[19] http://lila-erc.eu/ontologies/lila/Lemma.

[20] http://lila-erc.eu/ontologies/lila/.

[21] http://lila-erc.eu/ontologies/lila/hasPOS.

[22] http://lila-erc.eu/ontologies/lila/lemmaVariant.

[23] Cf. Serianni, 1989, p.143.

[24] http://lila-erc.eu/ontologies/lila/hasHypolemma.

[25] http://lila-erc.eu/ontologies/lila/isHypolemma.

[26] http://lila-erc.eu/ontologies/lila/Hypolemma.

[27] https://dizionario.internazionale.it/ (cons. 08/04/2025).

[28] Universal Part of Speech Tagset, https://universaldependencies.org/u/pos/.

sono state riallineate manualmente ai tag UPOS. Inoltre, l'estrazione iniziale ha temporaneamente escluso alcune categorie marginali, come prefissi, segni di punteggiatura e alcune abbreviazioni, il cui trattamento è previsto in fasi successive del progetto.

4. Conclusione e lavoro futuro

In questo contributo sono stati illustrati i primi sviluppi del processo di pubblicazione, secondo i principi dei Linguistic Linked Open Data (LLOD), di una raccolta di forme canoniche di citazione (lemmi) per la lingua italiana. Tale raccolta costituisce il nucleo della Lemma Bank, componente centrale del progetto LiITA, una knowledge base pensata per favorire l'interoperabilità tra risorse linguistiche italiane, sulla scia dell'esperienza maturata con la knowledge base LiLa per il latino. LiITA si propone di colmare l'attuale frammentazione delle risorse linguistiche italiane, fungendo da nodo centrale per l'integrazione di lessici e corpora esistenti e futuri. A tal fine, la Lemma Bank è progettata per accogliere e armonizzare differenti criteri di lemmatizzazione adottati nelle varie risorse, senza alterarne la struttura originaria, ma valorizzandone la specificità e garantendone al contempo la massima riusabilità all'interno dell'ecosistema Linked Data.Il passo immediatamente successivo nello sviluppo della KB LiITA, entro i termini del finanziamento corrente, consiste nel collegamento di una almeno una risorsa testuale e di almeno una risorsa lessicale.

Poiché LiITA intende adottare un processo resource-driven di acquisizione dei dati, la costruzione della Lemma Bank per rendere interoperabili le risorse distribuite per l'italiano adottando il paradigma Linked Data rimane inevitabilmente un processo aperto. Il collegamento di sempre più risorse alla KB richiederà l'inclusione di nuovi lemmi e l'implementazione incrementale della Lemma Bank di LiITA, sia attraverso l'estrazione di lemmi da varie fonti lessicali sia, in modo controllato, dalle risorse testuali.

Oltre all'estensione della Lemma Bank e al collegamento delle prime risorse, il progetto LiITA si propone anche lo sviluppo di servizi online per semplificare l'accesso ai dati. Il processo di collegamento di un testo o corpus nella KB verrà infatti supportato da uno strumento che permette lemmatizzazione automatica, PoS-tagging e collegamento. A questo scopo, è stato recentemente addestrato un nuovo modello Stanza (Qi et al. 2020) testato su tutte le treebank italiane disponibili in Universal Dependencies. Questo modello servirà come base per il processo di collegamento delle risorse testuali da includere nella KB di LiITA.[29]

L'interrogazione avanzata dei dati offerta da tutte le risorse interconnesse in LiITA sarà facilitata da un'interfaccia grafica che aiuterà nel compito di scrivere complesse query SPARQL.

Infine, data l'architettura language-independent e l'adozione di un vocabolario di descrizione della conoscenza basato su modelli e ontologie ampiamente usati nella comunità LOD, il progetto LiITA può determinare un sostanziale avanzamento nell'accesso, pubblicazione ed utilizzo delle risorse linguistiche per l'italiano, come il progetto pioniere LiLa ha fatto per il latino, applicando una metodologia focalizzata sull'interoperabilità basata sull'approccio Linked Open Data.

RINGRAZIAMENTI

Il presente contributo è finanziato dall'Unione Europea – Next Generation EU, Missione 4 Componente 1, CUP J53D2301727OOO1. Il progetto PRIN 2022 PNRR "LiITA: Interlinking Linguistic Resources for Italian via Linked Data" è realizzato congiuntamente dall'Università Cattolica del Sacro Cuore di Milano e dall'Università di Torino.

BIBLIOGRAFIA

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The Semantic Web. Scientific American, 284(May 2001), 34–43.

Chiarcos, C., Moran, S., Mendes, P. N., Nordhoff, S., & Littauer, R. (2013). Building a Linked Open Data Cloud of Linguistic Resources: Motivations and Developments (I. Gurevych & J. Kim, Eds.; pp. 315–348). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-35085-6_12

Cimiano, P., Chiarcos, C., McCrae, J. P., & Gracia, J. (2020). Linguistic Linked Data: Representation, Generation and Applications. Springer International Publishing. https://doi.org/10.1007/978-3-030-30225-2

[29] Il modello è disponibile sulla repository Github del progetto: https://github.com/LiITA-LOD/LiITA_NLP_Models

McCrae, J. P., Gil, J., Gràcia, J., Bitelaar, P., & Cimiano, P. (2017). The OntoLex-Lemon Model: Development and Applications. https://www.semanticscholar.org/paper/The-OntoLex-Lemon-Model%3A-Development-and-McCrae-Gil/3ab2877e3cf9d8f7bad3a4fb9a03602010e00691

Favretti, R. R., Tamburini, F., & Santis, C. (2002). CORIS/CODIS: A corpus of written Italian based on a defined and a dynamic model.

Ide, N., & Pustejovsky, J. (2010). What Does Interoperability Mean, Anyway ? Toward an Operational Definition of Interoperability for Language Technology. in: Proceedings of the Second International Conference on Global Interoperability for Language Resources. Hong Kong, China, 2010.

Mauri, C., Ballarè, S., Goria, E., Cerruti, M., & Suriano, F. (2019). KIParla Corpus: A New Resource for Spoken Italian. Italian Conference on Computational Linguistics.

Miller, E. (1998). An Introduction to the Resource Description Framework. *D-Lib Magazine*, Journal of library administration, 34 (2001) pp. 245–255.

Passarotti, M., Mambrini, F., & Moretti, G. (2024). The Services of the LiLa Knowledge Base of Interoperable Linguistic Resources for Latin. In C. Chiarcos, K. Gkirtzou, M. Ionov, F. Khan, J. P. McCrae, E. M. Ponsoda, & P. M. Chozas (Eds.), Proceedings of the 9th Workshop on Linked Data in Linguistics @ LREC-COLING 2024 (pp. 75–83). ELRA and ICCL. https://aclanthology.org/2024.ldl-1.10/

Petrov, S., Das, D., & McDonald, R. (2012). A Universal Part-of-Speech Tagset. In N. Calzolari, K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, & S. Piperidis (Eds.), Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC`12) (pp. 2089–2096). European Language Resources Association (ELRA). https://aclanthology.org/L12-1115/

Pianta, E., Bentivogli, L., & Girardi, C. (2002). MultiWordNet: Developing an aligned multilingual database. In First International Conference on Global WordNet (pp. 293–302).

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In A. Celikyilmaz & T.-H. Wen (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations (pp. 101–108). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-demos.14

Roventini, A., Marinelli, R., & Bertagna, F. (2016). ItalWordNet v.2. http://www.ilc.cnr.it/it/content/italwordnet.

Serianni, L. (1989). Grammatica Italiana, UTET.

Dai *limina* a LiMINA: un *database* per i *marginalia* alla *Commedia*

Serena Malatesta[1], Beatrice Mosca[2].
[1] Università degli Studi di Padova - Université de Tours, Italia, serena.malatesta@phd.unipd.it
[2] Università eCampus – Universidad Complutense de Madrid, Italia, beatricevmosca@gmail.com

ABSTRACT (ITALIANO)

Il progetto PRIN 2022 Dante LIMINA, diretto da Elisabetta Tonello e Ciro Perna, si concentra sull'analisi dei segni marginali dai manoscritti alle edizioni a stampa, evidenziando come essi possano rivelare percorsi di lettura e processi di appropriazione. L'intervento si concentrerà sull'iniziativa centrale del progetto, LiMINA (Lost in Manuscripts: Ideas, Notes, Acknowledgments), un database relazionale open access che integra strumenti digitali avanzati con solidi principi filologici e codicologici, offrendo nuovi orizzonti interpretativi e un approccio transdisciplinare allo studio della *Commedia*. Il progetto si volge a promuovere un dialogo sinergico tra università, ricerca e istituzioni di conservazione. Attraverso questo network e grazie alla digitalizzazione e catalogazione sistematica, LiMINA mira a documentare e valorizzare le pratiche di lettura storiche, aprendo nuove prospettive per l'interpretazione e la divulgazione del poema dantesco con l'obiettivo di garantire l'accesso a ogni tipo di utenza, grazie a una piattaforma progettata per essere facilmente esplorabile, interoperabile e fruibile.

Parole chiave: *Marginalia*; *Divina Commedia*; Manus Online; IIIF; Database.

ABSTRACT (ENGLISH)

From limina *to LiMINA: A Database for the Marginalia of the* Commedia

The PRIN 2022 project Dante LIMINA, directed by Elisabetta Tonello and Ciro Perna, focuses on the analysis of these marginal signs, spanning from manuscripts to printed editions, highlighting how they can reveal reading trajectories and processes of appropriation. The presentation will focus on the project's core initiative, LiMINA (Lost in Manuscripts: Ideas, Notes, Acknowledgments), an open-access relational database that combines advanced digital tools with robust philological and codicological principles, offering new interpretative horizons and a transdisciplinary approach to the study of the *Commedia*. The project collaborates with institutions, creating a synergistic dialogue between universities, research and preservation institutions. Through this network and systematic digitization and cataloging, LiMINA aims to document and enhance historical reading practices, opening new perspectives for the interpretation and dissemination of Dante's poem, with the goal of ensuring access for all types of users through a platform designed to be easily navigable, interoperable and accessible.

Keywords: *Marginalia*; *Divine Comedy*; Manus Online; IIIF; Database.

1. INTRODUZIONE

La presente ricerca si propone di illustrare il lavoro fin ora compiuto per indagare quegli spazi marginali, fisici e concettuali, che ospitano tracce di lettura, annotazioni e interventi vari lasciati da lettori e copisti nel corso dei secoli intorno la *Commedia* di Dante Alighieri, che vengono individuati con il termine omnicomprensivo di *limina*. Partendo dal progetto PRIN 2022 Dante LIMINA, diretto da Elisabetta Tonello e Ciro Perna, si intende esplorare le molteplici dimensioni della ricezione dell'opera dantesca attraverso i segni "marginali". Come sottolineano i direttori scientifici, «*limina* rappresenta un contenitore che comprende tutte le tracce, verbali e non, che, se messe a sistema, possono rivelare i percorsi dei lettori a contatto con l'opera» (Perna & Tonello 2023: VII). Questo approccio si basa su un'attenta analisi del materiale testimoniale che evidenzia la ricchezza della tradizione manoscritta della *Commedia*. La metodologia adottata combina strumenti digitali avanzati con metodi di catalogazione tradizionale. L'obiettivo è duplice: da un lato, ricostruire la storia culturale e materiale della *Commedia* attraverso l'analisi dei suoi *marginalia*; dall'altro, mettere a disposizione della comunità accademica e del pubblico un *hub* capace di facilitare nuove interpretazioni e connessioni. La comunicazione andrà ad illustrare il progetto centrale, ovvero LiMINA (Lost in Manuscripts: Ideas, Notes, Acknowledgments). Il progetto è articolato in diverse fasi, che includono la schedatura dei manoscritti in collaborazione con il portale Manus Online, la digitalizzazione in alta risoluzione dei codici, l'importazione dei dati e la descrizione sistematica dei *limina*. Inoltre, grazie all'integrazione tra front-end e back-end, l'infrastruttura del portale offre un'interfaccia intuitiva che permette un'agile schedatura e di esplorare i manoscritti e avere una visione complessiva e dettagliata dei relativi *marginalia*. L'obiettivo è definire un workflow strutturato che

permette di aprire nuovi orizzonti interpretativi che mettano in luce il ruolo centrale dei *limina* nella ricezione dell'opera dantesca e definire un oggetto di ricerca che si presta ad essere applicato ad altri testi e opere significative. I risultati attesi includono la creazione di una piattaforma che documenta e valorizza le pratiche di lettura e appropriazione del poema, offrendo strumenti per una lettura innovativa e transdisciplinare della *Commedia*.

2. Da DanteLimina.it a LiMINA

Esistono zone ai margini del testo – sia esso manoscritto o a stampa – in cui palinsesti di voci e scritture si stratificano; tracce che, a posteriori, permettono di superare la consueta analisi testuale per accedere a questi spazi inesplorati. Essi consentono di ricostruire la storia della ricezione del testo, la sua circolazione e le sue vicende storico-culturali. I *limina* rappresentano tutto ciò che circonda il testo, sia in senso concreto che astratto: elementi che facilitano un accesso o un ingresso all'opera, permettendo di sviluppare una lettura innovativa e inedita della *Commedia*. Da tali presupposti nasce il progetto ancora *in fieri dantelimina.it*, un portale che funge da *hub* open access, concepito per raccogliere e analizzare i vari aspetti dei *limina* nella *Commedia* di Dante attraverso diversi progetti interconnessi. Tra le anime che completano l'universo dantelimina.it, vi sono *Dante Matrix*, un'applicazione digitale progettata per analizzare le affinità stemmatiche nei codici della *Commedia*; *D.A.N.T.E. (Digital Archive and New Technologies for E-content)* per la catalogazione delle edizioni illustrate della *Commedia*, dagli incunaboli fino ai giorni nostri insieme ad un'esperienza immersiva in realtà virtuale ispirata ai disegni e alle xilografie del XV secolo; *D.verse (Dante in Vernacular Experience: Readings, Studies, Exegesis)* in cui il focus si sposta sulle traduzioni e i commenti della *Commedia* nei dialetti italiani entro e oltre l'Unità d'Italia; infine, *Dante Juyō* che invece esplora una ricezione 'estrema' della *Commedia*, quella in Giappone, analizzando non solo il suo arrivo ma anche le reinterpretazioni contemporanee di Dante nella cultura j-pop giapponese. Al centro dell'iniziativa PRIN 2022 si colloca *LiMINA (Lost in Manuscripts: Ideas, Notes, Acknowledgments)*, oggetto principale del nostro intervento, che si propone di indagare i *limina* dei manoscritti dell'antica vulgata della *Commedia*, soglie dei manoscritti, intese come margini che vivono e si configurano quale spazio privilegiato di comunicazione tra testo, autore e lettore. È proprio in tali soglie, negli elementi che incorniciano il testo - il *peritesto* (Genette 1982), che il libro assume la propria forma e si attua concretamente l'esperienza della lettura. L'analisi dei *marginalia* consente infatti di approfondire la ricezione testuale della *Commedia*, concentrandosi su quegli spazi bianchi nei quali il lettore instaura un rapporto dialogico con l'autore e, al contempo, agisce nel proprio presente. Comprendere l'identità dei lettori della *Commedia* e le loro possibili reazioni consente non solo di misurare l'impatto culturale dell'opera, ma anche di individuare le modalità con cui ciascun lettore — differente per cultura, estrazione sociale e interessi — si confronta con il testo (si veda anche Fasano 2017, 2019). L'esame di tratti codicologici, avvertenze di bottega, *marginalia* figurati, verbali e non verbali rende possibile la delineazione di un Lettore, il quale si manifesta proprio nelle soglie del codice. Un simile approccio si rivela non solo metodologicamente fondante, ma estremamente significativo dal punto di vista filologico e letterario: consente infatti di far emergere le dinamiche di lavoro all'interno delle botteghe, di ricostruire le pratiche dei copisti e di cogliere le differenze nei modi di trasmissione e ricezione del testo tra ambienti colti e di committenza borghese o mercantesca. In tal modo, lo studio dei margini e del peritesto si configura come uno strumento privilegiato per restituire un autentico spaccato della storia letteraria, evidenziando come il testo dantesco sia stato non solo letto, ma anche riscritto e annotato dai suoi lettori. Inoltre, la disponibilità di questo materiale in formato digitale apre prospettive nuove per la ricerca e la divulgazione: consente un accesso più ampio e democratico a una documentazione altrimenti difficilmente consultabile, favorendo il riuso critico da parte di studiosi e utenti interessati e facilita in modo decisivo la sistematizzazione e la condivisione dei risultati, aggiungendo un nuovo tassello al dialogo tra filologia, storia della cultura e *digital humanities*.

3. Landing

Il progetto LiMINA è accessibile attraverso il portale dantelimina.it cliccando sull'icona predisposta (Fig. 1) nella home page del sito, l'utente viene reindirizzato alla *landing page* dedicata (Fig. 2). Questa è stata progettata per garantire una coerenza grafica che si integrasse con l'estetica generale del portale, rafforzando l'identità del progetto e promuovendo al contempo un dialogo tra tradizione e innovazione. Il design UX/UI è stato concepito per offrire un'interfaccia chiara e intuitiva, capace di accogliere sia studiosi che appassionati in uno spazio digitale che unisce rigore scientifico e accessibilità. Un elemento centrale del design è rappresentato dalle icone personalizzate, realizzate da Angelo Rauso, illustratore, comic artist

e graphic designer. La scelta di affidare a un giovane la progettazione di questi elementi non è stata puramente funzionale, ma nasce dalla volontà di valorizzare e promuovere una visione integrata che collega il mondo accademico al territorio, le iniziative scientifiche alle esperienze artistiche. Le icone, che costituiscono un tratto distintivo della *landing page*, sono il risultato di un dialogo continuo tra l'artista e il team di lavoro. Durante questo processo, sono state tradotte in elementi visivi le suggestioni e le esigenze concettuali del progetto, dando vita a un linguaggio grafico unico, capace di combinare semplicità formale e profondità simbolica. Le icone, disegnate con linee essenziali e un'estetica moderna, rappresentano un equilibrio ideale tra immediatezza comunicativa e coerenza visiva, rendendosi facilmente riconoscibili e funzionali per orientare gli utenti nella navigazione del portale.



Fig. 1 dantelimina.it

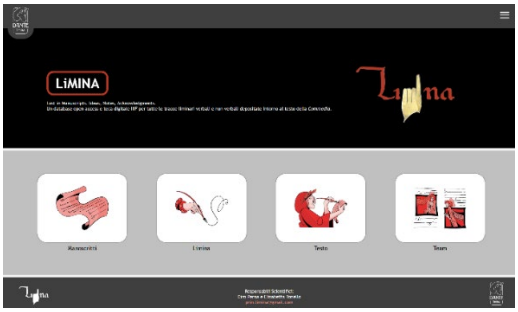


Fig. 2 Landing page limina.dantelimina.it

4. Front-end

Per agevolare l'esplorazione del sito, la landing page è suddivisa in quattro sezioni principali che permettono una navigazione fluida e facilitano l'accesso ai diversi contenuti del progetto:

- *Manoscritti*: in cui è possibile consultare la scheda paleografica e codicologica del manoscritto, navigare tra i suoi *limina* e i relativi riferimenti bibliografici;
- *Limina*: offre la possibilità di consultare i *marginalia* attraverso diversi filtri di ricerca che consentono di adattare l'indagine alle esigenze dell'utente. È possibile effettuare una ricerca testuale libera, esplorando il contenuto del *database* senza vincoli, oppure selezionare un manoscritto tra quelli già schedati e inseriti. Si può inoltre specificare la tipologia del *limen*, distinguendo tra *limina verbali* e *non verbali*, oppure focalizzarsi su uno specifico copista, annotatore o lettore. Un ulteriore strumento di ricerca è il filtro "testo della *Commedia*", che consente di ricercare in modo mirato un luogo specifico del poema. Infine, è possibile indicare una datazione per periodi, uno strumento utile per svolgere un'analisi cronologica delle tracce marginali e comprendere meglio la loro evoluzione nel tempo.
- *Testo*: questa parte del database permette di consultare il testo della *Commedia*. I direttori scientifici del progetto, consapevoli delle numerose edizioni della *Commedia*, hanno deciso di rendere disponibile il testo della *Commedia* di Giorgio Petrocchi (Petrocchi 1966-1967) consci che a oggi è l'edizione scientifica di riferimento a livello nazionale, nonostante certi presupposti assunti dell'editore ad oggi sembrino superabili.
- *Team*: in cui si trovano tutte le informazioni sul progetto e i relativi collaboratori.

a. Limina

La sezione centrale del progetto è rappresentata da *Limina*, l'interfaccia che consente di esplorare il database tramite query preimpostate. Se, ad esempio, si seleziona un manoscritto specifico, viene restituita una lista ordinata dei *limina* identificati, secondo la sequenza delle carte. L'interfaccia mostra un'anteprima dell'area annotata e consente l'accesso diretto alla digitalizzazione completa tramite Mirador, nonché alla scheda descrittiva redatta dal catalogatore.

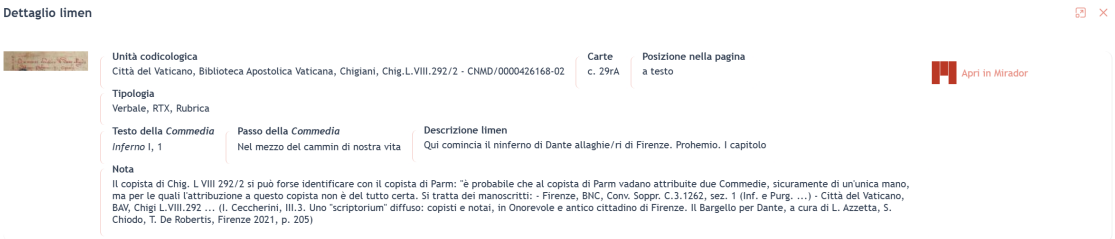


Fig. 3 Esempio di dettaglio del *limen* (Marzo 2025).

Ciascun *limen* è visualizzabile singolarmente e corredato da informazioni come segnatura, tipologia, datazione (puntuale o approssimativa), lingua, autore o annotatore. Oltre alla segnatura del manoscritto, vengono fornite tutte le informazioni: il tipo di classificazione in base alle sue caratteristiche, distinguendo tra macro-categorie *verbali* e *non verbali*, e corredandola di una datazione, che può essere puntuale o desumibile a seconda delle informazioni ricavabili dal testimone. Inoltre, è inclusa l'indicazione dell'annotatore. In caso di personalità non identificate, si utilizza la formula "Mano di (segnatura del manoscritto)". Al contrario, qualora si riconosca la mano di chi interviene sul testo o se il testimone stesso fornisce indicazioni mediante una firma o una nota, viene creato un autority file.

I *limina* sono suddivisi in *verbali* e *non verbali*, e ulteriormente classificati in base alla relazione o meno con il testo e con il testimone secondo le categorie sottostanti:



Fig. 4 Classificazione dei *limina*.

Questa varietà offre allo schedatore la possibilità di descrivere in modo puntuale il *limen*, mentre consente allo studioso di identificare con precisione le caratteristiche del *limen* con cui si sta confrontando, indirizzando così la propria ricerca in maniera sempre più mirata. Lo schedatore può anche avvalersi delle sezioni dedicate alle note, per garantire un'esposizione chiara e scientificamente fondata. Per garantire la massima flessibilità e accuratezza, si è deciso, di comune accordo tra i direttori scientifici del progetto e i collaboratori, di introdurre una sezione *"altro"* per ciascuna *macro-tipologia* di *limen*. Questa scelta mira a evitare forzature interpretative e classificazioni improprie di *limina* che, a causa della loro originalità o univocità, non rientrano in nessuna delle *tipologie* predefinite.
È importante sottolineare che le *tipologie* descrittive adottate non sono state definite a priori, ma rappresentano il risultato di un lavoro collaborativo e dinamico. Questo processo ha coinvolto informatici, esperti del settore, direttori scientifici e catalogatori, i quali, nel corso delle loro campagne di studio, ricerca e sistematizzazione dei materiali, hanno segnalato tutte le casistiche che inizialmente non prevedevano un'etichetta specifica nel database. Tale approccio ha consentito un miglioramento progressivo e continuo del progetto, garantendo una maggiore precisione nella descrizione e nella classificazione dei *limina*.

5. Workflow informatico e back-end catalogatori

Il progetto LiMINA si sviluppa attraverso un workflow strutturato in quattro fasi principali volti al raggiungimento degli obiettivi di ricerca e alla creazione di un *database* open access e interoperabile. L'intero sistema è stato realizzato come un'unica applicazione web integrata, sviluppata in PHP 8.2, che gestisce sia il back-end sia il front-end in modo modulare. L'applicazione è strutturata secondo una logica MVC (Model–View–Controller) e si appoggia a un database MySQL, con interazione mediata da Eloquent ORM e query personalizzate via facade DB. Di particolare rilievo è il sistema Ariadne, sviluppato per semplificare la gestione back-end di relazioni complesse tra tabelle (uno-a-molti, molti-a-molti), integrato in interfacce dinamiche via JavaScript e con supporto AJAX. Ariadne si compone di modelli predefiniti per controller, view e servizi, e consente di operare agevolmente con griglie di selezione, checkbox, campi disabilitati e interazioni asincrone centralizzate.

Schedatura dei manoscritti.

La prima fase del progetto si concentra sulla catalogazione dettagliata dei manoscritti. Per garantire un approccio uniforme e sistematico, il team ha scelto di collaborare con l'ICCU (Istituto Centrale per il Catalogo Unico) avvalendosi del portale Manus Online (MOL). Questa collaborazione ha portato alla creazione di un progetto speciale accessibile all'indirizzo https://manus.iccu.sbn.it/limina e all'introduzione di nuovi campi specifici nella catalogazione, pensati per segnalare e descrivere i *marginalia* contenuti nei codici descritti estendendo l'oggetto di ricerca oltre i manoscritti della *Commedia*.

Digitalizzazione in alta risoluzione del materiale.

La digitalizzazione in alta risoluzione costituisce un pilastro essenziale del progetto, per consentire l'accesso remoto ai manoscritti da parte di studiosi e appassionati e garantire uno studio approfondito dei segni marginali da parte dei catalogatori. In collaborazione con Space S.p.A. e con le biblioteche partecipanti, sono stati definiti standard rigorosi per la digitalizzazione. I manoscritti non ancora digitalizzati sono stati acquisiti in formato .tiff piramidale, che assicura un'elevata qualità delle immagini e una gestione ottimale per la visualizzazione online. I file vengono archiviati in un repository interno, predisposto per comunicare attraverso il protocollo IIIF (International Image Interoperability Framework - https://iiif.io/) e consentire la visualizzazione tramite *Mirador* (https://projectmirador.org/index.html) il viewer open-source, web based integrato tramite CDN (Content Delivery Network).

Importazione dei contenitori (manoscritto).

La struttura dati di LiMINA è stata inizialmente concepita su fogli di calcolo CSV, successivamente convertiti da Luigi Tessarolo in un *database* MySQL per la gestione e l'interrogazione dei dati. I dati relativi ai manoscritti vengono importati da MOL tramite le API JSON messe a disposizione dalla stessa piattaforma, in modo tale che ogni manoscritto sia corredato da una scheda catalografica completa e da un CNMD, un identificatore univoco, che assicura l'interoperabilità tra LiMINA e Manus Online. A ciascun item manoscritto viene associato un URI IIIF, generato internamente o già disponibile in un sito esterno. Questo approccio permette la piena integrazione con il visualizzatore Mirador, che è stato incorporato nell'interfaccia del portale per consentire agli utenti di esplorare e manipolare in dettaglio le immagini digitalizzate e operare confronti tra più carte

Descrizione dei limina.

Il cuore del progetto è costituito dallo studio e dalla catalogazione degli elementi liminari, come descritto nel paragrafo 4. Il risultato dello studio dei catalogatori viene inserito nel backend dei catalogatori, che permette di immettere tutte le informazioni rilevanti per ogni *limen* con un ID univoco e di comunicare con il database in modo efficace e preciso. Quest'ultimo è strutturato in una serie di tabelle interconnesse che permettono di organizzare, catalogare e visualizzare i dati in modo coerente e strutturato.

Tutte le tabelle principali condividono un approccio uniforme, con campi essenziali come un identificativo univoco (id), timestamp di creazione e aggiornamento (created_at e updated_at), e il riferimento all'utente che ha creato il record tramite la chiave esterna user_id. Ciò garantisce tracciabilità e controllo nella gestione dei dati. Il cuore di questa sezione è la tabella *marginals*, che raccoglie le informazioni principali relative a ciascuna annotazione. Ogni record descrive non solo il contenuto del marginale, ma anche la sua precisa localizzazione fisica (es. lato del foglio, colonna, ecc.) e testuale, la lingua utilizzata, l'autore, la datazione, eventuali note descrittive e il manoscritto e il testo a cui l'annotazione è riferita.

Il sistema è arricchito da tabelle collegate che permettono una classificazione articolata delle annotazioni. La tabella *marginal_types* consente di indicare la tipologia (glossa, nota lessicale, rimando, ecc.), mentre *marginal_nature* ne descrive la natura funzionale o stilistica (esplicativa, scolastica, decorativa, ecc.). Inoltre, grazie all'associazione con la tabella *names* tramite la pivot *marginal_name*, è possibile collegare ciascun limen a figure specifiche come copisti, commentatori o personaggi menzionati. Infine, la relazione con la tabella *text_name* consente di contestualizzare ogni annotazione rispetto al contenuto testuale di riferimento. Questo sistema relazionale garantisce una descrizione accurata e multilivello degli elementi liminari e consente anche una navigazione trasparente e scalabile dei dati, a beneficio sia dei ricercatori coinvolti nella catalogazione sia degli utenti finali che fruiranno delle edizioni digitali e delle schede descrittive.

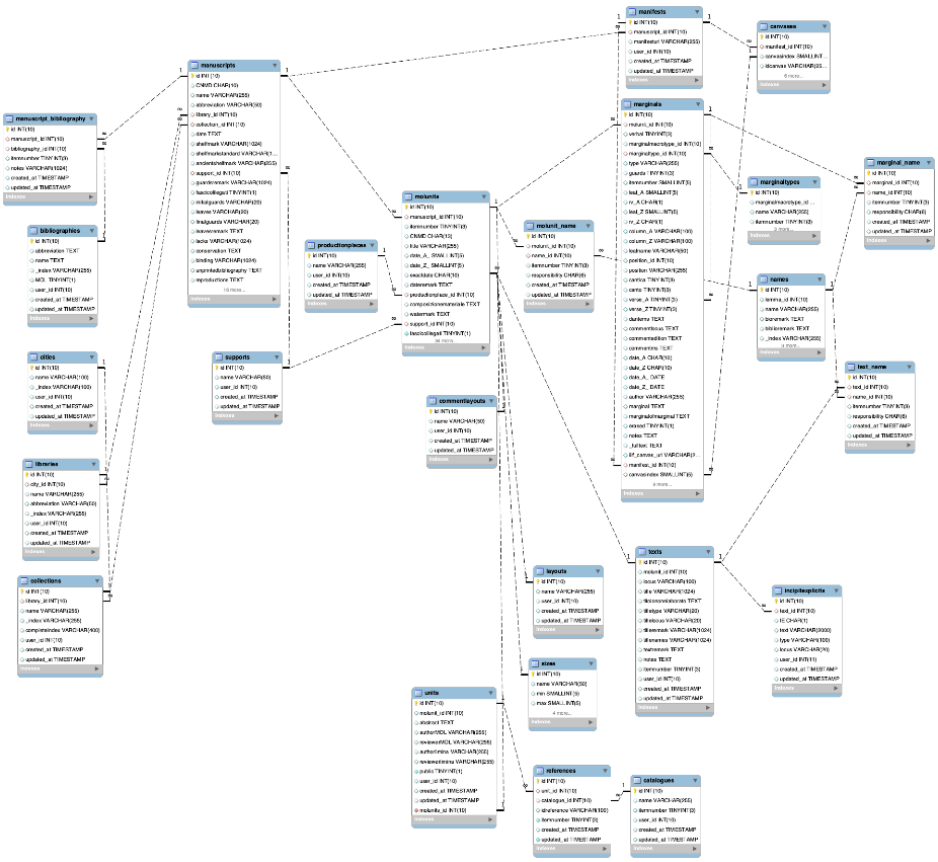


Fig.5. Schema ER (Entity-Relationship) del database di Limina

Conclusioni

L'approccio interdisciplinare adottato dal progetto LiMINA ha permesso di affrontare in modo efficace le sfide legate alla conservazione, catalogazione e valorizzazione dei marginalia. Gli elementi liminari, pur trovandosi ai margini del testo, si rivelano tutt'altro che secondari: essi restituiscono informazioni fondamentali sulla storia materiale del manoscritto, sull'interazione dei lettori/copisti con il testo, e sull'evoluzione della ricezione della *Commedia* nel tempo. La loro corretta descrizione e classificazione consente di ricostruire una fitta rete di pratiche culturali, intellettuali e interpretative, che arricchiscono la comprensione dell'opera e della sua fortuna.

Per quanto riguarda gli sviluppi futuri, una possibile direzione sarà quella di affiancare al modello relazionale attuale un sistema di RDF, rendendoli compatibili con il paradigma dei Linked Open Data e massimizzando i vantaggi dell'architettura esistente e delle tecnologie semantiche.

Un'ulteriore linea di sviluppo riguarda il potenziale analitico del progetto: una volta consolidata la fase di raccolta e normalizzazione dei dati, sarà possibile avviare operazioni di analisi e confronto, anche con strumenti (semi)automatici basati su tecnologie di Natural Language Processing, reasoning ontologico o modelli di rappresentazione della conoscenza. In questa prospettiva, sarà importante tenere conto anche della possibilità di individuare dati in tensione (es. commenti o annotazioni che propongono interpretazioni divergenti dello stesso passo), fenomeno comune nella ricezione letteraria e fonte preziosa di riflessione critica. Infine, si immagina uno scenario in cui l'intelligenza artificiale potrà supportare l'interrogazione diretta del database, fornendo risposte precise e contestualizzate a domande complesse. In questa direzione, risultano particolarmente promettenti gli approcci basati sul *semantic parsing* per la generazione automatica di *semantic query graphs* a partire da domande formulate in linguaggio naturale (es. modello SPEDN come in Wei et al. 2023), che combina tecniche neurali e strutture grafiche per l'interpretazione e la traduzione semantica delle interrogazioni su basi di conoscenza RDF.

Questa prospettiva aprirebbe nuovi orizzonti per la fruizione e la ricerca, rendendo il progetto uno strumento di conservazione e un laboratorio dinamico per l'interpretazione del patrimonio testuale.

BIBLIOGRAFIA

Fasano, G. 2017. *La ricezione della* Commedia*: Studio dei* marginalia *nei manoscritti medievali*. Revista de la Sociedad Española de Italianistas, 11, 91-109.

Fasano, G. 2019. Marginalia *figurata e ricezione testuale della* Commedia *dantesca*: González Martín, V., Núñez García, L., Bianchi, M., García-Pérez, M. (Eds.) 2019. *Literatura y cultura italianas entre Humanismo y Renacimiento*. Ediciones Universidad de Salamanca, pp. 91-109.

Genette G., *Palimpsestes: La littérature au second degré*, Seuil, Paris, 1982.

Perna, C., & Tonello, E. 2023. *Premessa*. Storie e Linguaggi, 9, VII-VIII.

Petrocchi, G. (Eds.). 1966-1967. *La Commedia secondo l'antica vulgata*. Le Lettere, Firenze.

Wei, S., et al. 2023. *Semantic Parsing for Question Answering over Knowledge Graphs*. 10.48550/arXiv.2401.06772.

Dai Materiali Didattici alle Piattaforme FAIR: Costruire un'Infrastruttura di Training in H2IOSC

Giulia Pedonese[1], Francesca Frontini[2], Roberta Ottaviani [3], Federico Boschetti [4], Alessia Spadi[5], Lucia Francalanci[6], Alessia Scognamiglio[7], Pietro Restaneo[8], Antonina Chaban[9], Jana Striova[10], Laura Benassi[11]

[1] Istituto di Linguistica Computazionale "Antonio Zampolli" (CNR-ILC), Italia – giulia.pedonese@cnr.it
[2] Istituto di Linguistica Computazionale "Antonio Zampolli" (CNR-ILC), Italia – francesca.frontini@cnr.it
[3] Istituto di Linguistica Computazionale "Antonio Zampolli" (CNR-ILC), Italia – roberta.ottaviani@cnr.it
[4] Istituto di Linguistica Computazionale "Antonio Zampolli" (CNR-ILC), Italia – federico.boschetti@cnr.it
[5] Istituto Opera del Vocabolario Italiano (CNR-OVI), Italia – alessia.spadi@cnr.it
[6] Istituto Opera del Vocabolario Italiano (CNR-OVI), Italia – lucia.francalanci@cnr.it
[7] Istituto per la Storia del Pensiero Filosofico e Scientifico Moderno (CNR-ISPF), Italia – alessia.scognamiglio@cnr.it
[8] Istituto per il Lessico Intellettuale Europeo e Storia, delle Idee (CNR-ILIESI), Italia – pietro.restaneo@cnr.it
[9] Istituto Nazionale di Ottica (CNR-INO), Italia – antonina.chaban@cnr.it
[10] Istituto Nazionale di Ottica (CNR-INO), Italia – jana.striova@cnr.it
[11] Istituto Nazionale di Ottica (CNR-INO), Italia – laura.benassi@cnr.it

ABSTRACT (ITALIANO)

Questo contributo si propone di illustrare la progettazione e lo sviluppo di un'infrastruttura di formazione innovativa per le Scienze Umane e Sociali, basata sui principi FAIR e sulla promozione della Scienza Aperta, nell'ambito del progetto Humanities and cultural Heritage Italian Open Science Cloud (H2IOSC). L'obiettivo principale è la creazione di un ecosistema integrato che renda i materiali didattici facilmente reperibili, accessibili, interoperabili e riutilizzabili. A tal fine, sono state implementate due piattaforme: *H2IOSC Virtual Environment*, dedicata all'erogazione di corsi e risorse per studenti, e *H2IOSC Training Library*, un deposito per la conservazione e la condivisione di materiali didattici modulari. Entrambe le piattaforme si basano sulla metodologia "FAIR-by-Design" raccomandata dal progetto Skills4EOSC, che struttura il processo educativo in sei fasi, garantendo standard elevati di metadatazione e l'uso di formati aperti. Con l'implementazione di queste piattaforme, i cui servizi saranno resi disponibili a un livello di aggregazione più alto nel *Marketplace* di H2IOSC, il progetto intende favorire un approccio scalabile e sostenibile alla formazione, promuovendo al contempo la collaborazione tra docenti e studenti.
Parole chiave: formazione; gestione dei dati; infrastrutture di ricerca; principi FAIR; Scienza Aperta.

ABSTRACT (ENGLISH)

From Learning Materials to FAIR Platforms: Building a Training Infrastructure in H2IOSC.

This paper aims to illustrate the design and development of an innovative training infrastructure for the Social Sciences and Humanities based on FAIR principles and the promotion of Open Science, within the Humanities and cultural Heritage Italian Open Science Cloud (H2IOSC) project. The main objective is the creation of an integrated ecosystem that makes learning materials easily findable, accessible, interoperable and reusable. To this end, two platforms have been implemented: H2IOSC Virtual Environment, dedicated to the delivery of courses and student resources, and H2IOSC Training Library, a repository for the storage and sharing of modular teaching materials. Both platforms are based on the 'FAIR-by-Design' methodology recommended by the Skills4EOSC project, which structures the educational process in six steps, ensuring high standards of metadata and the use of open formats. By implementing these platforms, whose services will be made available at a higher aggregation level in the H2IOSC Marketplace, the project intends to foster a scalable and sustainable approach to education, while promoting collaboration between teachers and students.
Keywords: training; data management; research infrastructures; FAIR principles; Open Science.

1. CONTESTO E OBIETTIVI

Il contributo si inserisce nell'ambito della formazione gestita dal *Work Package* 8 (WP8) del progetto *Humanities and cultural Heritage Italian Open Science Cloud*[1] (H2IOSC), progetto PNRR che mira a creare un cluster federato di servizi e risorse sviluppati dai nodi nazionali di quattro infrastrutture di ricerca europee per la Scienza Aperta nel settore delle Scienze Umane e Sociali: CLARIN, DARIAH, E-RIHS e OPERAS (Degl'Innocenti, 2023). Nell'ambito della formazione, le iniziative sono spesso disperse su piattaforme differenti, non sempre descritte con metadati specifici e accessibili al pubblico. Per questo il

[1] https://www.h2iosc.cnr.it/

progetto H2IOSC ha previsto l'implementazione di una metodologia condivisa (Pedonese, 2024) per strutturare i materiali didattici in modo che siano utilizzabili in accordo con i principi FAIR di gestione dei dati e di un'infrastruttura composta da due piattaforme: *H2IOSC Virtual Environment* per l'erogazione e la fruizione del training da parte degli utenti (identificati come studenti universitari, ricercatori, docenti delle scuole secondarie, professionisti del settore culturale e professionisti a sostegno della ricerca come Data Stewards e Data librarians); e *H2IOSC Training Library*, un *repository* specifico per la pubblicazione di materiali didattici modulari in formato editabile che permette l'assegnazione di *Persistent dentifiers* (PID), licenze standard (ad esempio Creative Commons) e una gestione integrata per l'aggiornamento delle versioni, in modo che materiali didattici siano accessibili e riutilizzabili per docenti universitari e formatori professionali. Il contributo dettaglia la progettazione delle due piattaforme e il loro design integrato secondo la metodologia condivisa.

2. LA METODOLOGIA DIDATTICA CONDIVISA DI H2IOSC

Il progetto H2IOSC si è allineato al progetto Skills4EOSC[2] adottando la metodologia "FAIR-by-Design" (Filiposka, 2023) per lo sviluppo di risorse didattiche. Questa metodologia, basata sui principi FAIR (*Findable*, *Accessible*, *Interoperable*, *Reusable*), segue un processo in sei fasi (preparazione, scoperta, design, sviluppo, pubblicazione, e controllo qualità) e integra strumenti pratici come checklist e modelli per garantire un approccio strutturato e scalabile. In particolare, H2IOSC ha adottato la definizione di "*learning object*" proposta da Skills4EOSC, considerandolo come una lezione modulare caratterizzata da obiettivi di apprendimento specifici che include non solo attività, ma anche strumenti di valutazione delle competenze acquisite. Inoltre, tutte le risorse didattiche di H2IOSC sono descritte con il *Minimal Metadata Set for Learning Resources* proposto dalla Research Data Alliance (Hoebelheinrich, 2022) così come arricchito dal modello CLARIN (van der Lek, 2025), che aggiunge campi come ad esempio *contributors*, *workload in* ECTS e *Persistent Identifier* (PID) per migliorare l'utilizzabilità e la documentazione delle risorse. Per garantire la massima cercabilità dei materiali didattici, si è deciso di compilare il set di metadati di ogni risorsa inizialmente in lingua inglese (per mantenere l'allineamento con Skills4EOSC), prevedendo comunque una gestione multilingue dei metadati, utilizzando anche le risorse terminologiche multilingui di CLARIN (Frontini, 2021; Borek, 2020). Per una maggiore riusabilità, H2IOSC ha scelto di convertire il materiale esistente in formati aperti come *plaintext*, per garantire flessibilità e riuso, e di utilizzare Markdown per la creazione di nuovi contenuti. Le licenze saranno il più aperte possibile, in linea con le raccomandazioni di Skills4EOSC, per favorire il riutilizzo. Inoltre, H2IOSC ha sviluppato pratiche di citazione formalizzate per consentire la tracciabilità e il riuso dei materiali, lavorando in stretta collaborazione con comunità scientifiche come AIUCD per promuovere standard condivisi e sostenibili.
Uno dei punti cardine della metodologia di Skills4EOSC è l'applicazione dei principi FAIR alla progettazione didattica sotto due punti di vista: da un lato quello dello studente, che fruisce del prodotto finito su piattaforme di e-learning; dall'altro quello del formatore, che ha la necessità di poter trovare, accedere e riutilizzare materiali didattici in formati editabili a seconda delle proprie esigenze. Per questo il progetto H2IOSC ha previsto la produzione di due diverse tipologie di oggetti digitali: corsi ad accesso aperto in modalità e-learning rivolti agli studenti in formati compatibili con i maggiori programmi di lettoscrittura e materiali didattici modulari disponibili per il download in formati aperti e modificabili secondo le necessità dei formatori. Per la fruizione di questi due tipi di oggetti digitali, il progetto H2IOSC ha previsto l'implementazione di due diverse piattaforme: un *learning management system* chiamato *H2IOSC Training Environment* per l'accesso ai corsi da parte degli studenti e un *repository* di materiali didattici per formatori chiamato *H2IOSC Training Library*.

3. H2IOSC TRAINING ENVIRONMENT

La piattaforma *H2IOSC Training Environment* è progettata per la pubblicazione e la fruizione di materiale didattico strutturato in corsi, moduli e lezioni. I corsi sono destinati alle persone (studenti, professionisti, ricercatori ecc.) che intendono acquisire o migliorare le proprie competenze nell'ambito delle scienze umane e sociali, con particolare attenzione agli strumenti messi a disposizione dalle infrastrutture di ricerca afferenti ad H2IOSC: per questo i nodi italiani di CLARIN, DARIAH, E-RIHS e OPERAS hanno prodotto un'offerta di corsi in autoformazione relativi, ad esempio, alla gestione di dati linguistici secondo i

[2] https://www.skills4eosc.eu/

principi FAIR e Open Science, alle Digital Humanities per la didattica e all'uso di modelli per comprendere la pittura a olio e il patrimonio artistico.[3]
La piattaforma è stata sviluppata da ETT S.p.A. in collaborazione con l'Istituto CNR di Linguistica Computazionale "Antonio Zampolli" di Pisa (CNR-ILC) sulla base di una progettazione condivisa dalle quattro infrastrutture in modo da offrire un'esperienza completa sia per gli studenti sia per i docenti afferenti ai diversi domini di ricerca.
Dal lato *front-end*, la piattaforma consente agli studenti la creazione di un profilo personale che raccoglie i corsi seguiti e le certificazioni conseguite, l'iscrizione spontanea a corsi, sia ad accesso libero sia ad accesso chiuso (in quest'ultimo caso mediante un codice fornito dal docente) e l'interazione con altri utenti tramite funzionalità integrate di chat e forum.
Dal *back-end*, la piattaforma mette a disposizione dei docenti un sistema di *editing* intuitivo per organizzare il materiale didattico su tre livelli gerarchici: 1) il livello corso, che consente di definire il *syllabus* generale utilizzando una griglia conforme al modello di metadati RDA (Sezione 2); 2) il livello modulo, che rappresenta un livello intermedio dedicato alla descrizione di un sottogruppo di attività didattiche seguendo il medesimo schema di metadati del corso; 3) il livello dei materiali didattici, che riguarda i contenuti specifici di ciascun modulo, da organizzare in cartelle per le singole lezioni oppure da caricare direttamente come file in formato testo, video o immagine; in alternativa, è possibile integrare collegamenti a lezioni online tramite servizi esterni (es. Big Blue Button, Teams, Zoom, Google Meets), che verranno gestiti come oggetti della piattaforma nella sezione "eventi" e nel calendario integrato. Grazie a queste funzionalità, *H2IOSC Training Environment* rappresenta una soluzione modulare che rende il processo di insegnamento-apprendimento interattivo e versatile.

4. H2IOSC LIBRARY

H2IOSC Training Library è una piattaforma in fase di sviluppo grazie alla collaborazione tra M.E.T.A. S.r.l. e CNR-ILC. Si tratta di un repository, attualmente a nostra conoscenza unico nel suo genere, esclusivamente dedicato al deposito FAIR di materiale didattico modulare, organizzato secondo la metodologia proposta da Skills4EOSC opportunamente adattata alle esigenze di H2IOSC (Sezione 2).
Il design della piattaforma è stato testato su un materiale didattico di esempio, cioè il corso *Introduction to Language Data: Standards and Repositories* (van der Lek, 2023) originariamente sviluppato da CLARIN per il progetto UPSKILLS[4], in seguito tradotto e riadattato per la comunità italiana nell'ambito di H2IOSC. Seguendo le linee guida adottate, i contenuti del corso tradotto, dal titolo *Introduzione ai Dati Linguistici: Standard e Archivi Digitali* (van der Lek, 2024), sono stati convertiti in file .md (*Markdown*) e strutturati in una gerarchia che consente agli utenti di individuare e scaricare singole parti del corso in base alle proprie necessità. La piattaforma, infatti, è stata progettata per offrire un *backoffice* che semplifica il deposito del materiale didattico, supportando chi desidera rendere i propri contenuti disponibili.
L'interfaccia della *H2IOSC Training Library* consente di selezionare vari contenitori organizzati gerarchicamente, dal più generale al più specifico: il *learning path*, cioè la radice che si collega al catalogo delle infrastrutture di ricerca partecipanti a H2IOSC; il *Corso*, che è il contenitore più ampio, e relativi contenitori intermedi: *sezione*, *unità* e *modulo*, che permettono di selezionare porzioni specifiche di materiale da scaricare, fino al contenitore minimo che rappresenta la porzione didattica più elementare, chiamata *learning object* (Sezione 2). Gli unici oggetti obbligatori per il deposito sono: 1) il *learning path*, che funge da radice e include documenti essenziali per facilitare il riuso del materiale didattico (quali il file README, la citazione in formato .cff, il *code of conduct*, la *facilitation guide* e la licenza applicata al materiale, insieme al *syllabus* e a eventuali note di rilascio che accompagnano la gestione delle versioni successive); e 2) il *learning object*, cioè il nodo foglia che ha la funzione di contenere i materiali didattici in formato aperto e riproducibile. Ognuno di questi sei contenitori è descritto con il set di metadati RDA, con l'opzione di ereditare i metadati già compilati nelle sezioni sovraordinate.
La piattaforma include inoltre un'interfaccia di ricerca con funzionalità avanzate per la scoperta del materiale didattico depositato che si rivolge a utenti interessati al riuso. Tutti i materiali didattici saranno consultabili liberamente, mentre il *download* sarà possibile solo previa autenticazione tramite un sistema di autenticazione federata. L'interfaccia di ricerca offrirà la possibilità di effettuare ricerche semplici in linguaggio naturale, con i risultati ordinati in base alle parole chiave inserite nella sezione apposita dei metadati. Inoltre, un sistema a faccette permetterà di affinare ulteriormente i risultati della ricerca,

[3] Il catalogo completo e aggiornato dell'offerta formativa disponibile sulla H2IOSC Training Environment è disponibile all'indirizzo: https://www.h2iosc.cnr.it/training-infrastructure/
[4] https://upskillsproject.eu/

utilizzando una gamma più ampia di criteri come la data di pubblicazione, la lingua, il nome dell'autore, il formato e l'argomento principale.
Le funzionalità di entrambe le piattaforme, relative alla fruizione, al deposito e alla scoperta di materiale didattico modulare puntano a massimizzarne il riutilizzo in accordo con i principi FAIR di gestione dei dati e le buone pratiche di Scienza Aperta. Nella prossima sezione si anticiperanno le modalità di accesso integrato a queste funzionalità attraverso il Marketplace H2IOSC.

5. IL TRAINING NEL MARKETPLACE H2IOSC

Il *Marketplace* H2IOSC è una piattaforma progettata con l'obiettivo di offrire un punto di accesso centralizzato a dati, strumenti e servizi relativi alle scienze umane, sociali e culturali (Sichera, 2024). Consente l'integrazione e l'aggregazione di risorse interne ed esterne grazie a protocolli standard e API dedicate che permettono agli utenti di condividere, tra le altre risorse, anche materiali formativi. La sua struttura si ispira al *Social Sciences and Humanities Marketplace*[5], con cui condivide una tipologia simile di oggetti (Barbot, 2024).
In un tale contesto, il *Marketplace* di H2IOSC accoglierà, insieme ad altre tipologie di dati, i metadati con cui sono state descritte le risorse didattiche sulle due piattaforme, rispettivamente *H2IOSC Training Environment* per i **corsi** diretti agli studenti e *H2IOSC Training Library* per i **materiali didattici** riutilizzabili da docenti e formatori. Tali metadati saranno allineati con quelli presenti nel modello dati del *Marketplace* anche se, naturalmente, le rispettive piattaforme mostreranno una metadatazione a grana più fine rispetto a quella del *Marketplace*, che si configura come una *discovery platform.* Inoltre, per quanto riguarda i vocabolari controllati associati a specifici metadati, questi saranno tratti dai vocabolari dello SSHOC *Marketplace*, con eventuali traduzioni in italiano.
Nel Marketplace saranno integrate funzionalità di ricerca tra oggetti di tipo diverso, per cui l'utente avrà a disposizione un'interfaccia di che potrà restituire, sulla base di un'interrogazione con chiave di ricerca, diverse tipologie di oggetti, comprendendo i materiali didattici di *H2IOSC Training Library* (suddivisi in corsi, sezioni, unità, moduli e *learning object*) ed eventi formativi o corsi in autoformazione disponibili su *H2IOSC Training Environment*, che saranno integrati nel *Marketplace* attraverso un calendario. Il *Marketplace* inoltre supporta la creazione di *workflow* complessi, in cui i materiali didattici possono essere aggregati in sequenze di apprendimento. Questo è particolarmente utile per formatori che desiderano integrare più risorse in un unico percorso formativo e garantirà l'interoperabilità con altre infrastrutture di *training*, come DARIAH-Campus,[6] CLARIN Learning Hub[7] e SSHOC Training Discovery Toolkit.[8]
In questo modo, il *Marketplace* H2IOSC permetterà alle piattaforme di training di creare un ecosistema integrato e dinamico, in grado di supportare l'apprendimento continuo e la condivisione delle conoscenze nel contesto delle Scienze Umane, Sociali e Culturali.

RINGRAZIAMENTI

Gli autori e le autrici dichiarano di avere utilizzato l'intelligenza artificiale generativa, in particolare ChatGPT, allo scopo di riassumere le informazioni, tradurre e migliorare il testo del presente abstract.

Progetto H2IOSC - Humanities and cultural Heritage Italian Open Science Cloud finanziato dall'Unione europea NextGenerationEU - Piano Nazionale di Ripresa e Resilienza (PNRR) - Missione 4 "Istruzione e Ricerca" Componente 2 "Dalla ricerca all'impresa" Linea di Investimento 3.1 "Fondo per la realizzazione di un sistema integrato di infrastrutture di ricerca e innovazione" Azione 3.1.1 "Creazione di nuove IR o potenziamento di quelle esistenti che concorrono agli obiettivi di Eccellenza Scientifica di Horizon Europe e costituzione di reti" - Codice progetto IR0000029 - CUP B63C22000730005. Soggetto attuatore CNR.

BIBLIOGRAFIA

Barbot, L., Dolinar, M., Gray, E. J., Grisot, C., Illmayer, K., Kurzmeier, M., & McGillivray, B. (2024). Contextualizing Research Tools & Services Through Workflows in the SSH Open Marketplace. *Journal of Open Humanities Data*, *10*(1). https://doi.org/10.5334/johd.192

[5] https://marketplace.sshopencloud.eu/
[6] https://campus.dariah.eu/
[7] https://www.clarin.eu/content/learning-hub
[8] https://sshopencloud.eu/ssh-training-discovery-toolkit

Borek, L., Hastik, C., Khramova, V., & Geiger, J. (s.d.). *TaDiRAH Taxonomy of Digital Research Activities in the Humanities*. Social Sciences & Humanities Open Marketplace. Recuperato 21 gennaio 2025, da https://marketplace.sshopencloud.eu/tool-or-service/w6OC9p

Degl'Innocenti, Emiliano, Monica Monachini, Alberto Bucciero, Enrico Pasini, Bruno Fanini, e Francesca Frontini. «H2IOSC: Humanities and Heritage Open Science Cloud». In La memoria digitale: forme del testo e organizzazione della conoscenza. Atti del XII Convegno Annuale AIUCD, a cura di Emmanuela Carbé, Gabriele Lo Piccolo, Alessia Valenti, e Francesco Stella, 63–64, 2023. https://iris.unive.it/retrieve/0f226d38-e332-418b-9b14-d5558d1a0d9d/AIUCD2023.pdf.

Filiposka, S., Green, D., Mishev, A., Kjorveziroski, V., Corleto, A., Napolitano, E., Paolini, G., Di Giorgio, S., Janik, J., Schirru, L., Gingold, A., Hadrossek, C., Souyioultzoglou, I., Leister, C., Pavone, G., Sharma, S., Mendez Rodriguez, E. M., & Lazzeri, E. (2023). *D2.2 Methodology for FAIR-by-Design Training Materials*. https://zenodo.org/records/8305540

Frontini, F., Gamba, F., Monachini, M., & Broeder, D. (2021). *SSHOC Multilingual Data Stewardship Terminology* [Dataset]. http://hdl.handle.net/20.500.11752/ILC-567

Hoebelheinrich, N. J., Biernacka, K., Brazas, M., Castro, L. J., Fiore, N., Hellström, M., Lazzeri, E., Leenarts, E., Martinez Lavanchy, P. M., Newbold, E., Nurnberger, A., Plomp, E., Vaira, L., van Gelder, C. W. G., & Whyte, A. (2022). *Recommendations for a minimal metadata set to aid harmonised discovery of learning resources*. https://doi.org/10.15497/RDA00073

Pedonese, G., Frontini, F., Ottaviani, R., Boschetti, F., Spadi, A., Francalanci, L., Scognamiglio, A., Restaneo, P., Chaban, A., Striova, J., & Benassi, L. (2024). *Materiali didattici come oggetti digitali FAIR: Una metodologia condivisa per la formazione in H2IOSC*. https://doi.org/10.6092/unibo/amsacta/7927

Sichera, P., Marras, C., & Pasini, E. (2024). *ORCHESTRAZIONE API PER WORKFLOW APPLICATIVI NELL'AMBITO DELLE DIGITAL HUMANITIES*. Zenodo. https://doi.org/10.5281/zenodo.14187534

van der Lek, Iulianna; Fišer, Darja. (2023). *Introduction to Language Data: Standards and Repositories.* In UPSKILLS Learning Content.https://upskillsproject.eu/project/standards_repositories/. CC BY 4.0.

van der Lek, I., Fišer, D., Frontini, F., & Pedonese, G. (2024, agosto 31). Introduzione ai Dati Linguistici: Standard e Archivi Digitali. Zenodo. https://doi.org/10.5281/zenodo.13911935

van der Lek, I., Frontini, F., Fiser, D., & König, A. (2025). Making the CLARIN Training Materials FAIR-By-Design. CLARIN Annual Conference (CLARIN2023), Leuven. Zenodo. https://doi.org/10.5281/zenodo.14699078

IlluminAI: un sistema di navigazione interattivo per i manoscritti miniati rinascimentali

Valeria Minisini[1], Giorgio Gosti[2], Bruno Fanini[3]
[1] Sapienza Università di Roma, Italia – valeria.minisini@uniroma1.it
[2] CNR-ISPC, Italia – giorgio.gosti@cnr.it
[3] CNR-ISPC, Italia – bruno.fanini@cnr.it

ABSTRACT (ITALIANO)

Nell'ultimo decennio, la digitalizzazione massiccia del patrimonio culturale ha reso consultabile una quantità inedita di volumi manoscritti, difficilmente accessibili per ragioni conservative, riversando in rete migliaia di pagine e dando nuovo impulso agli studi comparativi sull'evoluzione dei loro apparati decorativi. La conseguente necessità di sviluppare strumenti automatici e intuitivi per agevolare la ricerca di specifiche componenti ha portato a individuare nell'intelligenza artificiale un imprescindibile campo di sperimentazione non solo per il riconoscimento dei caratteri ma anche per l'analisi del layout e l'identificazione dei soggetti rappresentati nelle miniature figurative. Il contributo presenta "IlluminAI", un'interfaccia per il confronto tra manoscritti miniati basata su una rete neurale profonda addestrata con un dataset originale al fine di creare un sistema di visualizzazione immersiva per la valorizzazione dei codici rinascimentali attraverso le raffigurazioni contenute e la loro iconografia specifica. Dopo una breve introduzione per contestualizzare il progetto, verrà descritto il modello con particolare attenzione ai dati e alla tecnica usata per l'addestramento prima di esporre i risultati finora ottenuti. Concluderemo quindi presentando il sistema di visualizzazione in fase di elaborazione e i futuri sviluppi previsti.
Parole chiave: transfer learning; reti neurali profonde; manoscritti miniati; iconografia; realtà virtuale

ABSTRACT (ENGLISH)

IlluminAI: an interactive navigation system for Renaissance illuminated manuscripts

In the last decade, the massive cultural heritage digitization has made available an unprecedented quantity of manuscript volumes that would be difficult to access for conservation reasons, pouring thousands of pages online and giving new impetus to comparative studies on the evolution of their decorative apparatuses. The resulting need to develop intuitive automatic tools to facilitate the search for specific components has led to identify artificial intelligence as an essential field of experimentation not only for character recognition but also for layout analysis and the identification of subjects represented in figurative miniatures. The paper presents "IlluminAI", a comparison interface for illuminated manuscripts based on a deep neural network trained on an original dataset with the aim of creating an immersive visualization system for the enhancement of the Renaissance codices through the illustrations contained and their specific iconography. After a brief introduction to contextualize the project, the model will be described, with particular attention to the data and the technique used for training before presenting the results obtained so far. We will conclude by exposing the visualization system under development and the project future steps.
Keywords: transfer learning; deep neural network; illuminated manuscripts; iconography; virtual reality

1. INTRODUZIONE

L'autonomia della miniatura come genere artistico è una conquista relativamente recente, consolidata dagli studi dello storico dell'arte Otto Pächt (1902-1988) e dal relativo ciclo di lezioni tenute presso l'Università di Vienna nel semestre invernale del 1967-68, in seguito pubblicate in un celeberrimo volume (Pächt, 1984). Sebbene nell'immaginario collettivo la produzione dei codici miniati sia tradizionalmente associata al periodo medievale, alcuni dei volumi più riccamente decorati giunti fino a noi risalgono al Rinascimento che segna l'ultima grande stagione della miniatura in Europa, resa fiorente da artisti specializzati e pittori come Lorenzo Monaco, Taddeo Crivelli e Beato Angelico (Alexander, 1994).

Questa particolare tipologia di bene culturale, difficilmente accessibile fisicamente a causa della fragilità dei suoi materiali, ha trovato nuova vita grazie a diversi progetti di digitalizzazione (Cappellotto, 2020) che stanno mettendo a disposizione degli studiosi e del grande pubblico un'inedita quantità di opere raramente in precedenza uscite dai depositi. L'inevitabile perdita parziale delle informazioni strettamente legate alla fisicità dell'oggetto, non replicabili nel facsimile virtuale (Campagnolo, 2024) anche quando acquisito con tecniche avanzate (Pietroni et al., 2023), è compensata dalla visibilità ottenuta su portali facilmente

raggiungibili e dotati di visualizzatori come Mirador[1] capaci di comparare esemplari ripresi da fonti differenti grazie all'ormai ampia adozione dell'*International Image Interoperability Framework* (IIIF)[2]. Nonostante ciò, l'interesse collettivo verso i codici miniati è ancora contenuto rispetto all'attrazione esercitata da altre tipologie di opere continuando a provenire, per la maggior parte, da un numero relativamente esiguo di esperti a causa delle barriere linguistiche esistenti e della complessità insita nel decifrare la scrittura manoscritta.
Sono osservabili oggi due diverse tendenze nella digitalizzazione dei volumi, legate alla tipologia di istituzione proprietaria del bene e al suo pubblico di riferimento: se le biblioteche riproducono infatti l'intero codice, i musei si limitano a creare copie digitali solamente delle pagine più interessanti dal punto di vista artistico, ovvero quando presenti scene con figure, decorazioni e capilettera ornamentali. Mentre il primo approccio preserva l'artefatto nella sua completezza fisica e contenutistica rendendolo però un prodotto impegnativo da esplorare anche se opportunamente indicizzato, il secondo è più idoneo a catturare l'attenzione dell'utente comune che naviga quotidianamente sul web ma, per farlo, riduce drasticamente il volume alla sua sola qualità estetica.
Oltre alla componente testuale che costituisce, al tempo stesso, il principale veicolo e il maggiore ostacolo alla comunicazione dei contenuti, proprio le miniature che più affascinano il pubblico non accademico possono essere utilizzate per costruire percorsi esperienziali accessibili all'interno dei codici attraverso un linguaggio visivo padroneggiato indipendentemente dal paese di provenienza e dal grado di istruzione posseduto.
Il contributo presenta la rete neurale profonda alla base di "IlluminAI", un'interfaccia di visualizzazione immersiva in fase di sviluppo progettata per valorizzare i volumi manoscritti digitalizzati agevolando lo spostamento tra le pagine grazie al riconoscimento automatico degli elementi in esse contenuti, comparare esemplari diversi e mettere in relazione le eventuali illustrazioni presenti con riproduzioni di opere d'arte dal medesimo soggetto sfruttando la comune iconografia.
Dopo alcuni brevi cenni allo stato dell'arte, si procederà a presentare la composizione del dataset appositamente formato e il sistema di classificazione delle immagini adottato prima di esporre la tecnica elaborata per addestrare il modello attraverso un approccio di apprendimento automatico interattivo (Porter et al., 2013) impiegato per annotare rapidamente tre gruppi di immagini dalle dimensioni crescenti. Infine, nella sezione conclusiva dopo i risultati finora ottenuti verranno prospettati i futuri sviluppi.

2. DEEP LEARNING E MANOSCRITTI

L'impiego del machine learning e delle reti neurali profonde per facilitare lo studio dei documenti storici è un campo di ricerca ormai consolidato (Lombardi & Marinai, 2020) che si differenzia da altri settori della computer vision per problemi specifici, aggiungendo agli ostacoli propri del riconoscimento automatico dei caratteri quelli dell'individuazione di particolari elementi all'interno delle immagini (Esposito et al., 2008). La maggioranza dei dataset riguardanti i testi manoscritti e a stampa antichi sono formati a partire da una lingua condivisa, dal periodo di creazione e l'istituzione proprietaria delle risorse (Nikolaidou et al., 2022) qualificandosi spesso come poco rappresentativi e non performanti su materiale diverso da quello d'addestramento. A differenza di quanto accade poi con le immagini naturali, i giornali (Lee et al., 2020) e le collezioni museali (Ypsilantis et al., 2022), la dimensione è sempre relativamente contenuta con pochi esemplari completi al loro interno come nel dataset HBA (Mehri et al., 2017) formato a partire da undici volumi disponibili su Gallica, il portale digitale della Biblioteca Nazionale di Francia (BnF). Nonostante l'aumento delle risorse disponibili, i dataset possono necessitare anche di venire integrati con nuovi elementi creati sinteticamente (Rahal et al., 2024), strategia particolarmente utile nella ricerca focalizzata sulle decorazioni e sull'analisi del layout.
Oltre al riconoscimento dei caratteri (Islam & Iacob, 2023), l'applicazione più studiata con decine di sperimentazioni pubblicate, l'attenzione su elementi grafici come i capilettera ornamentali per l'estrazione delle lettere contenute (Coustaty et al., 2011) o sulle figure per la classificazione dei gesti attraverso rilevatori basati sul template (Schlecht et al., 2011) ha contribuito a spostare gradualmente l'interesse dalla parola alle decorazioni che l'accompagnano. Per quanto riguarda nello specifico le miniature, la loro diversa cromia è stata sfruttata per individuarle (Borghesani et al., 2014) e tentativi di riconoscimento

[1] Per approfondire la piattaforma open-source Mirador: https://projectmirador.org/ (cons. 09/04/2025).

[2] Maggiori informazioni sul Framework IIIF sono disponibili sul sito ufficiale: https://iiif.io/ (cons. 09/04/2025).

iconografico sono stati effettuati impiegando l'algoritmo di object detection YOLO (Manoni, 2023) con buoni risultati, soprattutto se usato contestualmente il sopracitato standard IIIF (Aouinti et al., 2022). Una piattaforma online particolarmente interessante con il medesimo scopo, inoltre, è Mandragore (Jacquot, 2018) che attinge materiale dalle collezioni della BnF e dotata, dal 2022, di una nuova versione con una buona ma tradizionale interfaccia grafica[3].
Nel campo storico-artistico, tra le diverse tipologie di reti neurali (Zhao et al., 2021), l'architettura ResNet (He et al., 2016) è stata spesso utilizzata (Milani & Fraternali, 2021) per le buone performance normalmente ottenute nell'attività di etichettatura e rilevazione oltre a essere facilmente addestrabile, anche su esigue quantità di dati, in quanto la sua implementazione di funzioni residue consente una migliore propagazione degli errori.

3. IL PROGETTO

3.1 La composizione del dataset

Per rendere interattivo il sistema di visualizzazione immersiva "IlluminAI", un'apposita rete neurale profonda come Minisini et al. (2024) è stata addestrata su un dataset originale al fine di individuare le pagine manoscritte in gruppi eterogenei di riproduzioni, identificare i diversi elementi di layout al loro interno oltre a distinguere le eventuali illustrazioni presenti anche da motivi decorativi e iniziali ornamentali. A tale scopo, il training del modello è stato effettuato non soltanto su riproduzioni di codici ma considerando un più articolato repertorio contenente testi a stampa, opere d'arte per la maggior parte bidimensionali quanto oggetti frequentemente presenti nei cataloghi museali e spesso decorati anch'essi con soggetti figurativi.
In totale, il dataset si compone ad oggi di 46.014 immagini provenienti esclusivamente da database istituzionali come la Library of Congress di Washington, il Metropolitan Museum of Art di New York e la J. Paul Getty Museum Collection di Los Angeles, ma anche da realtà europee come la Staatsbibliothek di Berlino, il Catalogo Generale dei Beni Culturali italiano e la Koninklijke Bibliotheek dei Paesi Bassi la cui collezione è ampiamente visibile sul portale Europeana. Il ricorso a fonti diverse ha consentito di avere un buon grado di eterogeneità nelle condizioni di scansione e scatto nonché nella qualità delle riproduzioni, sempre comunque materiale in pubblico dominio o protetto da licenze Creative Common CC-BY e CC-BY-NC scelte in previsione del rilascio finale.
Essendo il progetto incentrato sui manoscritti, il numero di immagini a loro dedicate costituisce il 68% del totale a cui si aggiungono le riproduzioni di incunaboli e volumi a stampa, attestate sul 15%, mentre il restante 17% è formato da opere d'arte e oggetti musealizzati. Le pagine, in particolare, provengono da più di duecentotrenta volumi differenti, centotrentacinque dei quali sono stati inclusi nella loro interezza, con una media di trecento immagini ciascuno, per coprire la più ampia varietà di layout e stili decorativi possibile. Compresi in un ampio arco cronologico, esteso dal IX al XX secolo, hanno principalmente tema sacro e presentano illustrazioni sia a pagina intera che dalle dimensioni più contenute e inserite nel corpo del testo.
Ogni riproduzione è stata classificata utilizzando un sistema alfanumerico che permette di risalire rapidamente all'ente proprietario e ottenere informazioni in merito al soggetto raffigurato, al periodo storico di realizzazione quanto all'autore o, se sconosciuto, almeno alla sua area geografica d'origine. Tramite poi l'aggiunta di un prefisso relativo alla tipologia di bene digitalizzato, il materiale è stato inoltre suddiviso in due macrocategorie, volumi (P) e arte (A), quanto in dieci sottogruppi: fogli manoscritti (P.01), pagine stampate (P.02), dipinti (A.01), incisioni (A.02), disegni (A.03), sculture (A.04), vetrate (A.05), arazzi (A.06), stampe d'arte (A.07) e oggetti di natura differente (A.08) come monete, ceramiche antiche e porcellane.
Un esempio di classificazione così prodotta è "P.01_S(Mark)_14BMGetty" dove "P.01" indica la natura di foglio manoscritto, "S(Mark)" identifica il soggetto della miniatura presente come San Marco Evangelista mentre "14" è il secolo di creazione, "BM" le iniziali dell'autore e "Getty" è la collezione di provenienza. Per le pagine, la sequenza che segue la specifica della raffigurazione può essere comune a più esemplari in quanto associa in maniera univoca tutti i fogli originariamente parte dello stesso volume con il codice identificativo a esso assegnato, a sua volta collegato a un record contenente le principali informazioni bibliografiche, il permalink alla risorsa primaria e l'eventuale IIIF Manifest.

[3] Il portale è raggiungibile al seguente indirizzo: https://mandragore.bnf.fr/ (cons. 09/04/2025).

Il sistema sinora esposto ha consentito di contenere significativamente il rischio di inserire copie multiple di un'opera in casi come le incisioni e, allo stesso tempo, associare con facilità le immagini dotate di una medesima raffigurazione per aumentare i soggetti scarsamente rappresentati.
Le pagine prive di miniature sono state invece classificate sulla base della presenza di decorazioni o, in assenza di queste, di testo e spartiti musicali considerando la caratteristica dominante per bilanciare prima la composizione del dataset e, successivamente, i tre gruppi usati per addestrare il modello dalle dimensioni progressivamente crescenti di 400, 4.000 e 41.614 elementi.

3.2 L'addestramento del modello

Il primo passo nella creazione del modello è stato implementare un ciclo d'addestramento incrementale iterativo articolato in quattro fasi per limitare l'apporto umano e accelerare il processo di etichettatura che, in analoghi progetti, costituisce uno dei momenti più dispendiosi temporalmente.
Il procedimento adottato è il seguente: innanzitutto, un esperto di dominio annota o corregge le etichette di un dataset più piccolo che un esperto d'intelligenza artificiale impiega in seguito per addestrare la rete neurale utilizzando la tecnica del transfer learning e l'ottimizzazione degli iperparametri. Il modello ottenuto, quindi, viene usato per prevedere le etichette di un dataset più grande che l'esperto di dominio successivamente controlla per identificare specifici bias del sistema e, quando necessario, provvede a correggere manualmente ricorrendo alla piattaforma open-source Label Studio (Tkachenko et al., 2025) prima che il processo venga ripetuto per l'insieme successivo.
Nello specifico, inizialmente sono state annotate manualmente solo le immagini facenti parti del più contenuto gruppo da 400 esemplari per segnalare la presenza o co-presenza nelle pagine di illustrazioni (Fig), decorazioni (Deco), testo (Text), spartiti musicali (Mus) e capilettera ornamentali (Let) con un minimo di una fino a un massimo di quattro etichette per un totale di 818. L'annotazione "Fig" è stata usata per contraddistinguere anche tutte le immagini della categoria "Arte" oltre a venire assegnata alle lettere abitate e istoriate.
Quanto ottenuto è stato ripartito destinando il 70% per il training e il restante 30% al test del modello, basato sull'architettura ResNet, al quale sono stati dati come compiti nella prima sperimentazione la distinzione tra volumi e arte, fogli manoscritti e pagine a stampa oltre al rilevamento della presenza dei cinque elementi di layout sopra elencati.
La tecnica del transfer learning, ricorrendo sia al Fine Tuning su tutti i pesi (Fine-Tuning) che mantenendo le caratteristiche degli strati più profondi fisse e riaddestrando solamente l'ultimo strato completamente connesso (Fixed Features), è stata impiegata con una ResNet-18 preaddestrata su ImageNet (Russakovsky et al., 2015) scelta tenendo in considerazione le dimensioni contenute del dataset a disposizione e la tendenza dei modelli con pochi parametri a essere meno soggetti a incorrere in un regime di sovraparametrizzazione. Sono stati anche applicati due metodi distinti di addestramento con discesa del gradiente: discesa del gradiente stocastica (SGD) e Adam. Per SGD abbiamo utilizzato un exponential learning rate scheduler con velocità di apprendimento iniziale di 0,001, momentum 0,9, dimensione dello step learning rate scheduler 7 e gamma 0,1. Nel caso di Adam, invece, è stato fissato un learning rate pari a 0,001.
In quasi tutti i casi, il Fine Tuning su tutti i pesi e SGD hanno dato i risultati migliori anche se la differenza è contenuta. Il tasso di accuratezza registrato con questa combinazione è particolarmente elevato nella distinzione delle pagine dall'arte (97%) e del materiale manoscritto da quello a stampa (95%) quanto nell'individuazione di testo (96%) e musica (97%). Il valore più basso, invece, è stato registrato per le iniziali miniate (82%) a causa principalmente dall'ambiguità intrinseca tra semplici lettere ornamentali e iniziali istoriate.
Considerati i risultati promettenti, i modelli con il più elevato tasso di accuratezza per ogni target sono stati utilizzati per prevedere automaticamente le etichette del secondo gruppo, ottenendo buoni risultati nel rilevamento di elementi visivi diversi dal testo come decorazioni e illustrazioni, soprattutto nel caso di fogli con bordi ornamentali importanti oppure occupati da scene a pagina intera. Una certa confusione, però, persiste tra le iniziali e le miniature inserite nel corpo del testo, spesso confuse dal sistema e classificate come "Let".
Su 4.000 immagini solo 1.529 (38%) hanno richiesto un intervento manuale, risparmiando diverse ore di lavoro per raggiungere un risultato di 7.204 annotazioni complessive. Confrontando le previsioni con le etichette corrette, una particolare criticità è stata rilevata per gli spartiti musicali a stampa, non riconosciuti in quanto tali, con un impatto significativo sul numero di etichette assegnate dal sistema per "Mus", solo 251 sulle 358 esatte, quanto ripercussioni su "Text", passato da 3.170 a 2.855. Anche

l'etichetta "Let" ha subito una flessione, da 966 a 933, mentre le classi "Fig" e "Deco" sono aumentate, rispettivamente da 1.488 a 1.762 e da 1.214 a 1.296.
La precisione delle previsioni è stata complessivamente considerata soddisfacente con un tasso del 97% per il riconoscimento delle miniature figurative, dello 88% per il testo, del 83% per le decorazioni ma solamente del 66% per le iniziali miniate.
Nonostante lo sbilanciamento di alcune categorie, abbiamo comunque ottenuto dei modelli con un buon livello di discriminazione e si è quindi proceduto ad addestrare nuovamente la rete neurale adottando la medesima suddivisione di 70% per il training e 30% per il test.
Mentre con il primo gruppo abbiamo potuto però mantenere il learning rate costante, nel secondo sia per SGD che per Adam è stato necessario variare il valore tra 0,00001 e 0,1 per raggiungere i risultati visibili in Figura 1. Nel caso del Fine Tuning e SGD, rispetto l'insieme precedente è stato osservato un miglioramento significativo nell'accuratezza registrata nel riconoscimento delle decorazioni (92%) e delle lettere ornamentali che, seppur rimanendo il valore più basso, raggiungono anch'esse il 91%.

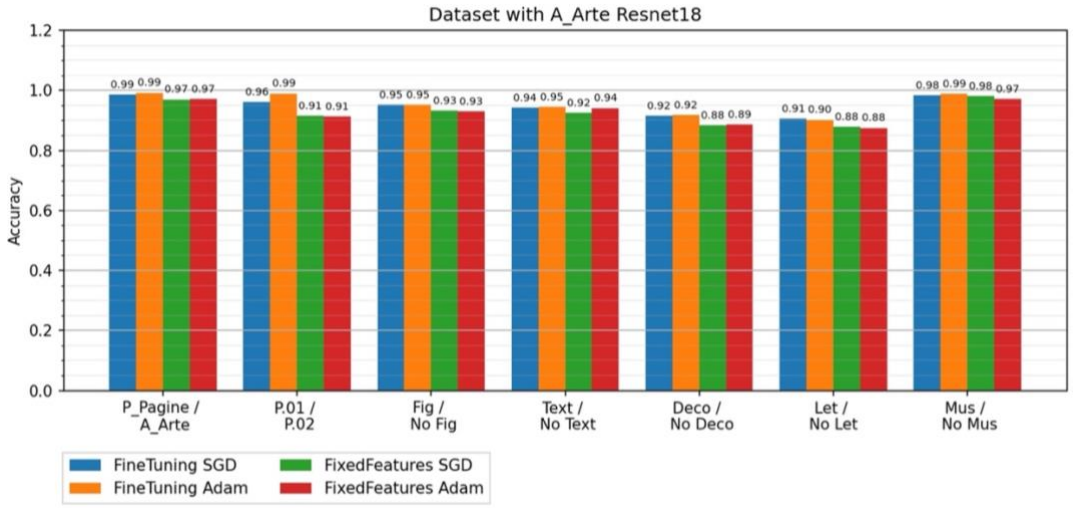


Figura 1. Grafico dei risultati ottenuti con il secondo gruppo da 4.000 immagini.

L'ultimo passaggio nel processo di addestramento ha visto ripetere quanto finora esposto per le rimanenti immagini del terzo e più consistente gruppo in cui, a differenza dei precedenti incentrati sulla figura umana, sono confluiti in numero consistente paesaggi e vedute cittadine quanto tipologie di opere quali vasi, reperti archeologici, arredi e rilegature di volumi, tutti parte della sottoclasse generica A.08.
Questa volta, solamente il 35% delle previsioni ottenute sulle 41.614 immagini analizzate è incorsa in modifica, portando a un totale di 63.677 etichette di cui 48.719 assegnate alle sole pagine manoscritte con una media di due a esemplare. Una flessione negativa tra tag previsti e corretti è stata osservata per "Let", passato con la revisione da 6.949 a 6.737, mentre per le altre etichette si è registrato un aumento con "Fig" che sale da 12.978 a 13.665, "Deco" da 9.032 a 9.217, "Text" da 29.394 a 29.428 e "Mus" da 4.604 a 4.630. Attraverso un esame approfondito dei falsi positivi è emersa la tendenza del modello a interpretare erroneamente gli elementi vegetali e le architetture quanto il persistere delle difficoltà nel riconoscere la presenza di note musicali nelle pagine a stampa oltre al testo.
Il tasso di precisione del modello, infine, si è contraddistinto per un generale miglioramento raggiungendo il 99% per il testo, le illustrazioni e gli spartiti musicali, il 96% per le decorazioni mentre per le iniziali miniate è stato registrato un più netto incremento arrivando a 88%.

4. CONCLUSIONI E FUTURI SVILUPPI

Ad oggi, la prima fase del progetto si è conclusa e i test di riconoscimento sugli elementi di layout svolti finora ci hanno dato risultati molto incoraggianti poiché livelli di accuratezza superiori all'80% non si ottengono spesso con dataset così piccoli e, ancora più raramente, vengono mantenuti incrementando l'entità e la varietà all'interno dei gruppi. La presenza di alcune categorie sottorappresentate non sembra avere impattato significativamente sui valori e, pertanto, non si è ritenuto per il momento necessario ricorrere alla data augmentation o a metodi generativi per la creazione di ulteriori esempi sintetici.
L'obiettivo, una volta riaddestrato il modello per l'ultima volta con tutte le immagini etichettate, è passare gradualmente dalla classificazione alla segmentazione semantica delle miniature utilizzando le zone di

attivazione ottenute come guida per posizionare automaticamente le caselle sulle aree delle pagine da elaborare. Le maschere così ricavate saranno utilizzate per arricchire un sistema web3D immersivo e interattivo, basato sul framework open-source ATON (Fanini et al., 2021), sviluppato dal CNR-ISPC e tra i servizi del progetto H2IOSC - *Humanities and cultural Heritage Italian Open Science Cloud*, che metterà l'utente in condizione di ispezionare e interrogare spazialmente grandi quantità di pagine in un ambiente 3D virtuale, accostare immagini simili e confrontarle con altre opere di analogo soggetto attraverso l'uso di comuni web browser.
Inoltre, le riproduzioni saranno corredate da informazioni e spiegazioni per contestualizzare i manoscritti non solo a livello storico ma culturale. In questo modo intendiamo dimostrare come, attraverso l'uso delle nuove tecnologie, sia possibile elaborare strategie di valorizzazione innovative capaci di riutilizzare materiale già acquisito e rendere artefatti, originariamente prerogativa di pochi, accessibili a tutti liberando un delicato bene culturale dai suoi limiti fisici.

RINGRAZIAMENTI

Questa ricerca è finanziata dal Progetto H2IOSC - *Humanities and cultural Heritage Italian Open Science Cloud* (https://www.h2iosc.cnr.it/, ultimo accesso 09/04/2025), finanziato dall'Unione europea NextGenerationEU - Piano Nazionale di Ripresa e Resilienza (PNRR), Mission 4 "Education and Research", Componente 2 "From research to business", Investment 3.1 "Fund for the realization of an integrated system of research and innovation infrastructures", Action 3.1.1 "Creation of new research infrastructures strengthening of existing ones and their networking for Scientific Excellence under Horizon Europe", codice progetto IR0000029-CUP-B63C22000730005. Ente di implementazione CNR.

BIBLIOGRAFIA

Alexander, J. G. J. (1994). *The Painted Page: Italian Renaissance Book Illumination, 1450-1550*, Prestel Pub.

Aouinti, F., Eyharabide, V., Fresquet, X., & Billiet, F. (2022). Illumination Detection in IIIF Medieval Manuscripts Using Deep Learning. *Digital Medievalist*, *15*(1), 1–18. https://doi.org/10.16995/dm.8073

Borghesani, D., Grana, C., & Cucchiara, R. (2014). Miniature illustrations retrieval and innovative interaction for digital illuminated manuscripts. *Multimedia Systems*, *20*, 65–79. https://doi.org/10.1007/s00530-013-0315-3

Campagnolo, A. (2024). Codificare il codice. La materialità del libro e la sua digitalizzazione verso una codicologia digitale. *Umanistica Digitale*, *8*(17), 1–23. https://doi.org/10.6092/issn.2532-8816/16694

Cappellotto, A. (2020). From codex to apps: the medieval manuscript in the age of its digital reproduction. *Umanistica Digitale*, *4*(9), 1–18. https://doi.org/10.6092/issn.2532-8816/11459

Coustaty, M., Pareti, R., Vincent, N., & Ogier, J.-M. (2011). Towards historical document indexing: extraction of drop cap letters. *International Journal on Document Analysis and Recognition (IJDAR)*, *14*(3), 243–254. https://doi.org/10.1007/s10032-011-0152-x

Esposito, F., Ferilli, S., Basile, T. M. A., & Di Mauro, N. (2008). Machine Learning for Digital Document Processing: from Layout Analysis to Metadata Extraction. In S. Marinai & H. Fujisawa (Eds.), *Studies in Computational Intelligence: Vol. 90. Machine Learning in Document Analysis and Recognition* (pp. 105-138). https://doi.org/10.1007/978-3-540-76280-5_5

Fanini, B., Ferdani, D., Demetrescu, E., Berto, S., & D'Annibale, E. (2021). ATON: An Open-Source Framework for Creating Immersive, Collaborative and Liquid Web-Apps for Cultural Heritage. *Applied Sciences*, *11*(22), Article 11062. https://doi.org/10.3390/app112211062

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 770–778. https://doi.org/10.1109/CVPR.2016.90

Islam, M. A., & Iacob, I. E. (2023). Manuscripts Character Recognition Using Machine Learning and Deep Learning. *Modelling*, *4*(2), 168–188. https://doi.org/10.3390/modelling4020010

Jacquot, O. (2018, October 22). Mandragore, base des manuscrits enluminés de la BnF. *Carnet de recherche de la Bibliothèque nationale de France*. https://doi.org/10.58079/m3f3

Lee, B. C. G., Mears, J., Jakeway, E., Ferriter, M., Adams, C., Yarasavage, N., Thomas, D., Zwaard, K., &

Weld, D. S. (2020). The Newspaper Navigator Dataset: Extracting Headlines and Visual Content from 16 Million Historic Newspaper Pages in Chronicling America. *CIKM '20: Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, New York, NY, USA, 3055–3062. https://doi.org/10.1145/3340531.3412767

Lombardi, F., & Marinai, S. (2020). Deep Learning for Historical Document Analysis and Recognition—A Survey. *Journal of Imaging*, *6*(10), Article 110. https://doi.org/10.3390/jimaging6100110

Manoni, P. (2023). AI4MSS: un esperimento di intelligenza artificiale alla Biblioteca Apostolica Vaticana. In G. Bergamin & T. Possemato (Eds.), *Guardando oltre i confini: partire dalla tradizione per costruire il futuro delle biblioteche: studi e testimonianze per i 70 anni di Mauro Guerrini* (pp. 231-244). AIB - Associazione Italiana Biblioteche. https://doi.org/10.1400/294065

Mehri, M., Héroux, P., Mullot, R., Moreux, J.-P., Coüasnon, B., & Barrett, B. (2017). HBA 1.0: A Pixel-based Annotated Dataset for Historical Book Analysis. *HIP '17: Proceedings of the 4th International Workshop on Historical Document Imaging and Processing*, New York, NY, USA, 107–112. https://doi.org/10.1145/3151509.3151528

Milani, F., & Fraternali, P. (2021). A Dataset and a Convolutional Model for Iconography Classification in Paintings. *Journal on Computing and Cultural Heritage*, *14*(4), Article 46. https://doi.org/10.1145/3458885

Minisini, V., Gosti, G., & Fanini, B. (2024). Transfer learning for Renaissance illuminated manuscripts: starting a journey from classification to interpretation. In R. Damiano, S. Ferilli, M. Striani, G. Silvello & B. Sassoli de' Bianchi (Eds.), *CEUR Workshop Proceedings: Vol. 3865. Artificial Intelligence for Cultural Heritage 2024* (pp. 30-35). https://ceur-ws.org/Vol-3865/04_paper.pdf

Nikolaidou, K., Seuret, M., Mokayed, H., & Liwicki, M. (2022). A survey of historical document image datasets. *International Journal on Document Analysis and Recognition (IJDAR)*, *25*, 305–338. https://doi.org/10.1007/s10032-022-00405-8

Pächt, O. (1984). *Buchmalerei des Mittelalters: eine Einfuḧrung (D. Thoss & U. Jenni)*. Prestel-Verlag.

Pietroni, E., Chirivì, A., Fanini, B., & Bucciero, A. (2023). An Innovative Approach to Shape Information Architecture Related to Ancient Manuscripts, Through Multi-layered Virtual Ecosystems. From Codex4D to DataSpace. In L. T. De Paolis, P. Arpaia & M. Sacco (Eds.), *Lecture Notes in Computer Science: Vol. 14219. Extended Reality: International Conference, XR Salento 2023* (pp. 247-267). Springer. https://doi.org/10.1007/978-3-031-43404-4_16

Porter, R., Theiler, J., & Hush, D. (2013). Interactive Machine Learning in Data Exploitation. *Computing in Science and Engineering*, *15*(5), 12–20. https://doi.org/10.1109/MCSE.2013.74

Rahal, N., Vögtlin, L., & Ingold, R. (2024). Approximate ground truth generation for semantic labeling of historical documents with minimal human effort. *International Journal on Document Analysis and Recognition (IJDAR)*, *27*, 335–347. https://doi.org/10.1007/s10032-024-00475-w

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, *115*, 211–252. https://doi.org/10.1007/s11263-015-0816-y

Schlecht, J., Carqué, B., & Ommer, B. (2011). Detecting gestures in medieval images. *2011 18th IEEE International Conference on Image Processing*, Brussels, Belgium, 1285–1288. https://doi.org/10.1109/ICIP.2011.6115669

Tkachenko, M., Malyuk, M., Holmanyuk, A., & Liubimov, N. (2020-2025). Label Studio: data labeling software. Retrieved from https://github.com/HumanSignal/label-studio

Ypsilantis, N.-A., Garcia, N., Han, G., Ibrahimi, S., Van Noord, N., & Tolias, G. (2022). The Met Dataset: Instance-level Recognition for Artworks. arXiv preprint arXiv:2202.01747

Zhao, W., Zhou, D., Qiu, X., & Jiang, W. (2021). Compare the performance of the models in art classification. *PLoS ONE*, *16*(3). https://doi.org/10.1371/journal.pone.0248414

Making Germanic Cultural Heritage accessible to students: a proposal for a case study

Chiara De Bastiani[1], Giulia Fabbris[2]
[1] Università Ca'Foscari Venezia, Italy – chiara.debastiani@unive.it
[2] Università Ca'Foscari Venezia, Italy – giulia.fabbris@unive.it

ABSTRACT (ENGLISH)
This poster presents a proposal for a case study involving two User Interfaces (UIs) that present Italian and Germanic Cultural Heritage (CH) to a diversified audience. More specifically, this proposal is designed to be applied with students of Germanic Philology and Linguistics and shows how the two UIs can be employed to investigate and visualize data about Germanic CH in Italy in an accessible way.
Keywords: Linked Open Data; Semantic Web; Aggregators; FAIR data; Germanic Philology.

ABSTRACT (ITALIANO)
Rendere il patrimonio culturale accessibile alle e agli studenti: una proposta per un caso di studio.
Questo poster presenta una proposta per un caso di studio che coinvolge due interfacce utenti che presentano il patrimonio culturale italiano e di origine germanica a un pubblico diversificato. Più nello specifico, questa proposta è pensata per essere messa in pratica con studenti e studentesse di Filologia e linguistica germanica e mostra come le due interfacce possano essere usate per indagare e visualizzare dati sul patrimonio culturale germanico in Italia in modo accessibile.
Parole chiave: Linked Open Data; web semantico; aggregatori; dati FAIR; Filologia Germanica.

1. INTRODUCTION
The aim of this poster is to present a proposal for a case study involving two UIs that collect Germanic and Italian Cultural Heritage resources and make them accessible to a diversified audience. In the present contribution, the sense of the keyword accessibility is understood as spelled out in the FAIR acronym: the resources presented in the two UIs involved in the case study are appropriately identified through hyperlinks; they are described through standard metadata and are open. Moreover, we understand accessibility also as presenting data to a non-expert audience, according to the Faro Convention.[1] Opening data to a non-expert audience is a first step in making data more inclusive and democratic; this step is aided and supported by the digital paradigm, which provides the good practices at the basis of the present proposal for a case study.
In Section 2, we will present the two UIs, and in Section 3 we will present a proposal for a case study consisting in the application of the two UIs in a didactic environment. Section 4 concludes the contribution.

2. THE CHIt AND ONTOVE-SAMPO UIs
Both UIs were developed by reusing existing tools and models and aggregate resources found in different online repositories through the same platform. The projects behind the two UIs presented are respectively CHIt "Cultural Heritage of Italy" and OntoVE "Ontologies for the description of Germanic Cultural Heritage in Veneto within the perspective of the European Open Science Cloud (EOSC)".
Both UIs adhere to the FAIR principles (Wilkinson et al. 2016); in particular, they exploit already existing resources and make CH Findable and Accessible. Moreover, the resources are interoperable and can dialogue with national and international similar initiatives. The two UIs were developed independently of each other, but the CHIt interface moreover presents a subset of OntoVE data through a dedicated thematic path,[2] referred to here as CHIt-OntoVE. Given the fact that both projects reuse open-source tools and models, the workflow adopted in both, and the workflow that led to the creation of the CHIt-OntoVE thematic path, can be potentially applied to other domains of CH. In the following, we describe the two UIs

[1] https://www.coe.int/en/web/culture-and-heritage/faro-convention

[2] For the sake of simplicity, we call the path presenting the OntoVE data the CHIt-OntoVE thematic path. More precisely, the OntoVE data can be accessed both by selecting OntoVE as provider in the CHIt interface, and this is what we refer to as CHIt-OntoVE thematic path, and the OntoVE data will also be accessible through the thematic path aggregating Lombard and Gothic CH.

and the CHIt-OntoVE thematic path in particular, and briefly describe the main goals of the CHIt and OntoVE projects in order to contextualize the present contribution.
The CHIt project, Cultural Heritage of Italy, consists in the creation of an aggregator of CH items produced in Italy. The research behind the CHIt aggregator started from the observation that there does not seem to exist a common methodology in the creation of cultural Heritage aggregators (cf. for example Siqueira and Martins 2021 or Smith 2022). It would be beyond the scope of this contribution to discuss in depth previous research on aggregators, but it suffices to note that data aggregation and integration still offers some challenges in terms of reusability or interoperability of data, as different projects usually create their own workflow and models to aggregate and describe their data. One of the outcomes of the CHIt project is therefore the development of a methodology that can be replicated and applied to different domains of interest, especially because it reuses open-source tools and models.
CHIt reuses the Muruca framework (Aiola and Lombardo 2022), developed with a three-layer architecture comprising a backend, a middleware and a frontend. For the CHIt-UI, the Muruca source code was adapted to the specific goals of the projects. The digital objects retrieved by CHIt are already available on the web, but scattered on different websites. Muruca allows for their recovery and reorganization within a platform that makes them homogeneous and interoperable. The metadata of the resources are either stored in the backend by manual addition (for stand-alone projects and GitHub) or retrieved via API (for Europeana and Zenodo) or via a SPARQL query (for OntoVE) directly from the middleware. From the backend, the items are sent to the middleware, where the business logic runs. Once the metadata are all in this central layer, they are mapped according to an output schema that makes them interoperable. This schema adheres to the DCMI. The mapping of the metadata into a uniform format proved useful, among the other things, for the creation of thematic paths. In fact, these are created by aggregating resources according to e.g. the provenance of the data. Figure 1 shows how resources are aggregated from different providers, among them the OntoVE SPARQL Endpoint,[3] mapped in the middleware and sent to visualization.

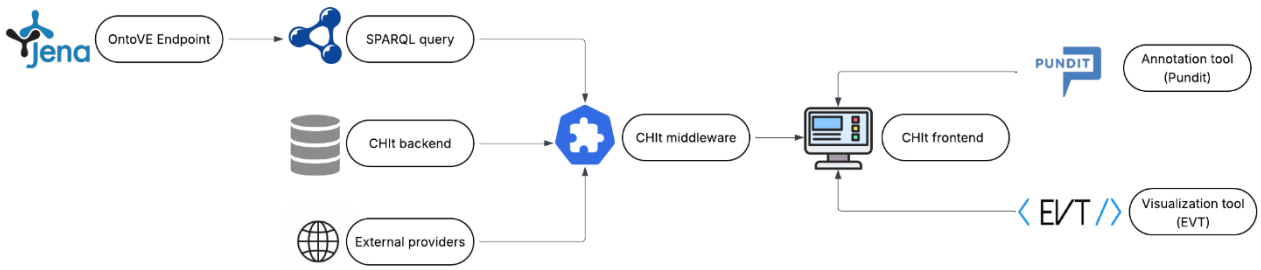


Figure 1. Aggregation of resources, mapping and visualization in CHIt

A further outcome of the project consists in the presentation of textual data to users, thanks to the addition of EVT3 (Cacioli et al. 2022) and the possibility for users to interact with the data. The CHIt interface is in fact enriched with the tool Pundit,[4] a browser extension that allows to annotate and comment data, creating notebooks. This functionality is particularly relevant for the case study proposed, since it can be exploited by students to explore and annotate the data according to specific questions they need to answer and to cooperate, since notebooks can be shared among users.[5]
The CHIt-OntoVE thematic path collects a subset of data retrieved from the OntoVE Knowledge Base, namely archaeological findings produced by Germanic populations and related to the Veneto region (Figure 2); the data are displayed in the same way as the other resources aggregated through CHIt and can be annotated with Pundit. This thematic path enriches both projects, since it provides more data for the CHIt interface, and makes the OntoVE data comparable to the data collected in CHIt that have a similar origin but are found outside of Veneto. A further thematic path relevant for the present proposal collects Italian CH items of Lombard and Gothic origin.

[3] The OntoVE SPARQL Endpoint is built with Apache-Jena-Fuseki (https://jena.apache.org/documentation/fuseki2/, last accessed April, 7th 2025) and can be found here: http://ontove-endpoint.chitontove.it/, last accessed April, 7th 2025. The Endpoint is only accessible with a username and a password, for security reasons.
[4] https://thepund.it/, last accessed April 7th, 2025.
[5] The CHIt interface can be explored through this link: https://chit.chitontove.it/risorse; the OntoVE-Sampo-UI can be retrieved here: https://ontove-ui.chitontove.it/it/. Last accessed April 7th, 2025.

The OntoVE UI reuses the Sampo Model (Ikkala et al. 2022, Rantala et al. 2023) and presents the OntoVE dataset through specific "User Perspectives" that have at their center different kinds of CH items and their properties, and make these properties searchable through dedicated filters. The main goal of the OntoVE project is the creation of a Knowledge Base that collects CH items produced by Germanic populations and related to Veneto; the dataset is therefore composed of historical artefacts of various nature (archaeological findings, sites such as castles, written documents). The items of the dataset are already present in different online repositories. The OntoVE project reuses different domain ontologies in order to devise an ontological description of the items comprising the dataset in order to answer different competency questions seeking to highlight the relationship between the Veneto region and the Germanic populations that transited through it. It is not the goal of this contribution to review the choices behind the models reused, but the reader can consult the OntoVE specification page for more information.[6]

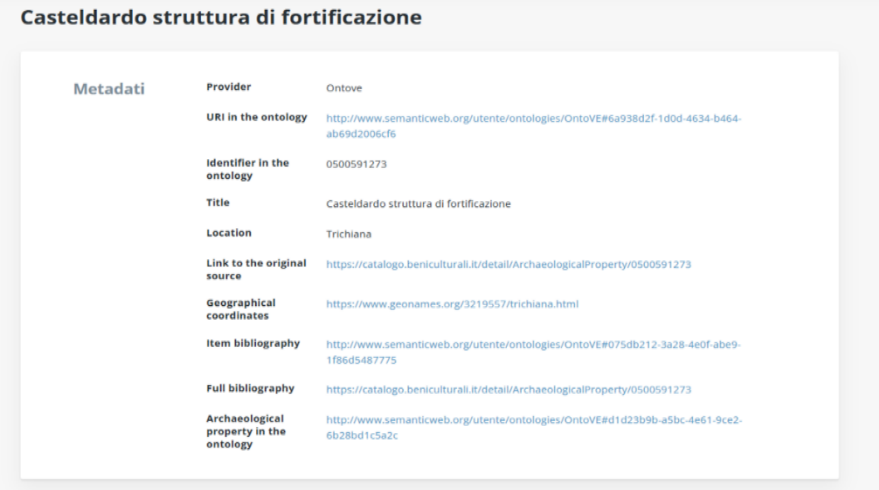


Figure 2. Visualization of one OntoVE resource through CHIt

This contribution focuses on the visualization of the OntoVE data. Whereas SPARQL is undoubtedly a very powerful way to query a Knowledge Base, the project also aims to offer different visualization outputs for users who do not have the relevant expertise.
To this end, two further visualization outputs are offered, namely the OntoVE-Sampo-UI and the thematic path provided by CHIt. The Sampo Model has been developed by the Semantic Computing Research Group (SeCo) at the Helsinki university; it is a reusable model, which has been tested on several Semantic portals built by the SeCo group. It allows to create portals on top of different RDF triplestores and is domain-agnostic. It provides different inspection tools, such as table-view, pie-charts or maps and allows users to select specific filters to further explore the data. The data are presented through so-called "search perspectives", i.e. aggregations of specific ontological classes and selected properties; at the basis of each search perspective lies a SPARQL query and a .json file with configurations for the filters to be visualized. The Sampo UI source code was reused in OntoVE to adapt it to the OntoVE dataset and seven search perspectives have been elaborated, each focusing on specific classes in the ontological modelling and related properties; at the moment, the OntoVE-Sampo-UI offers only table-view inspection and filtering function for the data, but we aim to implement it with interactive maps. For example, users can consult the Places perspective (Figure 3) and filter the texts produced by rulers of Germanic origin related to specific places in Veneto (or a historical administrative entity including Veneto). Instructions are provided on how to use the UI in a dedicated section and the portal is in English and Italian language.

[6] The OntoVE specification can be consulted here: https://w3id.org/ontove/ontology, last accessed April 7th, 2025.

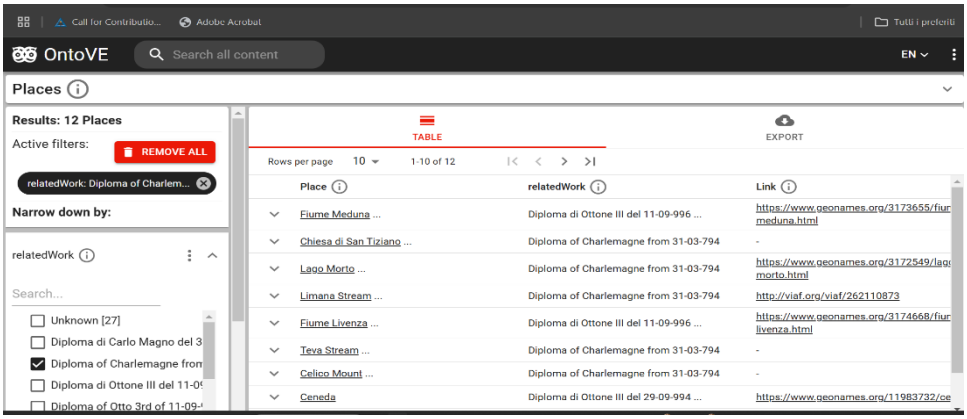


Figure 3. The "Places" perspective in the OntoVE-Sampo-UI

As illustrated in this brief overview, the two UIs handle different datasets and offer different ways of interacting with the data. The CHIt interface presents Italian CH items retrieved from various providers and allow users to create (collaborative) notebooks. The resources can be filtered according to the provenance of the data, or the Cultural Scope they belong to and their metadata, including links to the original providers, are displayed following the DCMI module. CHIt moreover allows to visualize XML data thanks to EVT3.

The OntoVE UI is designed on top of an RDF triplestore; the data that can be visualized are only provided by the OntoVE Knowledge Base, and the resources are displayed through different perspectives that are inspired by the modelling of the data. In other words, heterogeneous data are first described within the OntoVE project in RDF format and the resulting graph is stored in the Apache-Jena-Fuseki SPARQL Endpoint. The OntoVE-UI then retrieves the data from the Endpoint and presents it through the designed search perspectives. The perspectives cater to audiences with different (research) questions. Users can select a perspective and filter the data, similarly to the CHIt frontend, and links to the original source for the metadata are provided, if users wish to know more about an object of specific interest. The filters and the data displayed all reflect the classes and properties of the OntoVE ontology. The OntoVE-Sampo-UI does not however allow to annotate the data and does not include text visualization. The two UIs are therefore different in terms of their design and in the way their data are aggregated and structured, but they both adhere to the FAIR principles and offer interoperable and reusable data.

3. CASE STUDY PROPOSAL

The two UIs presented above can be employed in order to make Germanic Cultural Heritage accessible to students of Germanic philology and linguistics at the BA level. Course programs and textbooks about Germanic philology usually discuss the Germanic populations of the Lombards and Goths and the impact they had in shaping Italian history and language (cf. Francovich-Onesti 2002, Saibene & Buzzoni 2006, Battaglia 2021, Zironi 2022). We propose an activity that would let students explore the relevant sources and the cultural heritage left by the Lombards in Italy through the use of CHIt, Pundit and OntoVE-Sampo-UI. As a first step, students are asked to find and annotate on a relevant collaborative notebook the documentary sources produced by the Lombards. This can be done by following the CHIt path about the Lombards and the Goths. By way of demonstration, Figure 4 shows the relevant findings for Lombard CH in Italy, while Figure 5 shows how students could annotate the information about the codices preserving the Edict of Rothari in a notebook. Students can add keywords previously individuated with the lecturer to describe their data appropriately.

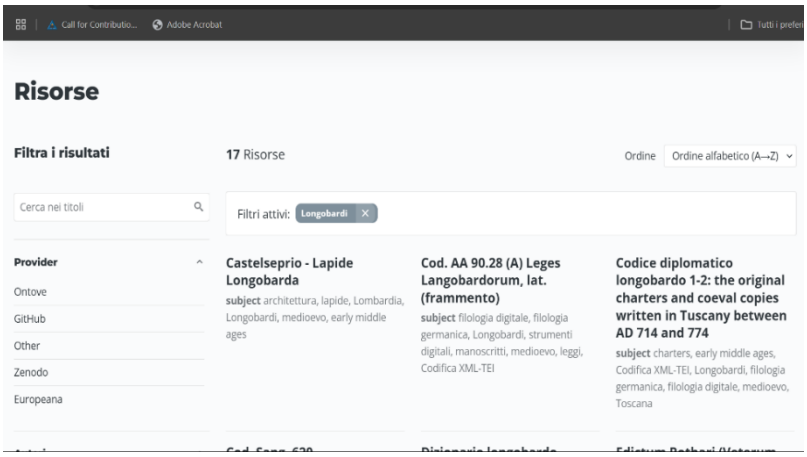


Figure 4. Lombard findings in the Lombard-Gothic thematic path in CHIt.

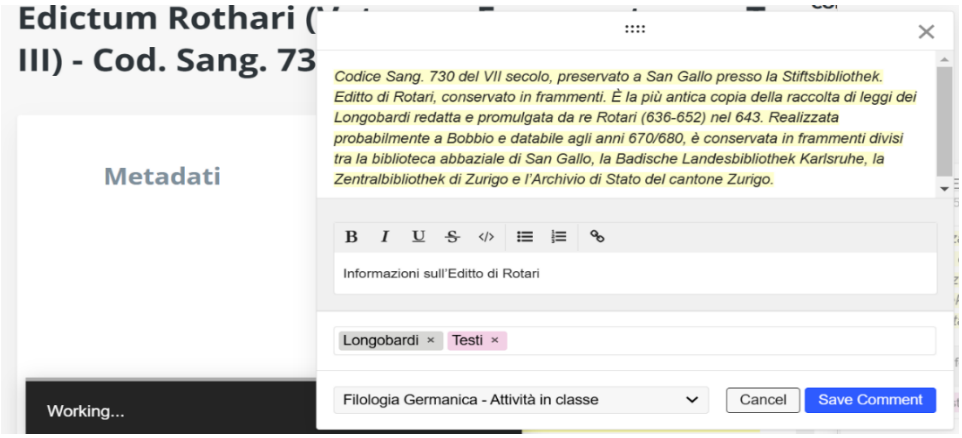


Figure 5. Example of annotation on the entry for the Edict of Rothari.

As we mean to test this case study on the courses offered at the Ca' Foscari University of Venice, the students will then be asked to locate the relevant historical artefacts left by the Lombards in Veneto and mark the areas where these are found. Then, by reading the *Historia Langobardorum* by Paolo Diacono (Zanella 1991), they will have to annotate on their notebooks whether the areas where the artefacts are located correspond to the areas of Veneto mentioned in the *Historia*. The OntoVE data retrieved through CHIt (cf. Figure 2) can be annotated as illustrated for the Edict of Rothari in Figure 5. Students will then be asked to locate similar historical artefacts in areas outside of Veneto and annotate whether they present similarities or differences. Finally, by using the OntoVE-Sampo-UI students will consult the "Events" perspective and they will be asked to first individuate the relevant acts issued by Charlemagne after conquering the Lombard reign (Figure 6), and then filter the locations in Veneto mentioned in the documents through the Subject filter provided in the "Texts" perspective (Figure 7), comparing them again to the places mentioned in the *Historia*.

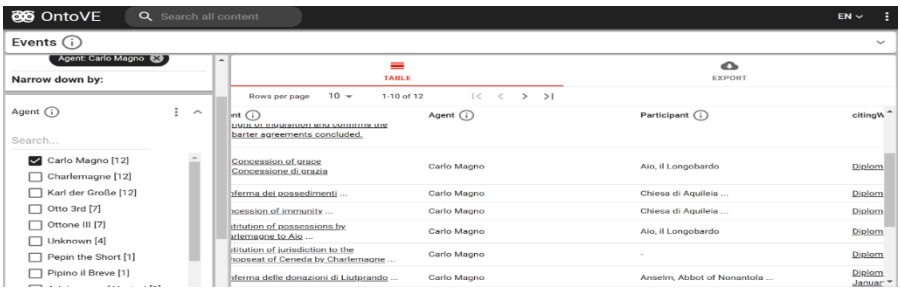


Figure 6. Filtering of events performed by Charlemagne in the "Events" perspective, OntoVE-Sampo-UI

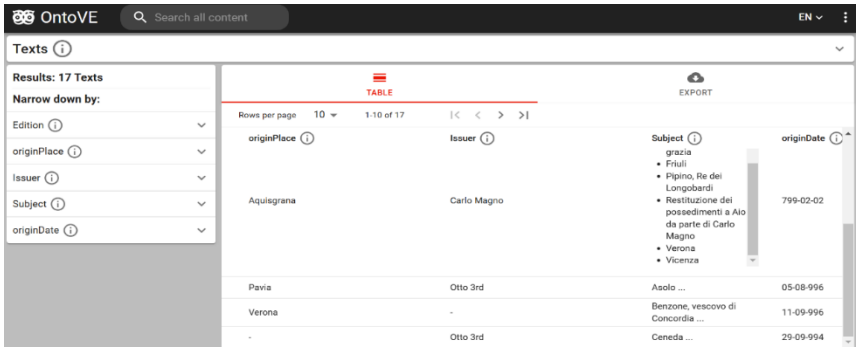


Figure 7. Visualization of the subjects contained in a selected single charter.

4. CONCLUSIONS AND FUTURE PERSPECTIVES

With this proposal, we have showed the potential of the CHIt and OntoVE-Sampo-UI for the inspection of Germanic CH by students. We argue that the two UIs presented let users visualize the information about CH items of their choice in an accessible way; since the original links to the resources retrieved through CHIt and described in OntoVE are provided, user can further broaden their research and visualize the information as provided in the original repository. The actual testing of the UIs and of the proposal with the students are left for future research. Despite the limited scope of the present proposal, our contribution shows how the OntoVE and CHIt resources can be easily retrieved, inspected and annotated in a collaborative way; especially through the CHIt interface and the CHIt-OntoVE thematic path, new activities for students can be devised. Moreover, we aim to add also the bibliographic data collected in OntoVE to the CHIt interface, to allow users to also annotate these data, and provide the OntoVE-Sampo-UI with maps, in order to offer a more dynamic visualization of the data. As far as the datasets are concerned, even though the OntoVE dataset is restricted to Germanic CH in Veneto, the OntoVE-Sampo-UI and the CHIt-OntoVE thematic path can be of interest to citizens and tourists willing to explore the history of the area. Finally, from the wider perspective of the methodology employed in both projects, we argue that the approaches to data aggregation, description and visualization can be applied to different CH domains, thus providing a shared, interoperable workflow and allowing to widen the pool of users interested in the UIs provided.

ACKNOWLEDGEMENTS

The OntoVE and CHIt projects are funded by the Italian National Operational Programme on Research and Innovation 2014-2020[7] and the European Social Fund[8] within the Recovery assistance for cohesion and the territories of Europe[9].

[7] PON, http://www.ponricerca.gov.it/pon-ricerca/programme/, last accessed April 7th, 2025.

[8] ESF, https://eur-lex.europa.eu/EN/legal-content/glossary/european-social-fund-esf.html, last accessed April 7th, 2025.

[9] REACT-EU, https://www.europarl.europa.eu/factsheets/en/sheet/215/recovery-assistance-for-cohesion-and-the-territories-of-europe-react-eu-, last accessed April 7th, 2025.

REFERENCES

Aiola, C., & Lombardo, T. (2022). Storytelling through Digital Scholarly Editions. DARIAH Annual Event 2022: Storytelling. Athens. Available at: https://zenodo.org/records/6594346

Battaglia, M. (2021). I Germani. Genesi di una cultura europea. Roma: Carocci Editore.

Cacioli, G., Cerretini, G., Di Pietro, C., Maenza, S., Rosselli Del Turco, R., & Zenzaro, S. (2022). There and Back Again: What to Expect in the Next EVT Version. AIUCD 2022 - Digital Cultures. Intersections: Philosophy, Arts, Media. Proceedings of the 11th National Conference. Lecce, pp. 212–217.

Council of Europe. (2005). Convention of the Value of Cultural Heritage for Society. Available at: https://www.coe.int/en/web/culture-and-heritage/faro-convention

Francovich-Onesti, N. (2002). Filologia Germanica. Lingue e culture dei germani antichi. Roma: Carocci Editore.

Ikkala, E., Hyvonen, E., Rantala, H., & Koho, M. (2022). Sampo-UI: A Full Stack JavaScript Framework for Developing Semantic Portal User Interfaces. Semantic Web – Interoperability, Usability, Applicability, vol. 13, no. 1, pp. 69-84, January, 2022. Online version published in 2021, print version in 2022.

Rantala, H., Ahola, A., Ikkala, E., & Hyvönen, E. (2023). How to create easily a data analytic semantic portal on top of a SPARQL endpoint: introducing the configurable Sampo-UI framework. VOILA! 2023 Visualization and Interaction for Ontologies, Linked Data and Knowledge Graphs 2023. CEUR Workshop Proceedings, Vol. 3508, October, 2023. https://ceur-ws.org/Vol-3508/paper3.pdf

Saibene, M.G., & Buzzoni, M. (2006). Manuale di Linguistica Germanica. Milano: Cisalpino.

Siqueira, J., & Martins, D. L. (2021). Workflow models for aggregating cultural heritage data on the web: A systematic literature review. Journal of the Association for Information Science and Technology, 73(2), 204–224. Available at: https://asistdl.onlinelibrary.wiley.com/doi/10.1002/asi.24498

Smith, M. (2022). Linked open data and aggregation infrastructure in the cultural heritage sector. A case study of SOCH, a linked data aggregator for Swedish open cultural heritage. In Koraljka, G. & Ying-Hsang, L. (Eds.), Information and Knowledge Organisation in Digital Humanities. Global Perspectives. Routledge: Oxon and New York.

Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016). Available at: https://doi.org/10.1038/sdata.2016.18

Zanella, P. (1991). Storia dei Longobardi di Paolo Diacono, Testo latino a fronte. Milano: Rizzoli.

Zironi, A. (2022). Filologia Germanica. Lingua, storia, cultura, testi. Firenze: Le Monnier Università.

/DH.arc Vocabularies: Making semantic artefacts more visible and accessible using SKOS

Laurent Fintoni[1]
[1] Dipartimento di Filologia Classica e Italianistica, Università di Bologna, Italia – laurent.fintoni2@unibo.it

ABSTRACT (ENGLISH)
This paper presents /DH.arc Vocabularies, a repository for semantic artefacts built using Skosmos and intended to make non-conventional research products such as ontologies and controlled vocabularies created by the Digital Humanities Advanced Research Centre at the University of Bologna more visible and accessible. The paper defines semantic artefacts, their importance, and the need to increase their FAIR-ness before reviewing related work and available technology and giving an overview of the repository, including why Skosmos was chosen, how it was implemented, and the publishing process which includes transforming ontologies written in OWL into SKOS in order to enable their browsing via the interface, a novel approach to existing uses of Skosmos in the fields of cultural heritage and Digital Humanities.
Keywords: Knowledge Organization Systems; Skosmos; Semantic Web; Ontologies; RDF

ABSTRACT (ITALIANO)
/DH.arc Vocabularies: rendere gli artefatti semantici più visibili e accessibili utilizzando SKOS. Questo articolo presenta /DH.arc Vocabularies, un repository per artefatti semantici creato utilizzando Skosmos e destinato a rendere più visibili e accessibili prodotti di ricerca non convenzionali come ontologie e vocabolari controllati creati da Digital Humanities Advanced Research Centre, Università di Bologna. L'articolo definisce gli artefatti semantici, la loro importanza e la necessità di aumentare la loro FAIR-ness prima di esaminare il lavoro correlato e la tecnologia disponibile e fornire una panoramica del repository, incluso il motivo per cui è stato scelto Skosmos, come è stato implementato e il processo di pubblicazione che include la trasformazione di ontologie scritte in OWL in SKOS per consentire la loro navigazione tramite l'interfaccia, un nuovo approccio agli usi esistenti di Skosmos nei campi del patrimonio culturale e delle discipline umanistiche digitali.
Parole chiave: Sistemi di organizzazione della conoscenza; Skosmos; web semantico; ontologie; RDF

1. INTRODUCTION
Over the past two decades the importance of Knowledge Organization Systems (KOS) within the humanities has grown, starting with the interoperability promised and enabled by Semantic Web standards in the 2000s and continuing with the emergence of FAIR Data Principles (Wilkinson et al., 2016) and the prevalence of data-intensive research in the 2010s, which requires data to be defined in a way that is understandable by both computers and humans. At the same time, these technological and research advancements have pushed against the limits of the term KOS. Defined by Hodge as any type of scheme "for organizing information and promoting knowledge management," (Hodge, 2000) KOS have come to cover a variety of digital objects, with the DCMI/NKOS Task Group listing 14 types[1] such as gazetteer, taxonomy, thesaurus, ontology, and controlled vocabulary, all of which enable interoperability at the semantic level by "providing a formal conceptualization of the data that can be shared, reused, and aligned" (d'Aquin & Noy, 2012). More recently, the term semantic artefact has emerged as an alternative (Jonquet et al., 2023), broadening the scope to a "machine-actionable formalization (represented using appropriate formats and serializations, including RDF and non-RDF standards) of a conceptualization, enabling sharing and reuse by humans and machines" (Corcho et al., 2024). This offers a more up-to-date definition that expands the types of objects covered, adding for example standards and metadata schemas to the list, and aligns with both private and public sector practices as well as the goals of international research efforts such as the European Union's Horizon 2020 program (Le Franc et al., 2020).
Regardless of how they are labelled, these research products hold a special importance in that they are both key to achieving FAIR-ness in the data being produced and are themselves research objects that should be FAIR (Cox et al., 2021). However, the latter is often less considered than the former (Poveda-Villalón et al., 2020). This is apparent in the Italian context, which interests us and where semantic artefacts hold a place of importance within humanities, and especially Digital Humanities (DH), research. Despite this importance they can often be difficult to share, find, and (re)use. These issues arise in part

[1] https://nkos.dublincore.org/nkos-type.html

due to the specificities of professional recognition and evaluation within the Italian academic system which prioritizes the processes (theoretical and practical) through which the object was produced rather than the object itself (Fintoni, 2024; Tammaro, 2014). A direct example of this is the lack of representation for semantic artefacts at the metadata level within IRIS institutional repositories where Italian academic research products are expected to be stored and where they are often assimilated into a generic dataset typology. A subtler example is the prioritization of research products that are easier to understand within the Italian academic system, such as digital editions, over semantic artefacts like ontologies.
For the past two decades, efforts to make semantic artefacts more visible and reusable have largely focused on the use of organizational systems that can facilitate their spread and which have been identified as an important component of Semantic Web infrastructure (Baclawski & Schneider, 2009). Due to the increasing number of these semantic artefacts and their usage, within the humanities and other fields such as the life sciences, there is an argument that such organizational systems should be mandatory (Jonquet et al., 2023).
In this paper, we present /DH.arc Vocabularies, a new repository developed to improve the visibility and FAIR-ness of the semantic artefacts produced by the Digital Humanities Advanced Research Centre (/DH.arc) at the University of Bologna and built using the open-source Simple Knowledge Organization System (SKOS) browser Skosmos[2]. Following a review of related work on the subject and available technology, we detail why Skosmos was chosen, how it was implemented, and the publishing process which includes transforming ontologies written in the Web Ontology Language (OWL) into SKOS in order to enable their browsing via the Skosmos interface, which, to our knowledge, offers a novel approach to existing uses of Skosmos in the fields of cultural heritage and DH.

2. RELATED WORK

Existing research into the distribution and accessibility of semantic artefacts has focused primarily on the aforementioned organizational systems, starting with a definition of ontology libraries in 2001 which already underlined the importance of re-usability (Ding & Fensel, 2001). By the end of the 2000s, the concept of ontology repository was introduced to expand these libraries beyond the mere description and listing of items and include other features, by then more common in web-based systems, such as searching, browsing, and querying of content via Application Programming Interfaces (API) (Hartmann et al., 2009). Two reviews of ontology libraries by d'Aquin & Noy (2011) and Naskar & Dutta (2016) built upon this previous work highlighting different categories of libraries such as directories, registries, application platforms, and repositories, with d'Aquin & Noy noting that as libraries all provide means for discovery and access to ontologies "they all serve as ontology repositories" (d'Aquin & Noy, 2011). More recently, Corcho et al. called for the need to adopt an inclusive definition of where semantic artefacts are stored, settling on the term "catalogue of semantic artefacts" for their maturity model, which they define as "a dedicated web-based system that fosters the availability, discoverability and long-term preservation and maintenance of semantic artefacts" (Corcho et al., 2024). We have chosen to refer to /DH.arc Vocabularies as a repository, in line with the features first outlined by d'Aquin & Noy, though semantic artefact catalogue would also be applicable.
As noted by Jansen, the biomedical and wider life sciences domains of research have done much to contribute to ontology development and this is reflected in the common usage of repositories for semantic artefacts in those domains such as OBO Foundry[3] and the NCBO BioPortal[4] (Jansen, 2009). Outside of the life sciences, notable efforts to make semantic artefacts accessible and usable come primarily from public administration and memory institutions such the ARDC Research Vocabularies Australia[5], the National Library of Finland's finto.fi[6], the RDA Registry[7], and the EU's EU-Vocabularies[8]. Looking more specifically at the DH, the Austrian Centre for Digital Humanities and Cultural Heritage's (ACDH-CH) Vocabs[9] is a

[2] https://skosmos.org/

[3] https://obofoundry.org/

[4] https://bioportal.bioontology.org/

[5] https://vocabs.ardc.edu.au/

[6] https://finto.fi/

[7] https://www.rdaregistry.info/

[8] https://op.europa.eu/en/web/eu-vocabularies/

[9] https://vocabs.dariah.eu/en/

repository that offers a single access point for vocabularies related to the Digital Research Infrastructure for the Arts and Humanities (DARIAH), including the latest version of the TaDiRAH taxonomy, popular in the DH. Linked Open Vocabularies[10] is another popular source of semantic artefacts for DH projects. Of the 26 catalogues that Corcho et al. assessed in their research, among the most recent available, nine can be considered to cover semantic artefacts related in part to the humanities, with a particular focus on cultural heritage.

3. SEMANTIC ARTEFACT REPOSITORY TECHNOLOGY

In terms of the technology that powers these repositories, a useful reference is the Software for controlled vocabularies wiki page[11] maintained by the Basic Register of Thesauri, Ontologies & Classifications (BARTOC) (Voß, 2016). Currently, the wiki lists upwards of 35 software and despite the title many of these can be used to edit, view, analyze, and make available most types of semantic artefacts with a particular focus on ontologies and controlled vocabularies. Among these, notable and relevant to our work are solutions that include both services for and services that use semantic artefacts, a key shift from the functionality of early ontology libraries: OntoPortal[12] (developed in Ruby and based on the NCBO BioPortal technology) and OLS[13] (developed in JavaScript and Java and deployed by the EBI Ontology Lookup Service), both used primarily in the life sciences; Skosmos (based on the SKOS data model and developed in PHP[14]), used by finto.fi and ACDH-CH's Vocabs; VocBench and ShowVoc[15] (developed in JavaScript and Java and managed by the Publications Office of the EU), used for the EU-Vocabularies service; TemaTres[16] (developed in PHP), used primarily in the Spanish-speaking world; and Semantic MediaWiki[17] (an extension of the software that powers Wikipedia developed in PHP) used by the semantic web portal Ontology Design Patterns[18].

4. /DH.ARC VOCABULARIES OVERVIEW

The original driver for the creation of the repository was the need to make the controlled vocabularies created by the KNOT project[19] available, thus completing the checklist from Cox et al. to ensure their FAIR-ness (Fintoni, 2024). This led us to consider Skosmos for a number of reasons: (i) it is intended for controlled vocabularies and based on SKOS, which was used to create the KNOT vocabularies, and enables the underlying RDF serializations to be served using HTTP content negotiation when URIs are accessed, thus simplifying the Linked Data publishing infrastructure needed; (ii) it is an open-source solution, which makes potential customizations easier at the cost of requiring more technical skills, and it has a clear separation between code and data with a simple configuration and importing system given the expected technical knowledge of DH practitioners; (iii) it offers multilingual user interfaces and a clean and intuitive presentation layer; (iv) it is already in use within a DH context for ACDH-CH's Vocabs, giving us a use case whose code could be referenced; (v) it offers a REST API for read-only access; (vi) and it is still being developed and supported by the team at the National Library of Finland, for whom the finto.fi service, built on Skosmos, serves as a central hub for shared authority files of all national memory institutions in the country (Suominen et al., 2015). The primary drawbacks of using Skosmos were the lack of an integrated editor, the lack of support for using different triple stores than the one recommended by the developers, and its inability to handle non-SKOS formatted files. We were eventually able to find some solutions to these last two points, which are detailed below.

[10] https://lov.linkeddata.es/dataset/lov/

[11] https://github.com/gbv/bartoc.org/wiki/Software-for-controlled-vocabularies

[12] https://ontoportal.org/

[13] https://github.com/EBISPOT/ols4

[14] While Skosmos is referred to as a browser and publishing tool for vocabularies in its documentation, the project from which it emerged, ONKI, was an ontology repository used for the Finnish Linked Open Ontology Cloud KOKO and several international vocabularies (Suominen et al., 2015).

[15] https://vocbench.uniroma2.it/

[16] https://vocabularyserver.com/web/

[17] https://www.semantic-mediawiki.org/wiki/Semantic_MediaWiki

[18] http://ontologydesignpatterns.org/wiki/Main_Page

[19] https://icdp-digital-library.github.io/KNOT/

Skosmos's architecture is relatively simple with a SPARQL endpoint acting as the backend data store and the front-end interface rendered using the Twig templating engine for PHP with additional functionality, such as browsing of vocabulary contents, provided by client-side JavaScript code. Due to the limitations of SPARQL 1.1. for text-based queries, the developers recommend using Apache Jena Fuseki with the jena-text index as the endpoint to handle larger vocabulary needs alongside Varnish as an HTTP proxy cache to improve performance. This approach allows for fast updates to the underlying data and fast page generation, alongside ensuring data is always up to date (Suominen et al., 2015). While it is possible to run Skosmos with a different triple store as backend, which would have been our preferred method due to previous issues with Apache Jena Fuseki within the existing /DH.arc infrastructure, there is no supporting documentation to do so requiring a larger technical investment that was not possible. As such we opted to implement Skosmos using Docker, which mimics the baseline installation[20]. Currently, /DH.arc Vocabularies uses Skosmos 2.18.1[21] which is the last stable release before the current 3.0 version that is still in development. Our production workflow involves a GitHub repository hosting a copy of 2.18.1, which can be run on personal computers using the desktop Docker app to act as a development environment for testing and implementing new functionalities and styles or semantic artefacts[22]. The repository is in turn connected to the /DH.arc web infrastructure where it is used as the production version of the application. Both the development and production versions use Docker Compose, as the dockerized version of Skosmos involves three containers: one for Fuseki, one for Varnish, and one for the application itself. Docker Compose commands are used to push changes from development to live. Though the developers do not recommend using Docker for production we have so far found it to be appropriate for our needs, as our semantic artefacts are relatively small in terms of the number of triples contained and it allows us to bypass internal technical difficulties with Apache Jena Fuseki while having a manageable production workflow that can be easily documented and handed over if need be.

All semantic artefacts made available via Skosmos need to be formatted using SKOS, though there is also support for non-SKOS properties and classes defined using RDF (discussed in more detail in Section 5). The publishing process we follow is similar to that suggested by the developers (Suominen et al., 2015), starting with a pre-process of the files using the Skosify Python library[23] to clean and validate them. This is then followed by a manual check and the addition of any necessary triples or information (for example Italian and English text strings, as these are the two languages used by the repository). Files are then imported into the triple store using cURL commands as per the existing Docker documentation. The last step is the addition of the semantic artefact to the Skosmos configuration file (which uses the Turtle syntax), where it is represented as a dataset and additional options can be set such as supported languages, display options, and download links. A last round of checking is then conducted via the front-end to ensure no information is missing or that there are no display issues. Wherever possible, source files for the semantic artefacts in /DH.arc Vocabularies are hosted in the same GitHub repository as the application. Any update to the files requires the cURL commands to simply be rerun for changes to be reflected in the front-end. Changes to the config file (needed if the URI of the source file changes) or any development changes require the Docker container to be relaunched via Compose commands.

On the front-end, all instances of *skos:Concept* are given their own page, with SKOS mapping relationships shown in a separate section, while instances of *skos:ConceptScheme* act as the de facto 'homepage' of the semantic artefact providing information about the artefact itself. In addition to the standard information one would expect to find at the concept scheme level (agents, creation dates, title, description, publisher etc.) we have chosen to add two additional properties for now: *foaf:homepage* to indicate the homepage of the project which created the artefact, and *dcat:downloadURL* to indicate where a non-SKOS version of the file may be found if it exists. The latter is particularly relevant when dealing with ontologies originally created in OWL as we will see in Section 5. Following early tests of this Skosmos implementation using controlled vocabularies from the KNOT and National Edition of Aldo Moro's Works projects we decided to try and widen its usage to include all semantic artefacts created by /DH.arc, including ontologies. Currently, /DH.arc Vocabularies is intended to host only semantic artefacts created by /DH.arc members and associated research projects and which are already formatted in SKOS or can be.

[20] https://github.com/NatLibFi/Skosmos/tree/main/dockerfiles

[21] https://github.com/NatLibFi/Skosmos/releases/tag/v2.18.1

[22] https://github.com/laurentfintoni/dharc-skosmos

[23] https://github.com/NatLibFi/Skosify

5. ADDING OWL ONTOLOGIES TO SKOSMOS

In the Semantic Web, OWL is intended to be used for the representation of formal ontologies while the SKOS data model was designed to port KOS such as thesauri or taxonomies without formal semantics (Belk et al., 2015). While the W3C indicates that SKOS may be used with OWL[24], as the data model itself is an OWL ontology, existing research into the usage of both languages together is rather limited with early work on moving from KOS to OWL (Hepp & de Bruijn, 2007)[25] and, more relevant to our interests, some explorations of the combination of both published shortly before and after the release of SKOS: usage patterns to help identify modeling errors and infer additional information (Belk et al., 2015), and use cases for combining both (Jupp et al., 2008).

While the majority of existing Skosmos implementations are dedicated to making only controlled vocabularies and similar KOS available[26], finto.fi does make explicit reference to ontologies in its front-end and documentation. As such, we investigated how these SKOS representations of ontologies differed from controlled vocabularies by looking at the General Finnish Ontology YSO[27], one of the central ontologies in finto.fi. This ontology consists mainly of concepts represented as instances of *owl:Class* that are subclasses of *skos:Concept*, alongside a handful of OWL object and data type properties used to indicate labelling types and deprecated concepts. This approach is well suited to SKOS as the concepts in the YSO are traditional KOS elements that don't require the inclusion of more formal semantics as might be found in an OWL formatted ontology where, for example, object properties could be used to describe binary relationships between individuals or a class might make use of restrictions. However, the ontologies created by /DH.arc are all written in OWL and it is the formal semantics the language makes possible and the ontological commitment in a domain these ontologies express (Bruseker et al., 2017) that should be made evident to the end user for their publication and browsing to be useful (beyond making them FAIR).

To solve this issue, we took inspiration from a 2008 W3C document[28] (which predates the 2009 SKOS recommendation and is connected to Jupp et al.) that indicates four patterns for working with SKOS and OWL including going from more (OWL) to less (SKOS) formal by either overlaying the two languages or transforming one into the other. We tested both approaches and settled on overlaying, whereby triples are merged, as this allowed us to make our ontologies viewable in the Skosmos browser (which requires SKOS classes and properties in order to display the underlying RDF triples) while ensuring that the formal semantics they offer are still understandable to both humans and machines. As noted by the aforementioned W3C document, this approach implies two key points: the resulting ontologies are necessarily OWL Full and *skos:Concept* and *owl:Class* cannot be disjoint. In effect, this approach lets us provide taxonomy- or thesaurus-like representations of OWL ontologies, which Jupp et al. noted as a sensible option for specific applications like navigation systems (Jupp et al., 2008).

Our overlaying approach for OWL ontologies is as follows: (i) the ontology is given the *skos:ConceptScheme* class, in addition to *owl:Ontology*, as well as the *skos:hasTopConcept* property (which points to all classes in the ontology as these are now also instances of concepts as explained next), and where needed or appropriate generic information properties are added such as *dc:license*, *foaf:homepage,* and *dcat:downloadURL* as noted in Section 4; (ii) each class and property in the original ontology is given the additional *skos:Concept* class and the *skos:prefLabel* (a duplicate of *rdfs:label*) and *skos:inScheme* properties to correctly reflect the structure of the ontology as a concept scheme as well as correctly enable front-end browsing; (iii) definitions of OWL specific properties, such as *rdfs:domain* and *rdfs:range,* are also added in order for them to be displayed in the front-end; (iv) lastly the *rdfs:subClassOf* property is kept (and defined so as to be displayed) and used in place of SKOS semantic relation properties (*skos:broader* and *skos:narrower*) so as to not confuse any potential subclass structure within the ontology with a more rigid, hierarchical KOS structure.

Figure 1 offers a simplified example of this approach in practice, using the Turtle syntax to show how the F Entry ontology and one of its classes were converted during the development phase of /DH.arc

[24] https://www.w3.org/TR/skos-reference/

[25] https://www.heppnetz.de/projects/skos2owl/

[26] This includes the ACDCH's Vocabs as well as the CNRS' Lottere, the UNESCO thesaurus, and the ELSST's social sciences thesaurus, links to which can all be found at https://skosmos.org/

[27] https://finto.fi/yso/en/

[28] https://www.w3.org/2006/07/SWD/SKOS/skos-and-owl/master.html

Vocabularies[29]. Please note the description property has been removed to keep the example concise.

fentry: a owl:Ontology, skos:ConceptScheme ;
dc:title "F Entry (Scheda F) Ontology"@en, "Ontologia Scheda F"@it;
skos:prefLabel "F Entry (Scheda F) Ontology"@en, "Ontologia Scheda F"@it;
dc:creator "Marilena Daquino", "Silvio Peroni" ;
owl:imports skos:, fabio:, hico:, cito:, pro:, scoro: ;
dct:language "English"@en, "Italiano"@it ;
dct:created "2016" ;
dct:issued "2016-12-15"^^xsd:date ;
owl:versionInfo "1.0" ;
skos:hasTopConcept fentry:photograph, fentry:FEntry, fentry:shot ;
foaf:homepage <https://essepuntato.it/fentry/current/fentry.html> ;
dcat:downloadURL <https://svn.code.sf.net/p/dwellonit/code/FEO/feo.owl> .

fentry:FEntry a owl:Class, skos:Concept ;
skos:inScheme fentry: ;
skos:topConceptOf fentry: ;
rdfs:label "F Entry"@en, "Scheda F"@it ;
skos:prefLabel "F Entry"@en, "Scheda F"@it ;
rdfs:comment "It is a document containing metadata about a photograph and about the concrete object portrayed by the photograph."@en, "È un documento contenente metadati su una fotografia e sull'oggetto concreto rappresentato dalla fotografia."@it ;
rdfs:subClassOf fabio:EntityMetadata .

Figure 1. Simplified example of an OWL ontology and one of its classes overlaid with SKOS for inclusion in the /DH.arc Vocabularies repository

6. FUTURE WORK

The development and release of the /DH.arc Vocabularies semantic artefact repository offers an example of how existing software solutions can be adapted to the needs of smaller DH centers and institutions to enhance the FAIR-ness of their semantic artefacts by using existing infrastructure, staff, and technical skill levels to help make these research products more visible and reusable, a challenge that the humanities must contend with as the importance of these artefacts continues to increase. While the overlaying of OWL and SKOS we have detailed works well for simple OWL ontologies, such as the ones we have currently included in the repository, the approach nonetheless has limits for ontologies with more intricate formal semantics that simply cannot be reproduced in SKOS (such as restrictions).
The next phases of work for the repository include further exploration of how to handle more intricate OWL ontologies (for example by using SKOS notation properties to indicate semantics to an end user), evaluation of the Docker instance for continued production use, the addition of documentation for how to access the API, and the implementation of a documented internal workflow for conversion and publication of future ontologies and other semantic artefacts to ensure their timely addition to the repository and its continued use.

REFERENCES

Baclawski, K. (2009). The Open Ontology Repository initiative: Requirements and research challenges. *CEUR Workshop Proceedings*. https://ceur-ws.org/Vol-514/paper6.pdf

Belk, S., Wohlgenannt, G., & Polleres, A. (2015). Exploring and exploiting(?) the awkward connections between SKOS and OWL. International Workshop on the Semantic Web. http://ceur-ws.org/Vol-1486/paper_96.pdf

[29] The fully converted ontology and all others can be found at https://github.com/laurentfintoni/dharc-skosmos/tree/main/vocab_files

Bruseker, G., Carboni, N., & Guillem, A. (2017). Cultural heritage data management: The role of formal ontology and CIDOC CRM. In *Heritage and Archaeology in the DigitalAge* (pp. 93–131). Springer International Publishing.

Corcho, O., Ekaputra, F. J., Heibi, I., Jonquet, C., Micsik, A., Peroni, S., & Storti, E. (2024). A maturity model for catalogues of semantic artefacts. *Scientific Data, 11*(1), 479. https://doi.org/10.1038/s41597-024-03185-4

Cox, S. J. D., Gonzalez-Beltran, A. N., Magagna, B., & Marinescu, M.-C. (2021). Ten simple rules for making a vocabulary FAIR. *PLoS Computational Biology*, 17(6), e1009041. https://doi.org/10.1371/journal.pcbi.1009041

Ding, Y., & Fensel, D. (2001). Ontology Library Systems: The key to successful Ontology Reuse. International Conference on Semantic Web & Web Services, 93–112. https://files.ifi.uzh.ch/ddis/iswc_archive/iswc/ih/SWWS-2001/program/full/paper58a.pdf

d'Aquin, M., & Noy, N. F. (2012). Where to publish and find ontologies? A survey of ontology libraries. *Web Semantics*, 11, 96–111. https://doi.org/10.1016/j.websem.2011.08.005

Fintoni, L. A. (2024). Rethinking Scholarly Digital Objects as Cultural Heritage: The KNOT Project. In A. Di Silvestro & D. Spampinato (Eds.), Me.Te. Digitali. Mediterraneo in rete tra testi e contesti, Proceedings del XIII Convegno Annuale AIUCD2024 (pp. 582–587).

Hartmann, J., Palma, R., & Gómez-Pérez, A. (2009). Ontology Repositories. In *Handbook on Ontologies* (pp. 551–571). Springer Berlin Heidelberg.

Hepp, M., & de Bruijn, J. (2007). GenTax: A generic methodology for deriving OWL and RDF-S ontologies from hierarchical classifications, thesauri, and inconsistent taxonomies. In *Lecture Notes in Computer Science* (pp. 129–144). Springer Berlin Heidelberg.

Hodge, G. M., & Digital Library Federation. (2000). *Systems of knowledge organization for digital libraries: Beyond traditional authority files*. Washington, DC : Digital Library Federation, Council on Library and Information Resources.

Jansen, L. (2019). Ontologies for the Digital Humanities: Learning from the life sciences? *Joint Ontology Workshops*. https://ceur-ws.org/Vol-2518/paper-WODHSA5.pdf

Jonquet, C., Graybeal, J., Bouazzouni, S., Dorf, M., Fiore, N., Kechagioglou, X., Redmond, T., Rosati, I., Skrenchuk, A., Vendetti, J. L., Musen, M., & members of the OntoPortal Alliance. (2023). Ontology repositories and semantic artefact catalogues with the OntoPortal technology. In *The Semantic Web – ISWC 2023* (pp. 38–58). Springer Nature Switzerland.

Jupp, S., Bechhofer, S., & Stevens, R. (2008). *SKOS with OWL: Don't be Full-ish!* https://ceur-ws.org/Vol-432/owled2008eu_submission_22.pdf

Le Franc, Y., Parland-von Essen, J., Bonino, L., Lehväslaiho, H., Coen, G., & Staiger, C. (2020). *D2.2 FAIR Semantics: First recommendations*. Zenodo. https://doi.org/10.5281/ZENODO.3707984

Naskar, D., & Dutta, B. (2016). *Ontology libraries: A study from an ontofier and an ontologist perspectives*. http://docs.ndltd.org/collection/etd2016/etd16-hal-01398427-2.pdf

Poveda-Villalón, M., Espinoza-Arias, P., Garijo, D., & Corcho, O. (2020). Coming to terms with FAIR ontologies. In *Lecture Notes in Computer Science* (pp. 255–270). Springer International Publishing.

Suominen, O., Ylikotila, H., Pessala, S., Lappalainen, M., & Frosterus, M. (2015). *Publishing SKOS vocabularies with Skosmos*. https://skosmos.org/publishing-skos-vocabularies-with-skosmos.pdf

Tammaro, A. M. (2014). Evaluation of digital humanities: An interdisciplinary approach. In *Communications in Computer and Information Science* (pp. 136–146). Springer Berlin Heidelberg.

Voß, J. (2016). *Open Source web applications for Knowledge Organization Systems*. Zenodo. https://doi.org/10.5281/zenodo.61262

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3, 160018. https://doi.org/10.1038/sdata.2016.18

MODELING AN ONTOLOGY FOR HERITAGE SCIENCE: CHALLENGES AND KEY STRATEGIES

Erica Scarpa[1], Riccardo Valente[2], Irene Rossi[3]
[1] ISPC - CNR, Italy – erica.scarpa@cnr.it
[2] ISPC - CNR, Italy – riccardo.valente@cnr.it
[3] ISPC - CNR, Italy – irene.rossi@cnr.it

ABSTRACT (ENGLISH)
This paper presents the authors' contributions to the Humanities and Cultural Heritage Italian Open Science Cloud (H2IOSC) project, in particular the development of an ontology for Heritage Science. Given the inherently multidisciplinary nature of Heritage Science, defining precise boundaries for the domain is a complex task that can hinder effective data organization and limit research potential. After providing a brief overview of the H2IOSC project, the paper delves into the key challenges faced in the ontology development. It further outlines the initial strategies employed to address these challenges, emphasizing the establishment of a permanent interdisciplinary group of Heritage subject-matter experts and the creation of the H-SeTIS online database for surveying existing semantic resources in the Heritage field. The paper also highlights future developments and critical considerations for advancing the project, ensuring its continued relevance and impact and promotes the adoption of good practices for the development of semantic artefacts in the Heritage domain.
Keywords: Semantic Web; Ontologies; Heritage Science; Knowledge Elicitation

ABSTRACT (ITALIANO)
Modellare un'ontologia per l'Heritage Science: sfide e strategie.
Questo articolo presenta il contributo degli autori al progetto Humanities and Cultural Heritage Italian Open Science Cloud (H2IOSC), con particolare attenzione alla progettazione di un'ontologia per l'Heritage Science. Data la natura intrinsecamente multidisciplinare di questo dominio, la sua definizione può risultare complessa, con il rischio di compromettere l'organizzazione dei dati e limitare le potenzialità della ricerca. Dopo una breve introduzione al progetto H2IOSC, vengono illustrati i principali problemi nella modellazione di un'ontologia per l'Heritage Science. Vengono inoltre descritti i primi approcci e le strategie adottate per affrontare questa vasta tematica: particolare attenzione è dedicata alla costituzione di un gruppo interdisciplinare permanente di esperti del dominio dell'Heritage e alla creazione di un database online di risorse semantiche (H-SeTIS). Infine, vengono presentati gli sviluppi futuri e gli aspetti critici.
Parole chiave: web semantico; ontologie; Heritage Science; elicitazione della conoscenza

1. INTRODUCTION
This paper explores the ongoing development of an ontology for the Heritage Science (HS) domain: the semantic model will be a key component of the Humanities and Cultural Heritage Italian Open Science Cloud (H2IOSC) project. Funded by the Next Generation EU plan and the Italian Ministry of University and Research, the H2IOSC project involves twelve institutes of the Italian National Research Council (CNR). Its primary goal is to establish a collaborative, federated cluster connecting the four Italian nodes of European Research Infrastructures (RI) focused on Human Sciences and Cultural Heritage: CLARIN, DARIAH, E-RIHS, and OPERAS.
Specifically, the project aims to create a digital open ecosystem that supports advanced multidisciplinary research while facilitating the creation, access, and (re-)use of open scientific data. Within this framework, the Institute of Heritage Science (ISPC-CNR) plays a crucial role in ensuring resource interoperability for the E-RIHS infrastructure, particularly through the development of an ontology tailored to the needs of HS. This paper presents the methodology and challenges involved in designing this ontology, highlighting the collaborative efforts of subject-matter experts and the innovative approaches used to build a structured, interoperable knowledge framework that advances research and innovation in HS.

2. MODELLING CHALLENGES
HS is generally described as an interdisciplinary research field that integrates the social sciences, natural sciences, and humanities. The concept of HS was first introduced in 2006 by the Science and Technology Select Committee of the British House of Lords (House of Lords Science and Technology Select Committee,

2006). It aims to bridge traditional disciplinary divides by linking Cultural Heritage Conservation with the Social Sciences. However, its formal definition is relatively recent, having been jointly established in 2019 by the E-RIHS RI and the International Centre for the Study of the Preservation and Restoration of Cultural Property (ICCROM)[1].

The recognition of HS as a distinct domain is intended to bring together researchers from diverse disciplines and backgrounds who work on common heritage topics and objects, fostering deeper collaboration and interdisciplinary approaches.

However, the object of study of HS (*i.e.*, Heritage) is a multifaceted, intricate and dynamic concept, with its meaning and perception shifting significantly through space and time. Therefore, before exploring the specific strategy for modeling an ontology for HS, it is essential to understand the challenges involved in defining the scope of the domain itself.

Different approaches to Heritage can coexist, as its interpretation develops in response to local cultures, traditions, and geographical contexts. In some cases, these approaches share common roots, differing only in specific aspects: Heritage is seen as something valuable to preserve and pass down through generations. However, in other instances, theoretical foundations may be fundamentally opposed, as not all societies share the same principles for defining what is considered worthy of transmission. Moreover, the scope of Heritage is not uniform, as it encompasses a wide range of cultural assets—from traditional monuments and works of art to industrial archaeology—each carrying its own unique significance.

Change is also a defining characteristic of Heritage, as modifications in the global concept of Heritage occur rapidly, particularly throughout the second half of the 20th century and into the early 21st century (Harrison, 2020). However, the evolution of Heritage is neither uniform nor consistent. This variability, combined with the relatively recent establishment of HS as a domain, presents significant challenges in developing an HS ontology. Such an ontology must not only accommodate the diverse nature of existing Heritage but also remain adaptable to its future transformations. Ensuring the long-term sustainability of the ontology is crucial for supporting the entire RI and enabling the effective retrieval and integration of complex, evolving data over time.

For the purposes of the H2IOSC project and the development of an HS ontology, it was necessary to define the boundaries of Heritage—an essential step in identifying the domain's characteristics and requirements. It should be emphasized that the boundary definition outlined below is pragmatic, serving as both a reference and modeling strategy for establishing the scope of the ontology. While it is grounded in the most widely recognized and up-to-date theories, it is not necessarily intended to serve as a new formal definition of Heritage.

Among the most significant opposing pairs used for the classification of Heritage entities, Nature-Culture (with Culture here specifically referring to human-made elements) and Tangible-Intangible were identified. For what concerns the first pair, it must be highlighted that what may initially seem like a straightforward classification—a living elephant is different from a wooden statue of an elephant—has more complex nuances. A wooden statue carved by a human is indeed a cultural entity, despite its material being entirely natural. Similarly, a series of carved rock art drawings might appear to fall into a comparable category. However, the immovable nature of their rock support, which entirely belongs to the natural realm, shifts the classification threshold. In such cases, what exactly qualifies as cultural, and what remains natural? Are only the carved drawings considered cultural, while their physical support remains natural? That might seem reasonable, yet carved figures cannot exist independently of the surface that holds them. Does the rock, then, become a cultural entity simply because it bears cultural elements on its surface? If so, would it be correct to classify an entire massive rock as a cultural entity, even if only a small corner contains carvings? Should only the portion with the drawings be considered cultural? Or should the entire rock be viewed as a mixed natural-cultural entity?

These uncertainties deepen when exploring other examples in Heritage. A *bonsai*, for instance, is not entirely a human creation, nor it is purely natural; it is a living organism that humans shaped into a cultural expression. This borderline case illustrates the difficulty of applying a clear-cut distinction between natural and cultural entities within the domain of Heritage.

The contraposition of Nature-Culture was dismissed for the aforementioned reasons, while the pair Tangible-Intangible was kept as more definable and inherently connected to the scope of the E-RIHS RI. The definition of these boundaries was carried out also thanks to the direct involvement of Heritage subject-matter experts (see §5).

[1] https://www.e-rihs.eu/e-rihs-in-a-nutshell/.

3. MODELING AN ONTOLOGY FOR THE HERITAGE SCIENCE DOMAIN

Ontology engineering methodologies begin with knowledge elicitation—the systematic process of gathering, organizing, and refining knowledge to create a structured representation of the concepts, relationships, and semantics characterizing a particular domain (Leenheer, 2009). Subject-matter experts play a crucial role in this process, as their collaboration with knowledge engineers provides the theoretical foundation for an accurate representation (Shadbolt & Smart, 2015). Ontology modeling involves a delicate balance between achieving a shared understanding of key definitions and meanings within a domain and formalizing them into a machine-actionable model. Given the dynamic and evolving nature of the object of study in the HS domain, the negotiation between knowledge engineers and subject-matter experts becomes an even more critical phase, as it influences how data are represented and interpreted. No ontology engineering methodology specifically tailored for the Heritage or HS domain exists, as most methodologies have been developed for business and industry contexts. In recent years, Linked (Open) Data (Heath & Bizer, 2011) and FAIR guidelines (Wilkinson et al., 2016) have had a significant impact on how ontologies have been developed over the past decade. As a result, only the most recent methodologies incorporate these standards. The development of the HS ontology discussed here follows the Linked Open Terms (LOT) methodology (Poveda-Villalón et al., 2022), which incorporates the latest requirements for Linked Data and FAIR principles and promotes the reuse of existing ontologies and vocabularies.

The LOD methodology begins with defining the ontology's requirements specifications, starting with the scope definition. It also involves identifying case studies and real-world scenarios to assess the model's practical applicability. To support these initial steps, two preliminary goals were established: a comprehensive survey of semantic artefacts in the Heritage domain and the establishment of a permanent working group focused on defining the boundaries of HS.

For the first objective, the *Heritage – Semantic Tools and Interoperability Survey* (H-SeTIS, see §4) database served the purpose (Scarpa & Valente, 2024a, 2024b). H-SeTIS is a preliminary survey of all semantic artefacts—"machine-actionable and -readable formalizations of a conceptualization, enabling sharing and reuse by humans and machines" (Hugo et al., 2020, p. 12)—specifically designed for or used within the Heritage domain. Before modeling a complex domain like HS conducting a thorough and detailed review of the state of the art is crucial. This step ensures that the model can address relevant and current questions while meeting the evolving needs of the user community. Additionally, it helps avoid duplicating solutions already developed in existing models.

The second goal is the establishment of a permanent interdisciplinary group of subject-matter experts. This group, gathering around the Heritage Science Ontology RoundTable (‹H/SORT›, see §5), will provide a sustainable framework for collaboration among subject-matter experts from various HS disciplines, including epigraphy, history, archaeology, and conservation science. The group also aims to identify key research questions (KRQs), ensuring that decision-making processes are well-informed, integrating and balancing the different perspectives on Heritage from each discipline.

4. H-SeTIS

The preliminary survey of semantic artefacts within the Heritage field provided a clear state-of-the-art overview for developing these resources. The *Heritage – Semantic Tools and Interoperability Survey* database (https://h-setis.cnr.it/) serves as a robust foundation for ontology development, addressing gaps left by existing aggregators such as BARTOC (Ledl & Voß, 2016) and Linked Open Vocabularies (Vandenbussche et al., 2017), which often fail to capture the complexity of the Heritage domain. H-SeTIS catalogs five core types of semantic artefacts: ontologies, metadata standards, thesauri, application profiles, and software (Scarpa & Valente, 2024a, 2024b). Each resource is documented in detail, including its availability, development status, URI/repository, contributors, language, formats, licensing, keywords, and related bibliographic references. Relationships between resources are mapped to highlight integrations and dependencies. Additionally, editors of all identified semantic artefacts are being contacted to evaluate the responsiveness of each resource and to establish a dedicated contact person.

The H-SeTIS database also includes a curated bibliography of publications referred to the surveyed semantic artefacts and more general resources for semantic web development, managed through Zotero (https://www.zotero.org/). Zotero APIs enable seamless integration of bibliographic data into the H-SeTIS interface, with the shared library publicly accessible (https://www.zotero.org/groups/5434475/). Using Kerko (https://pypi.org/project/Kerko/), the bibliography is displayed within the H-SeTIS front-end.

Although the survey is still ongoing, it has already revealed a rich landscape of semantic artefacts in the Heritage domain. However, it has also exposed major shortcomings in adherence to good practices—one of the key reasons why aggregators like BARTOC and LOD fail to adequately represent semantic artefacts in this field. Many artefacts lack fundamental features such as persistent URIs, or exist solely as conceptual models published in scientific papers. Since its launch in December 2023, the survey has uncovered critical challenges in the Heritage sector, including insufficient documentation, poor adherence to FAIR principles, and limited application of Linked (Open) Data practices. Many resources lack semantic versioning, or active maintenance (Fig. 1). Preliminary findings show that at least half of the surveyed ontologies are inactive or lack status information.
To address these challenges, the ‹H/SORT› working group is developing a collection of best practices for the creation and maintenance of semantic resources in the Heritage field. This initiative, currently in progress, includes defining status descriptions for resources, inspired by the OBO Foundry framework (Smith et al., 2007) but adapted to accommodate the broader range of Heritage-specific artefacts.

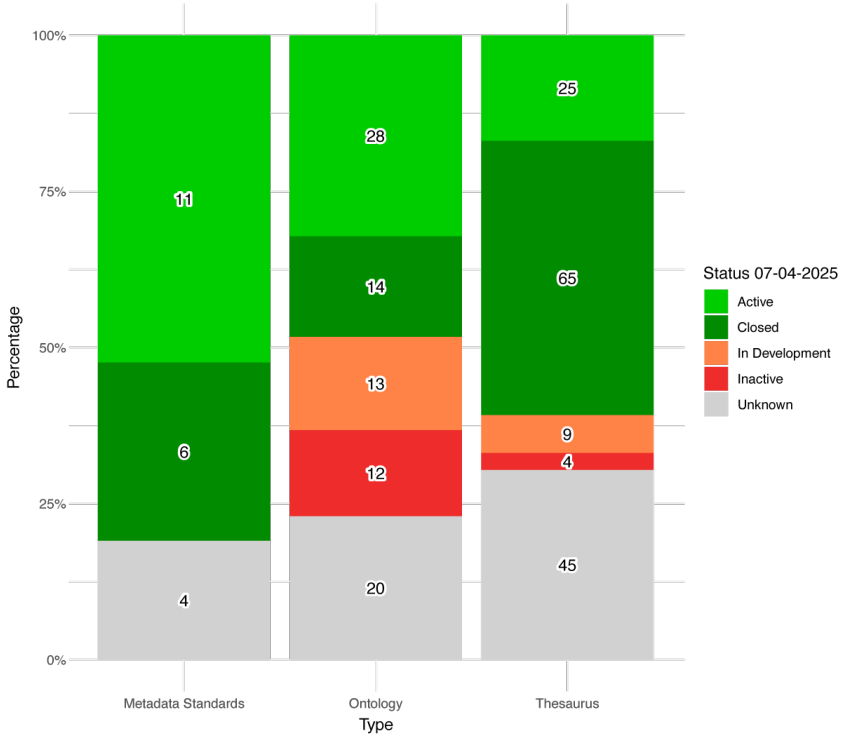


Figure 1 - Availability statuses of three kinds of semantic artefacts recorded in H-SeTIS. Digits in the bar segments display the number of artefacts.

5. ‹H/SORT›: Heritage Science Ontology RoundTable

The LOD methodology for ontology development (Poveda-Villalón et al., 2022) aligns with agile software development, a strategy that promotes close collaboration, face-to-face discussions, and flexible management to prioritize user needs and accelerate development. This approach emphasizes customer satisfaction, iterative improvements, and the creation of essential documentation[2].
Within this operational framework, the ‹H/SORT› working group gathers developers and subject-matter experts to explore the foundations of HS through seamless exchange and discussion, identifying key topics and case studies. The first roundtable of the working group, held in Venice on January 16-17, 2025, introduced an elicitation model focused on collaborative debate among subject-matter experts from specific HS subdomains, identifying diverging opinions on specific topics. The main focus of this first meeting was epigraphy, as the shared case study for all of the four RIs involved in the H2IOSC Work Package devoted to interoperability was the “inscribed object” (*i.e.*, any object carrying an inscription). A second roundtable, held in Milan on March 26, further addressed issues related to knowledge representation for HS, involving a group of experts from STEM disciplines, who brought a different and complementary approach compared to the Humanities.

[2] https://agilemanifesto.org/

The ‹H/SORT› hub (https://hsort-ec1af2.gitlab.io/) serves as a knowledge framework for developing the HS Ontology, where official, lightweight, and focused documentation is continuously updated based on insights from the roundtable discussions and feedback from the ‹H/SORT› working group. The hub also supports ongoing collaboration and tracks developments, ensuring that progress is monitored and results are effectively disseminated.
The ‘Topics’ section of the ‹H/SORT› hub collects a series of short posts that introduce key issues and themes related to Heritage. These posts, designed to be light, concise, and limited in bibliography, are carefully selected to stimulate discussion and debate. One of the key objectives of the ‘Topics’ section is to introduce the main themes and challenges encountered during the development of the HS Ontology to a broader audience. Some topics address core issues in Heritage Studies, such as the characteristics and transformations of Heritage, while others focus on the application of semantic technologies in the Heritage domain. For instance, one post explores the distinction between ‘Provenance’ and ‘Provenience’. The term provenance, commonly used by art historians, refers to the ownership history of an object. Conversely, provenience, typically used in archaeology, refers to the precise physical location where an object was discovered (Scarpa, 2025a). While the difference between these terms—the contrast between a “chain of custody” and a “specific place”—may seem minor or even irrelevant in certain contexts, it is a crucial distinction when dealing with Heritage data and its semantic representation. The varied, and at times contradictory, usage of these terms presents a significant challenge for accurate and consistent modeling.
Another section of ‹H/SORT› is dedicated to case studies, which serve as valuable tools for illustrating real-world applications and scenarios. An effective use of case studies involves selecting examples that differ in their characteristics. Diversity is essential for informing the correct modeling of classes and properties, ensuring that the ontology aligns with actual needs and use cases. For instance, one case study (Scarpa, 2025b) on ‹H/SORT› considers a standard text reproduced as an inscription on several, and quite different, supports (stone vessels, bowls, but even a natural shell). While scholars can easily recognize and interpret such variations, they present modeling challenges that must be carefully addressed to ensure proper representation within a semantic ecosystem.

6. CONCLUSIONS

The development of an ontology for the Heritage Science (HS) domain presents both unique challenges and significant opportunities for advancing research and represent the interdisciplinary nature of HS. By fostering a collaborative framework with subject-matter experts, the ‹H/SORT› working group aims to create a sustainable and interoperable knowledge structure that will support future research in the field. The adoption of current methodologies is central to integrating recent standards, such as Linked Open Data and FAIR principles, into the ontology model. Additionally, the incorporation of updated theoretical foundations seeks to provide an interpretive framework that has been largely unexplored to date. Moreover, the ‹H/SORT› working group is dedicated not only to documenting the ontology itself, with appropriate metadata, but also to thoroughly documenting the development process.

ACKNOWLEDGEMENTS

H2IOSC Project - Humanities and cultural Heritage Italian Open Science Cloud funded by the European Union NextGenerationEU - National Recovery and Resilience Plan (NRRP) - Mission 4 “Education and Research” Component 2 “From research to business” Investment 3.1 “Fund for the realization of an integrated system of research and innovation infrastructures” Action 3.1.1 “Creation of new research infrastructures strengthening of existing ones and their networking for Scientific Excellence under Horizon Europe” - Project code IR0000029 - CUP B63C22000730005. Implementing Entity CNR.

REFERENCES

Harrison, R. (2020). *Il patrimonio culturale. Un approccio critico* (V. Matera & L. Rimoldi, Eds.).
Heath, T., & Bizer, C. (2011). *Linked Data: Evolving the Web into a Global Data Space*. Springer Cham. http://linkeddatabook.com/editions/1.0/
House of Lords Science and Technology Select Committee. (2006). *Science and Heritage. Report with Evidence*. https://publications.parliament.uk/pa/ld200506/ldselect/ldsctech/256/256.pdf
Hugo, W., Le Franc, Y., Coen, G., Parland-von Essen, J., & Bonino, L. (2020). *D2.5 FAIR Semantics Recommendations Second Iteration*. https://zenodo.org/records/5362010
Ledl, A., & Voß, J. (2016). Describing Knowledge Organization Systems in BARTOC and JSKOS. In H. E. Thomsen, A. Pareja-Lora, & B. N. Madsen (Eds.), *Ledl, Andreas; Voß, Jakob (2016). Describing*

Knowledge Organization Systems in BARTOC and JSKOS. In: Thomsen, Hanne Erdman; Pareja-Lora, Antonio; Madsen, Bodil Nistrup (eds.) TKE 2016 the 12th International conference on Terminology and Knowledge Engineering (pp. 168-178). Denmark: Copenhagen Business School (pp. 168–178). Copenhagen Business School. https://research.cbs.dk/en/publications/term-bases-and-linguistic-linked-open-data-proceedings-of-tke-201

Leenheer, P. D. (2009). Ontology Elicitation. In L. Liu & M. T. Özsu (Eds.), *Encyclopedia of Database Systems* (pp. 1966–1972). Springer US. https://doi.org/10.1007/978-0-387-39940-9_1316

Poveda-Villalón, M., Fernández-Izquierdo, A., Fernández-López, M., & García-Castro, R. (2022). LOT: An Industrial Oriented Ontology Engineering Framework. *Engineering Applications of Artificial Intelligence*, *111*, 104755. https://doi.org/10.1016/j.engappai.2022.104755

Scarpa, E. (2025a, January 20). Provenance and Provenience. *Heritage Science Ontology RoundTable (‹H/SORT›)*. https://doi.org/10.5281/zenodo.15166679

Scarpa, E. (2025b, January 28). Case Study #1: Identical text on multiple objects. *Heritage Science Ontology RoundTable (‹H/SORT›)*. https://doi.org/10.5281/zenodo.15167824

Scarpa, E., & Valente, R. (2024a). A Resource Hub for Interoperability and Data Integration in Heritage Research: The H-SeTIS Database. *Archeologia e Calcolatori*, *35*(1), 543–562. https://doi.org/10.19282/ac.35.1.2024.32

Scarpa, E., & Valente, R. (2024b). *Heritage – Semantic Tools and Interoperability Survey* [Dataset]. https://h-setis.cnr.it/

Shadbolt, N., & Smart, P. R. (2015). Knowledge Elicitation: Methods, Tools and Techniques. In N. Shadbolt, P. R. Smart, & J. R. Wilson (Eds.), *Evaluation of Human Work* (pp. 163–200). CRC Press. https://eprints.soton.ac.uk/359638/

Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L. J., Eilbeck, K., Ireland, A., Mungall, C. J., OBI Consortium, Leontis, N., Rocca-Serra, P., Ruttenberg, A., Sansone, S.-A., Scheuermann, R. H., Shah, N., Whetzel, P. L., & Lewis, S. (2007). The OBO Foundry: Coordinated Evolution of Ontologies to Support Biomedical Data Integration. *Nature Biotechnology*, *25*(11), 1251–1255. https://doi.org/10.1038/nbt1346

Vandenbussche, P.-Y., Atemezing, G. A., Poveda-Villalón, M., & Vatant, B. (2017). Linked Open Vocabularies (LOV): A Gateway to Reusable Semantic Vocabularies on the Web. *Semantic Web*, *8*(3), 437–452. https://doi.org/10.3233/SW-160213

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., Da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Scientific Data*, *3*(1), 1–9. https://doi.org/10.1038/sdata.2016.18

Linked Open Data and IIIF for connecting manuscripts images with their transcriptions: a case study from the Veneranda Biblioteca Ambrosiana

Lorenza Talarico
Catholic University of the Sacred Heart, lorenzatalarico@gmail.com

ABSTRACT (ENGLISH)

This paper presents a case study that leverages Linked Open Data (LOD) and the International Image Interoperability Framework (IIIF) to enhance the accessibility of *De Educandis Ingeniis* by Federico Borromeo, a foundational manuscript of the Veneranda Biblioteca Ambrosiana. This 17th century treatise, reflecting Borromeo's dedication to promoting knowledge for the common good, has been digitized and linked with its textual transcriptions to create an innovative digital edition. The project culminated in a prototype developed on the Simple Annotation Server (SAS), allowing scholars to engage with the manuscript interactively.

The research builds on advancements in digital technologies that have transformed libraries into hubs for cultural dissemination. As the first Italian institution to adopt the IIIF standard, the Ambrosiana offers a rich collection of ancient works through its Digital Library. This study explores how IIIF and LOD can bridge the gap between historical artifacts and contemporary digital platforms, improving usability, accessibility and preservation.

The research was structured into three phases. The first addressed the theoretical foundations of LOD and the Semantic Web while defining the research problem. The second focused on technical implementation, employing Transkribus to create a Handwritten Text Recognition (HTR) model, TEI XML and ALTO formats for transcription, and IIIF standard for image delivery. The final phase developed a prototype on SAS, integrating annotations stored as linked data to ensure interoperability.

This case study demonstrates the integration of cutting-edge technologies to establish a replicable workflow for digital editions of manuscripts and the results highlight improved academic utility and accessibility while bridging tradition and innovation. By proposing a model for linking images and texts, the research addresses the lack of a unified workflow and opens pathways for future development in digital humanities.

Keywords: Linked Open Data; IIIF; Digital library; Transkribus; HTR

ABSTRACT (ITALIANO)

Linked Open Data e IIIF per collegare le immagini dei manoscritti alle loro trascrizioni: un caso studio dalla Veneranda Biblioteca Ambrosiana

Questo articolo presenta un caso di studio che utilizza i Linked Open Data (LOD) e l'International Image Interoperability Framework (IIIF) per migliorare l'accessibilità del *De Educandis Ingeniis* di Federico Borromeo, un manoscritto fondamentale conservato presso la Veneranda Biblioteca Ambrosiana. Questo trattato del XVII secolo, testimonianza dell'impegno di Borromeo nella promozione della conoscenza per il bene comune, è stato digitalizzato e collegato alla sua trascrizione testuale per creare un'edizione digitale innovativa. Il progetto si è concluso con lo sviluppo di un prototipo sul Simple Annotation Server (SAS), che consente agli studiosi di interagire con il manoscritto.

La ricerca si basa sugli sviluppi delle tecnologie digitali che hanno trasformato le biblioteche in centri di diffusione culturale. L'Ambrosiana, prima istituzione italiana ad adottare lo standard IIIF, offre una vasta collezione di opere antiche attraverso la sua biblioteca digitale. Lo studio esplora come IIIF e LOD possano colmare il divario tra artefatti storici e piattaforme digitali contemporanee, migliorando l'usabilità, l'accessibilità e la conservazione dei documenti.

Il lavoro si articola in tre fasi principali. La prima fase ha affrontato le basi teoriche dei LOD e del Web Semantico, definendo il problema di ricerca. La seconda si è concentrata sull'implementazione tecnica, utilizzando Transkribus per creare un modello di riconoscimento della scrittura a mano (HTR), i formati TEI XML e ALTO per la trascrizione e lo standard IIIF per la distribuzione delle immagini. La fase finale ha portato allo sviluppo di un prototipo su SAS, integrando annotazioni archiviate come linked data per garantire l'interoperabilità.

Questo caso studio dimostra come l'integrazione di tecnologie all'avanguardia possa stabilire un workflow replicabile per le edizioni digitali di manoscritti. I risultati evidenziano una maggiore utilità e accessibilità accademica, creando un ponte tra tradizione e innovazione. Proponendo un modello per collegare immagini

e testi, la ricerca affronta l'assenza di un workflow unificato e apre la strada a nuovi sviluppi nelle digital humanities.
Parole chiave: Linked Open Data; IIIF; Biblioteca digitale; Transkribus; HTR

1. INTRODUCTION

Nowadays, the integration of digital technologies into historical libraries is proving to be not only inevitable but essential, transforming these institutions from static knowledge repositories into dynamic instruments of cultural dissemination (Cusimano, 2015). This transformation is supported by advancements in technological standards such as Linked Open Data (LOD) and the International Image Interoperability Framework (IIIF), which have enabled the creation of interoperable digital editions accessible to researchers worldwide. Among the few Italian institutions leading this transition is the Veneranda Biblioteca Ambrosiana in Milan. Hence, this case study explores the use of LOD and IIIF standard to link images and text, making one of the Library's manuscripts, *De educandis ingeniis* by Federico Borromeo, accessible in a newly enhanced digital format. That is to say, this brief treatise represents an adequate starting point for testing the integration of digital images with textual transcriptions. Structured into three phases, the approach aims to improve searchability, facilitate academic use and establish a replicable workflow for other manuscripts. In other words, it seeks to answer a central research question: can modern technologies bridge the gap between ancient manuscripts and contemporary digital tools to enhance usability and preservation? The initial phase of the research contextualized the project by addressing the theoretical foundations of LOD and the Semantic Web. The second phase focused on technical implementation, encompassing the use of Transkribus to develop a Handwritten Text Recognition (HTR) model, the employment of TEI XML and ALTO formats and the adoption of IIIF standards. This process culminated in the creation of a final digital edition prototype on the Simple Annotation Server (SAS), enabling the side-by-side visualization of images and their corresponding transcriptions. In fact, the research involved uploading via SAS the annotations of the text transcriptions into the IIIF manifest, sourced directly from the Digital Library of the Ambrosiana, following the principles of Linked Open Data (LOD). As a matter of fact, the LOD paradigm offers significant advantages for language data, such as structural and conceptual interoperability, federation, the expressivity of semantic web models and the network effect. Given the context of this study and its placement within the realm of the Semantic Web, the use of the LOD paradigm was imperative for addressing the research problem. Therefore, the aim was to develop a model that can be replicated and adapted for other manuscripts, enhancing cultural heritage through innovation and accessibility. By bridging traditional practices with advanced technological tools, it addresses the absence of a unified data-structuring model in the community.

2. DATA AND TECHNOLOGIES EMPLOYED

The data at the center of this case study include the IIIF images of *De educandis ingeniis*, published on the Digital Library of the Veneranda Biblioteca Ambrosiana website, along with the corresponding IIIF manifest and the transcription of the manuscript. The Ambrosiana's Digital Library aims to enhance the primary funds, namely Inferior, Superior, S.P. and Trotti, through digitization and open online access. This initiative not only ensures the preservation of these invaluable manuscripts but also promotes their use for study and research. As the first library in Italy to adopt the IIIF standard, the Ambrosiana employs tools like the Mirador viewer and the image server Cantaloupe to provide broad accessibility. Indeed, the IIIF ecosystem is integrated with the library's OPAC (On-line Public Access Catalog), ensuring a direct link between a manuscript's catalog description and its corresponding digital resource (Cusimano, 2019). IIIF, a set of open standards supported by a global consortium of cultural institutions, offers significant advantages, including high-quality image delivery with advanced zooming and manipulation options, compatibility with diverse tools for viewing and comparing images and the flexibility to "publish once and reuse often" across different platforms (Salarelli, 2017).

The transcription of the manuscript was conducted using Transkribus, a platform developed during two EU-funded research projects and maintained by READ-COOP SCE since July 2019. In short, Transkribus, powered by Java technology, facilitates manual and automatic transcription, the training of custom AI models, advanced search tools and exporting documents in various formats. However, Transkribus is a "closed environment", with certain limitations, particularly compatibility issues that arise when attempting to use the work, produced within the platform, externally. These challenges, alongside bureaucratic and authorization obstacles and platform copyright restrictions, collectively hinder seamless adherence to LOD principles. Therefore, the goal of this specific study with *De educandis ingeniis* was to export the work

conducted and achieve a similar visualization result outside of the Transkribus environment, to ensure authentic interoperability and alignment with LOD paradigm. For *De Educandis Ingeniis*, a HTR model was created on Transkribus by transcribing the entire manuscript to produce a Gold Standard. The transcription was then exported in multiple formats, including TEI XML, a format widely used in Digital Humanities for encoding and managing digital data. Despite its flexibility and scholarly utility, integrating TEI with IIIF remains challenging and lacks standardization (Monella & Rosselli Del Turco, 2020). As a consequence, after evaluating various approaches, ALTO files exported from Transkribus were then selected for further use.

ALTO, an open XML Schema developed under the EU-funded project METAe, provides a standardized format for describing the layout and content of physical text resources, such as book pages or newspaper articles. The real strength of ALTO format in this case study lies in its ability to serve as a key connecting link, ensuring a smooth workflow without loss of work or redundancy during the annotation phase: it enables efficient transformation, adaptation and linking of data, all within a LOD perspective. In fact, the ALTO files were converted into annotations using the Alto2Annotations transformer and subsequently uploaded to SAS, the last technology employed in this project to create a visualizable prototype via Mirador. SAS, a Java-based annotation server developed by Glen Robson, the Technical Coordinator of the IIIF Consortium, offers an accessible and open-source solution, allowing users to work locally on their computers while sharing the results through links when needed. It supports both IIIF and the Mirador viewer, enabling the creation of annotations stored as linked data in an Apache Jena triple store. This approach aligns with Linked Open Data principles, enhancing interoperability (Ide & Pustejovsky, 2010) and semantic connections across the web and underscores the potential for linked data to serve as a foundation for future and similar projects. However, SAS can experience downtime and this issue arises because this server relies not only on its own stability but also on the stability of the website hosting the referenced IIIF manifest. Since SAS follows the LOD paradigm, everything is interoperable and interconnected, meaning also that if one component fails, it may affect the entire system. Regardless, SAS provides a configurable backend capable of leveraging various technologies, including Jena and Sesame for Resource Description Framework (RDF) support, as well as SOLR. It also supports custom annotation bodies formatted in RDFa, allowing for greater flexibility in how annotations are structured and utilized. Hence, future works could explore these features of the server to fully leverage Linked Open Data and enhance data analysis, enrichment and information retrieval.

3. WORKFLOW

The workflow and methodology of this project started with the study and application of Transkribus within the Veneranda Biblioteca Ambrosiana, but have progressively evolved through research, experimentation and overcoming various challenges. This progression is driven by the fact that the project involves a cutting-edge technology still under development.

The initial step was to select and review the manuscript *De educandis ingeniis* and its history, based on feasibility considerations, such as time available and the manageable amount of data. Subsequently, the manuscript images were correctly uploaded to the Digital Library in IIIF format, alongside the rest of the catalog already available on the platform. *De educandis ingeniis* model was created for the HTR of this type of handwriting, and all 92 pages of the manuscript were consistently and accurately transcribed line by line, forming a Gold Standard. The result, beyond the AI model, was the ability to visualize the manuscript pages, uploaded as JPEG files generated by the "master files", alongside their corresponding transcriptions within the platform, and the work downloadable in various formats. However, it became evident that for data preservation and maintenance, but primarily due to the inability to integrate Transkribus's workflow with the existing IIIF Mirador architecture on the Ambrosiana Digital Library website, further use of Transkribus was no longer necessary. Despite this, the work done on this platform remains relevant, as it provided the HTR model based on the manuscript, serving as the starting point for the entire case study.

Thereafter, exploring the various download formats, an attempt was made to integrate the IIIF images from the Digital Library, specifically the IIIF manifest, with the TEI XML file downloaded from Transkribus containing the transcription. This involved research into previous attempts by the community, particularly the work of Paolo Monella and Fabio Cusimano (Monella & Cusimano, 2019), as well as existing theoretical approaches. In theory, the TEI format meets all interoperability and extensibility requirements (Burnard, 2014), as the IIIF manifest of the manuscript could potentially reference it as an external resource by modifying the manifest with the referenced annotations technique (Monella & Rosselli Del Turco, 2020).

Within the TEI XML file a key element is <*facsimile*>, which connects the transcription to the corresponding images of the manuscript pages. This connection is essential for aligning the digital representation of the manuscript with its physical counterpart, enabling users to view the original image alongside its transcription on compatible platforms, such as EVT (Monella & Rosselli Del Turco, 2020). However, in this case study, EVT, despite being the current state-of-the-art within the community, was not employed. This decision was based on the fact that EVT is primarily tailored for critical editions of philological texts and is designed as a specialized system for a specific use case. As such, it was considered less appropriate for the broader needs of the Digital Library of the Ambrosiana. Consequently, the integration of TEI and IIIF is theoretically possible and experts are actively working on this process. However, for this specific project, no smooth or viable solutions have been found to date.

To address the initial research problem and achieve the project's objectives, an alternative solution was sought to leverage Linked Open Data for displaying manuscript images alongside their transcriptions. This research led to the discovery of SAS, which aligns with the LOD paradigm by storing annotations as linked data in an Apache Jena triple store. Compatible with IIIF and the Mirador viewer, SAS enhances interoperability and semantic connections across the web, reflecting LOD principles and maximizing the potential for data integration. Additionally, this server facilitates the annotation of IIIF manifests and enables their automatic modification while allowing the upload of JSON files as annotations. This ensures the preservation of previous work and enables a smooth workflow without unnecessary delays. Regarding *De educandis ingeniis*, 92 XML files, each corresponding to a manuscript page, were indeed exported from Transkribus in ALTO format, chosen for its compatibility with SAS. ALTO is widely utilized in OCR and its integration with the Alto2annotations transformer facilitated the conversion of these files into JSON annotations that are compliant with IIIF standards. This process ensured the seamless transfer of work from Transkribus to SAS, effectively linking the transcriptions to each manuscript page available online and maintaining an efficient workflow throughout the annotation process.

The objective was to establish a single prototype for this project utilizing the modified IIIF manifest through SAS: this, thus, involved structuring the data via the latter and revising the IIIF manifest obtained through a link from the Digital Library. The modification employs the other content technique, which incorporates links to Annotation Lists that point to external resources, namely, JSON files created by SAS containing the annotations written for each manuscript page. Before achieving this, it was necessary to convert all 92 ALTO files using the Alto2annotations transformer. However, the first visualization in Mirador showed that the rectangles did not correspond to the lines of text, resulting in disorganized transcriptions. To address this issue, line level annotations were utilized (instead of word level ones) and the coordinates were removed from each file, allowing the annotation for each page to refer to the entire corresponding page. SAS required JSON files compatible with Open Annotations IIIF v.2, necessitating the use of the "annotationListNoArt.xsl" XSLT from the transformer. Subsequently, the output JSON files were manually modified to include the URI of the corresponding canvas from the published IIIF manifest in the Digital Library, ensuring proper communication and interoperability. Once the files were prepared, all annotations were successfully uploaded to SAS. Hence, the server generated the modified IIIF manifest and allows for visualization in Mirador or Universal Viewer, as well as sharing the work via a link.

As mentioned before, SAS employed Annotation Lists to add annotations, specifically transcriptions, as this technique aligns with IIIF version 2. Annotation Lists are a method of grouping annotations, often at the canvas or page level. They may include resources like page transcriptions or OCR data formatted as annotations. These lists are linked to the specific canvas they pertain to, serving as independent resources that should be accessed separately when needed. This separation from the main manifest allows clients to display images quickly, enhancing user experience by adding additional content and commentary as the user interacts with individual canvases. Additionally, Annotation Lists are typically resolvable, meaning their *@id* can be entered in a web browser to retrieve the list directly. As a result, they can be generated dynamically, while the manifest remains static and easily cached. In the URI pattern, the *{name}* parameter should be unique to distinguish one Annotation List from others. This parameter often matches the canvas name, though it is not mandatory, as a single canvas might contain multiple lists categorized by resource type. Therefore, it is important to follow the specific URIs provided, rather than generating them in advance. Furthermore, each Annotation List must have a unique HTTP(S) URI specified in the *@id* field, which should return a JSON representation when accessed. Annotations are included in a *resources list* and the format should match the media type returned upon dereferencing the resource. The *content* type should be selected from a list in the Open Annotation specification or other widely recognized ontologies. For resources directly rendered, like images, transcriptions or musical performances, the

motivation should be *sc:painting*. Content resources can also contain additional fields, such as labels, descriptions, metadata, licenses and attribution.
In the context of this case study, the JSON file of the IIIF manifest published in the Digital Library includes the key elements which are characteristic of a manifest. Within the *sequences* property, the canvases are listed. Upon examining the modified and annotated IIIF manifest generated through SAS, it appears very similar to the original one. However, when analyzing the *sequences*, a new property named *otherContent* is found following the *thumbnail* property: it has been utilized by SAS to link additional content, specifically an Annotation List, to the manifest. This represents the main part of the new manifest, with the *@id* linking to another JSON file, specifically an Annotation List for each page of the manuscript. It is important to note that this structure is consistent across all 92 images, and thus applies to every canvas within the manifest in the same way. In analyzing the JSON of the Annotation List, the text of the page transcription in the text/html format and the motivation of the annotation, specifically *sc:painting*, are immediately evident and this confirms the theoretical framework discussed previously. With this structured, interoperable, precise and hierarchical data organization, the desired and final prototype becomes accessible on Mirador.

4. RESULTS AND DISCUSSION

In Mirador the manuscript can be virtually browsed page by page, with each image featuring an annotation corresponding to the text transcription completed at the project's outset using Transkribus. This approach successfully achieves the initial objective of the research: to add the transcription alongside the manuscript and link it to the IIIF manifest of the manuscript's images. This enhances interoperability and provides the Ambrosiana Library with a methodology to potentially further develop in the future its digital library system. The goal is thus to make its invaluable collection even more accessible, interoperable and reusable for researchers, students and enthusiasts worldwide (Cusimano, 2021).
Hence, this work proposes a solution for integrating images and text within the Ambrosiana Digital Library, employing LOD and IIIF technologies to make the manuscript *De educandis ingeniis* accessible in a newly digital and interactive format. The workflow was designed to overcome various bureaucratic, technical and copyright limitations associated with Transkribus, as well as the current lack of necessary Mirador plugins in the Ambrosiana Digital Library for annotating IIIF images. By progressing from transcription in Transkribus to annotation in SAS through multiple steps of format conversion, a digital edition prototype was developed, offering both accessibility and interactive research opportunities to scholars and enthusiasts. Although the results remain within the research domain and the prototype is not yet available on the Ambrosiana website, they highlight how LOD and IIIF could significantly facilitate the navigation, search and consultation of digitized manuscripts while preserving the quality and integrity of the original materials. This model also shows promise for replication with other manuscripts and even across other libraries, thus supporting a broader and more systematic dissemination of cultural heritage (Spina, 2023).
Nevertheless, certain constraints were encountered, as these technologies represent cutting-edge advancements that continue to evolve. Restrictions within Transkribus's closed ecosystem and the challenges of managing complex data partially limited the workflow's effectiveness in a standardized digitization context. As a consequence, the community shall prioritize research on the link between text and images and future developments could include collaborations with other platforms and institutions to enhance interoperability and accessibility, as well as establishing a unified and shareable data structuring model. However, future works should also encompass other technical improvements, such as supporting alternative encoding strategies to enable links to external resources, while ensuring that the entire edition, including XML, internal and external data and software, remains sustainable, durable and aligned with the FAIR principles (Wilkinson, Dumontier, Aalbersberg et al, 2016). Moreover, continuing to use the SAS platform, it would be beneficial to conduct an in-depth study on the coordinates required for accurate annotations. It would be valuable to establish and standardize a method to convert image coordinates from ALTO format via Transkribus to IIIF format. That is to say, while this essential conversion is technically feasible, its implementation remains challenging. If successfully completed, it would help meet the parameters required by the transformer on SAS, enabling word level annotation. This, in turn, would lead to a more sophisticated, precise and efficient prototype that closely resembles the output of Transkribus. Furthermore, it is clear that the future aim would be to apply this workflow to an increasing number of manuscripts, initially from the Ambrosiana Library and later from other institutions, continually refining the process to make it more streamlined and efficient: integrating manuscript images with a searchable text edition is a critical advancement for the academic world, as there currently appears to be

no established standard for achieving this goal (Spina, 2023). In this context, the adoption of LOD is not merely conceptual, but functional: the project embraces structural and conceptual interoperability, the expressiveness of RDF-based models and the potential for data federation. In future developments, data enrichment could be enhanced, cross-referencing with external authority datasets and improve information retrieval capabilities across the cultural heritage domain.
In conclusion, this project currently represents a small yet valuable contribution to the community and cultural resource management, fostering an ongoing dialogue between innovation and tradition. The developed prototype can facilitate the digital preservation of manuscripts while enhancing their accessibility and usability, echoing the ideal of *ad publicum commodum et utilitatem*, namely the common good, which inspired Federico Borromeo in founding the Veneranda Biblioteca Ambrosiana. Ultimately, history repeats itself: while contexts and technologies may evolve and change, the enduring spirit of public service and cultural sharing remains a constant in society.

REFERENCES

Burnard, L. (2014). What Is the Text Encoding Initiative?, OpenEdition Press, Marseille.

Cusimano, F. (2015). Bibliotecario digitale, umanista informatico. Bibelot. Notizie dalle biblioteche toscane, 21, 14-17.

Cusimano, F. (2019). La nuova biblioteca digitale della Veneranda Biblioteca Ambrosiana. L'Almanacco Bibliografico. Bollettino trimestrale di informazione sulla storia del libro e delle biblioteche in Italia, La questione, Edizioni CUSL – Università Cattolica del Sacro Cuore, Milano, 1-2.

Cusimano, F. (2021). The IIIF-based Digital Library of the Veneranda Biblioteca Ambrosiana. Umanistica Digitale, 10, 423-432.

Ide, N. & Pustejovsky, J. (2010). What Does Interoperability Mean, Anyway? Toward an Operational Definition of Interoperability for Language Technology. Proceedings of the Second International Conference on Global Interoperability for Language Resources, Hong Kong, China.

Monella, P. & Cusimano, F. (2019). Linking Text and image: TEI XML and IIIF, http://www.paolomonella.it/reires2019/index.html.

Monella, P. & Rosselli Del Turco, R. (2020). Extending the DSE: LOD Support and TEI/IIIF Integration in EVT. Marras, C., Passarotti, M.C., Franzini, G. & Litta, E. (2020). Atti del IX Convegno Annuale dell'Associazione per l'Informatica Umanistica e la Cultura Digitale (AIUCD). La svolta inevitabile: sfide e prospettive per l'Informatica Umanistica, Università Cattolica del Sacro Cuore, AIUCD, Milano.

Salarelli, A. (2017). International Image Interoperability Framework (IIIF): una panoramica. JLIS.it, 8, 50-66.

Spina, S. (2023). Handwritten Text Recognition as a digital perspective of Archival Science. AIDAinformazioni, 115-132.

Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018. https://doi.org/10.1038/sdata.2016.18

Describing Monastic Iconography Using Semantic Data: A Preliminary Investigation

Sofia Baroncini[1], Francesco Mele[2]
[1] Leibniz Institute of European History, Germany - Baroncini@ieg-mainz.de
[2] Pegaso University, Italy - francesco.mele@unipegaso.it

ABSTRACT (ENGLISH)

In recent years, significant developments have occurred in the Digital Humanities field in the analysis of artworks' iconography. Nevertheless, these studies often fall short of historians' specialised research, such as the studies about monastic iconography, which reconstructs the parallelism with hagiographies (i.e., texts describing the saints' lives), and their visual representation. Key moments in a saint's life described in hagiographies, namely hagiographic models, often reappear visually, indicating a complex reuse of models. Furthermore, saints' representations also show common visual patterns. A thorough description of the phenomena through semantic web tools would allow us to computationally trace the influences of visual and textual models in monastic culture. The current work examines how existing ontologies for art description, specifically ICON, can represent these aspects. Using St. Romuald's life as a case study, the findings show that while ICON covers many necessary features, additional extensions are needed to depict hagiographic models and the specific event typologies they illustrate. Future work will formalise these extensions and enrich them through literature that defines hagiographic models.
Keywords: Iconography; Hagiographic studies; Monastic culture; Semantic Web; Ontologies

ABSTRACT (ITALIANO)

Descrivere l'iconografia monastica attraverso dati semantici: un'indagine preliminare

Recenti sviluppi nel campo delle Digital Humanities riguardano l'analisi computazionale delle opere d'arte visiva e delle iconografie. Tuttavia, alcuni filoni specialistici della ricerca storico-artistica richiedono ulteriori approfondimenti. Tra questi, alcuni studi sull'iconografia di ambito monastico si concentrano sul collegamento tra testi agiografici (narrazioni sulla vita dei santi) e la loro rappresentazione visiva. In questo contesto, i momenti chiave della vita del santo descritti nei testi, noti come modelli agiografici, spesso ricompaiono visivamente, così che l'intricato riutilizzo di modelli presenti nella tradizione testuale agiografica di diversi santi si riflette nell'iconografia. Inoltre, ulteriori relazioni sono fornite dalla ricorrenza di prototipi visivi. Il presente lavoro esamina come le ontologie esistenti per la descrizione storico-artistica, in particolare ICON, possano rappresentare questi aspetti. Utilizzando la vita di San Romualdo come caso di studio, i risultati mostrano che, mentre ICON può descrivere la maggior parte delle caratteristiche del dominio, sono necessarie ulteriori estensioni per rappresentare i modelli agiografici e le specifiche tipologie di eventi che illustrano. Futuri sviluppi della presente indagine includeranno la formalizzazione di queste estensioni e l'arricchimento attraverso la letteratura che definisce i modelli agiografici.
Parole chiave: Iconografia; Studi agiografici; Monachesimo; Web semantico; Ontologie

1. INTRODUCTION[1]

In recent years, significant developments have occurred in the Digital Humanities field concerning the computational analysis of artwork iconography, as shown by the long-standing relevance of controlled vocabularies and by the more recent advancements in ontologies, Knowledge graphs, databases, and studies in computer vision for iconography detection.

Nevertheless, these resources often struggle to meet the requirements of art historians' specialised research. Among them, the studies in monastic iconography hinder a complexity going beyond the identification of a depicted saint in a visual work (e.g., a bearded man with the pastoral staff, chalice, book, and crow shows St. Benedict). Latin Hagiographies, namely, the descriptions of saints' lives, have a complex textual tradition in which narrative models of monastic life, traceable back to the Eastern world, shaped not only the dictates of the various monastic orders but also the visual depiction of the saints (Delcorno, 1991). Formally expressing such aspects in a semantic model would advance art-historical research by recognising mutual influences between common events characterising the saints' *vitae* and the variants in their visual representation.

[1] The authors contributed to the paper as it follows. Francesco is responsible for Sections 2 and 5.1. Sofia Baroncini is responsible for Sections 1, 3 and 7. Both authors are responsible for Sections 4, 5.2, and 6.

To the best of the authors' knowledge, no studies related to the hagiographic domain explore it through the lens of the semantic web. The current study aims to address the formalization of the hagiographic domain to support a structured methodological approach. In this paper, we examine whether existing ontologies allow the representation of the field's core questions to identify the existing gaps and propose a first modelling attempt. Therefore, the research questions we address are: 1) Can semantic web technologies aid art historians in detecting and examining the parallels existing between a saint's iconography and the narrative *topoi* present in hagiographical sources? 2) Do existing tools allow us to represent the core characteristics of hagiographical research? 3) How can we integrate current ontologies to better cover domain characteristics?
The study focuses on the case study of the depiction of St. Romuald, the founder of the Camaldolese order, and his relations with Pier Damiani's *Vita Beati Romualdi* (1042), a saint's hagiographical source presenting some narrative *topoi (*from *Vitae Patrum* literature*)*. This focus was chosen because the life of Saint Romuald, a work of significant literary importance, exemplifies the role of hagiography in shaping monastic ideals based on earlier (eastern) models. After an examination of the literature, we selected for the description of the case study the ICON ontology (Sartini et al., 2023), as it offers the highest level of detail and is domain-specific.

2. RELATED WORKS

2.1. Digital humanities and iconography

The growing interest in saints' iconography has led to projects focused on creating machine-readable data for analyzing it through various approaches[2] (Impett, 2020; Milani & Fraternali, 2021).
The digital edition of the "Index Lexicon der Christlichen Ikonographie" (LCI)[3] focus on Christian iconography, especially on saints. The Iconclass classification system[4] standardizes and describes (western) visual content through hierarchical notation. Similarly, Getty's Iconography Authority[5] enhances classification by defining relationships between iconographic narratives, allowing users to specify sources and facilitating accurate analysis.
While these tools help scholars describe artwork content, they fall short in addressing more complex relationships between parts of the iconography and reference sources. These tasks can be better described through ontologies. The VIR ontology (Carboni & de Luca, 2019) extends the CIDOC-CRM (CIDOC Conceptual Reference Model) (Bekiari et al., 2021) to support the recording of relationships and structural units in visual representations. It effectively describes connections between iconographical objects, their attributes, and prototypical objects in a visual source. The ICON ontology (Sartini et al., 2023), developed from Panofsky's theoretical framework, enables a semantic detailed description of artworks subjects and meanings, along with information about (textual) sources and the provenance of the interpretation.
Among knowledge graphs, IICONGRAPH (Sartini, 2024) describes through ICON art-related statements re-engineered from ArCo (Veninata, 2020) and Wikidata (Vrandečić & Krötzsch, 2014). However, the limited amount of information it includes concerning the lives of saints restricts its effectiveness for research about saints' iconography. The Iconology Dataset (Baroncini, 2024), which describes through ICON a selection of interpretations by Erwin Panofsky, contains some granular descriptions of visual representations of Saint's iconography. Relevant is also HyperReal (Sartini et. al., 2021), a knowledge graph describing symbols and symbolic meanings that includes descriptions of symbols of the Western Christian tradition.
The models and vocabularies discussed are valuable resources for classifying and describing the iconographic content of artworks related to saints. However, the role of texts and narrative models in the interpretation process (e.g., hagiographic text needed for understanding iconographic variants) is not addressed in detail.

2.2. Monastic Hagiography and Iconography

Research on monastic sanctity holds a significant place within the field of hagiographical studies (defined here as writings about saints and their cults; Chiapparini & Rochini, 2022; Leonardi, 2011; Rigo et al., 2018). In translating, adapting, and reinterpreting texts from Late Antiquity to the Middle Ages, various

[2] The various approaches involved include, but are not limited to: deep learning, network analysis, automatic recognition. https://datavirgo.huma-num.fr/a_propos.html (last consulted: 19/01/2025).

[3] https://brill.com/display/db/lcio (last consulted: 19/01/2025).

[4] https://iconclass.org/ (last consulted: 19/01/2025).

[5] https://www.getty.edu/research/tools/vocabularies/index.html (last consulted: 19/01/2025).

topoi related to monastic asceticism, which originated in Egypt and Syria before entering the medieval Latin world, were used by Latin authors to define new models of monasticism (Constantinou & Høgel, 2020; Wiśniewski, 2020).
Writers shaped these themes to meet their time's narrative, doctrinal, social, and devotional needs. This is especially evident in hermit holiness, which experienced a revival in the West within the 11th and 13th centuries (Wiśniewski, 2020), well exemplified by the figure and hagiography of Saint Romuald (Bannister, 2011). The concept of hagiographic *topoi* is the central theme of Thomas Pratsch's extensive study (Pratsch, 2012). This work aims to provide resources for identifying historical information in Byzantine hagiography and serves as a reference for recognizing these *topoi*, providing a systematic framework useful for classifying and interpreting Byzantine subjects in visual representations.
The collective work "Agiographie und Kunst" (Kerscher, 1993) highlights the importance of narrative and conceptual frameworks derived from hagiography and formalised in *topoi*, which serve as "interpretative models" (Kerscher, 1993) that can be also used as a lens for analysing visual representations.
The current research focuses on the sole visual representation of *hagiographic models* in saint's iconography, intending to understand how the textual tradition affects the visual arts. Although scholars use various terms to describe the *topoi* that characterise much of the hagiographic narrative (Pratsch, 2012; Kerscher, 1993), we adopt the term "hagiographic model" based on its prevalent usage in the literature (Del Corno, 1988; Degl'Innocenti, 1984; Leonardi, 1991; Petrus Damiani, 1957).

3. THE ICON ONTOLOGY

The ICON ontology describes interpretations of artwork content and meaning on three levels of understanding, from a more superficial one, in which simple objects, actions, and emotions can be recognized, to a deeper one, in which deep concepts and meanings are identified (e.g., the message of "wedding wish" that the artwork is supposed to give). Each subject is recognized through an icon:Recognition, and the overall interpretation is given by a collection of Recognitions (icon:InterpretationDescription). A key aspect of the ontology is that each recognition identifies the specific subject as manifested in the artwork (e.g., Venus with floating hair as it appears in Botticelli's painting *The birth of Venus*), identified by the class icon:VisualSubject and its subclasses at each level (icon:ArtisticMotif, icon:Image, icon:IntrinsicMeaning), which is then related to generic subjects, as they can be described by vocabularies (e.g., the ICONCLASS code for Venus), identified at each level by further subclasses (*Action, NaturalElement, ExpressionalMeaning* for level 1; *Character, Event, NamedObject, Personification, Symbol, Story, Allegory* at level 2; *CulturalPhenomenon* and *SocialObject* at level 3).
Relations among the Visual Subjects allow us to describe their appearance and meaning as they occur in the artwork analysed. For example, a Composition collecting the Artistic Motifs associated with the concepts *woman, fluttering hair, shell,* and *standing* (lev. 1), can be related to an Image (Venus as it appears in the painting), related to the icon:Character "Venus" (iconclass:92C4) and to the icon:Story "birth of Venus" (iconclass:92C411). In this way, it is possible to retrieve the specific appearance of the subject in the painting, knowing the pose, qualities, and, if present, further meanings of Venus, and, at the same time, retrieve all the artworks presenting the general subjects (e.g., the Character *Venus*, the Action *standing*, etc.). Furthermore, the reuse of relations from the CiTO ontology (Shotton, 2010) allows expressing the (textual) source used by the interpreter for making the recognition[6].

4. APPROACH

To the aim of verifying whether the available vocabularies and ontologies can be reused to describe the characteristics of hagiographical studies, we first define the requirements of the intended domain. Secondly, we verify whether the ICON ontology can fulfill such requirements through the aid of the description of the representative case study of St. Romuald. Finally, we provide some suggestions for possible extensions.
From examining the characteristics of the hagiographic models identified in the literature, and with the aid of the chosen case study, an (ontological) model for the study of iconography through the lenses of hagiography should be able to represent: 1) the hagiographic model (e.g., "Conversion") visually rendered in a work of art, 2) the relation with the source, 3) visual variants of the same model in different saints' *vitae*, 4) details of the scene that are meaningful for the narration (e.g., the presence of "Sign of

[6] Further details about the ontology are provided at: https://w3id.org/icon/docs/

sanctity"). We evaluate the coverage of these aspects by the ICON ontology, chosen as it offers the highest level of detail and is domain-specific.
As mentioned in section 2.2, reinterpreting the Lives of the first monk saints of Late Antiquity, particularly in (hermitic) sanctity narratives in the Latin West, involves reusing and adapting narrative modules. This is especially evident in hermit asceticism, which experienced a revival in the West during the 11th and 12th centuries (Dijk & Dijkstra, 2006). As the life of St. Romuald exemplifies such characteristics, it is chosen as a case study in this contribution. Therefore, we select one representative iconography depicting a meaningful moment of St. Romuald's *vitae* as a test case.

5. CASE STUDY

5.1 Description

The chosen case study is a print depicting St. Romuald's conversion. It illustrates the second chapter of Pier Damiani's *Vita Beati Romualdi* (Petrus Damiani, 1957), written in 1043, and is part of the printed series "Vita et Acta:SS:mi Patris et Patriarcha Romvaldi (...)" published in Vienna in 1726[7]. This case study represents the hagiographic model of conversion, a common theme in saints' lives (see the *Vitae Patrum* and the *Historia Monachorum;* Del Corno, 1991; Rufinus & Schulz-Flügel, 1990). The scene is set in the monastery of Sant'Apollinare in Classe, where Romualdo decides to become a monk after witnessing an apparition of the martyr saint. The print shows Romuald and another monk kneeling before Saint Apollinaris, who emerges from under the altar dedicated to the Virgin Mary and incenses the church with a thurible.
An exemplary parallel of the same hagiographic model is the conversion of St. Anthony. The iconography depicted in a 1372 fresco by the Florentine artist Niccolò di Tommaso and others, located in Pistoia at Museo Marino Marini (Fig. 2), illustrates the episode in which the saint experiences his conversion while listening to the sacred scriptures in a small church (Athanasius Alexandrinus, 1974). Above the altar, a triptych featuring the Virgin Mary in the central compartment.
The two art pieces depicting the conversions of St. Romuald and St. Anthony share common visual elements, specifically the setting and the representation of the Virgin Mary as both a statue and a painting. These different portrayals highlight our approach's potential in addressing research questions regarding typological patterns related to the connection between hagiography and iconography.



Figure 1. St.Apollinaris appears to St.Romuald (detail)[8]

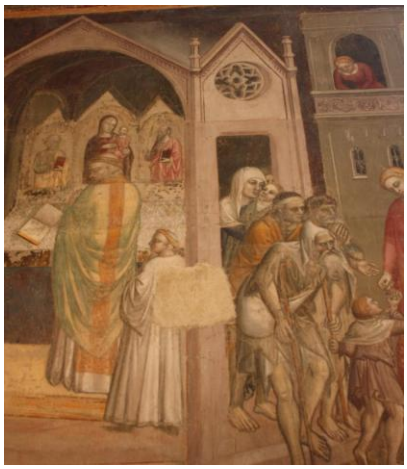


Figure 2. St. Anthony Listen to the Gospel passage[9]

5.2 Formal representation

In this paper, we provide a formal representation of the St. Romuald case study, focusing on the description of the main entity within the iconographic subjects with the ICON ontology (see Section 3). For the sake of brevity, we listed the items that could be described in the case study in Table 1.

[7] Vita et Acta:SS:mi Patris et Patriarcha Romvaldi (...), Vienna, 1726, Biblioteca Classense, 40.5.N.

[8] Courtesy of the Biblioteca Classense.

[9] Image 2 published under a, CC BY 3.0 license (author:Sailko), available at: https://commons.wikimedia.org/wiki/File:Niccol%C3%B2_di_tommaso_e_antonio_vite,_affreschi_dell%27oratorio_del_tau,_post_1372,_storie_di_s._antonio_01.jpg#metadata

	Class/relation		Instances
	Specific subject	**General subject**	
Lev 1	Artistic Motif or Composition	Natural Element	(1) kneeling man with joined hands; (2) church; (3) statue; (4) bearded man;(5) thurible; (6) kneeling monk with joined hands
		Action	(7) incensing the church;(8) pray while kneeling
		Quality	(1) youthful age
Lev 2	Image	Character	St. Romuald (1), Iconclass:11H(ROMUALD), St.Apollinaris (4)
		Event	St. Apollinaris appears to St.Romuald
		Place	St.Apollinaris Church (2)
		Named Object	Virgin altarpiece (3)
		Story	Story of St.Apollinaris who appears to St.Romuald
		cites for information	Vita Beati Romualdi, ch. 2

Table 1. Main entities relevant in the description of St.Apollinaris appears to St.Romuald

As we see in Table 1 and Figure 1, the majority of the identified traits of St. Romuald's depiction can be described with ICON ontology. Among the various significant iconographic entities, the class icon:Place allows the description of the location where the event occurs, and the class icon:NamedObject describes the presence of the simulacrum of the Virgin as a meaningful detail of the narrative. Furthermore, the property cito:citesForInformation enables the linkage of the depicted scene to its textual source of reference. At the first level, the class icon:Action provides for the description of the actions of specific characters as a visual counterpart to the iconographic representation of St. Anthony. The numbers associated with the instances indicate the relationships among the specific subjects (i.e., Images, Artistic Motifs, Compositions) recognized at the levels of the description. The artwork depicting St. Anthony can be described in a similar way,[10] highlighting the presence of the setting (a church) and Mary's altarpiece, features that are in common with the previous artwork.
Nevertheless, the ICON ontology presents some limits when representing core features of hagiography. Whereas the class icon:Event allows the description of every event, therefore including moments of the saint's life, a hagiographical analysis would require that such events are recognised as the hagiographical model shaped in the textual sources. To this end, the ICON ontology may be extended to represent such specific events depicting the saint's life, introducing the class *Hagiographic Event* as a subclass of icon:Event. This class would represent the iconographical rendering of the narrative *topoi* described in texts, allowing the identification of the influence of literary structures in the visual sources. Then, the class can be further specified with the definition of shared, recurrent moments in the saint's *vitae* (e.g., conversion, temptation). In the provided example, the *Hagiographic Event* enables us to associate this *topos* with certain common iconographic attributes and arrangements of the depicted scenes by relating the recognition of the Image (*ART1-IMG6*) to the ArtisticMotifs or Compositions representing the other subjects (e.g., a kneeling man) through the relation *icon:refersToArtisticMotif*, or other subjects grouped in the same Story (e.g., the Place *St. Apollinaris Church*). Specifically, the comparison pertaining to the conversion encompasses the saint's youthful age, the setting of a church, and the presence of the Virgin Mary (in both statue and painting) as one of the means of conversion. The separation between the generic and the artwork's specific subject provided by the ICON ontology allows a thorough description of the subjects depicted and their relations in a specific artwork (e.g., the kneeling man with joined hands is recognised at level 2 as the Character St. Romuald) and, at the same time, the retrieval of all the artworks presenting the subject (e.g., a query retrieving all the artworks in which the *Hagiographic Event* "Conversion of St. Romuald" can retrieve all the artworks, for which then more specific details about the Conversion are provided). Furthermore, the extension of the *Hagiographic Event* class with further specialised moments (e.g., Conversion, see Fig. 3) would allow an even broader retrieval of all the saints depicted in the conversion moment, allowing for a detailed comparison.

[10] We don't provide here a full description of this artwork for the sake of brevity.

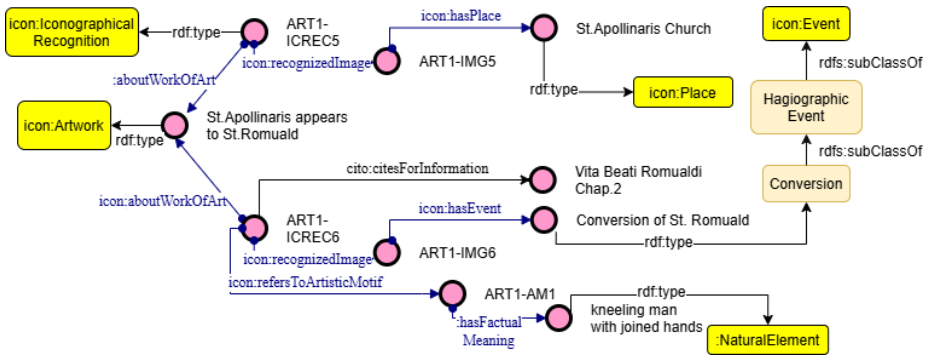


Figure 3. Section of the graphical representation of St.Apollinaris appears to St.Romuald (proposed classes in orange)

6. RESULTS

Although ICON can describe the key features of iconographic representations related to monastic themes found in hagiographic sources, it lacks means of descriptions for the hagiographic narrative models. These can serve as interpretative structures used to analyse visual representations. The process involves introducing a class, namely the *Hagiographic Event*, that corresponds to key terms (e.g., "conversion") into the artwork's description while implementing iconographic subject recognition.[11] These categories stem from common moments in saints' lives based on hagiographic narratives. The introduction of the *Icon:HagiographicEvent* class has enabled connections between meaningful elements and iconographic features related to the conversion of St. Romuald and the hagiographic *topos* (conversion), paralleling the representation of this event in a scene from the life of St. Anthony Abbot. This parallelism was selected due to the strong connections between the hagiographies of the two saints (Bannister, 2011).

7. CONCLUSIONS AND FUTURE WORKS

Narrowly interwoven with the textual tradition of hagiographies, namely, texts describing the saints' *vitae*, the iconography of (monastic) saints often depicts key moments of their life, known as *hagiographic models* (e.g., the moment of the conversion). Iconographies depicting them show not only the sharing of models, but also visual arrangements shared by art pieces depicting different saints. A detailed description of the phenomena through semantic web tools would allow us to computationally trace such influences. To this end, the current work examined the extent to which the currently available ontologies can represent the key aspects of the domain. While the ICON ontology addresses the most needed features, additional updates are required. We therefore suggested the extension of the ontology to include a class, named *HagiographicEvent*, intended to represent the hagiographic models as rendered in visual materials, namely, as events part of a saint's life. This class, further specialised in subclasses, representing the specific moments of a saint's biography, would allow the comparison of iconographic sources.

Future works include the formalisation and evaluation of the proposed extension, and its enrichment with further subclasses through the examination of literature that provides typological definitions of hagiographic models. Specifically, Pratsch's work systematically catalogues hagiographic models (*topoi*), which may be examined to extend the ontology or create a controlled vocabulary. Moreover, we consider to validate our model with empirical tests involving other hagiographic topoi and saints (e.g., the Hagiographic Event "Call to Withdrawal" linked to pictorial cycles related to anchorites stories). The extension would allow cross-referencing among a selection of representations of saint monks by computationally identifying and analysing structured data on the iconographic occurrences of specific modules.

Ultimately, this will deepen our understanding of monastic visual culture in relation to hagiographic narratives while also advancing digital art history and semantic web technologies within the field of iconography.

[11] Claims that are made by an agent (a person or a computer) about an artwork to recognize a single subject are individuals of the class Recognition. The subclasses of Recognition express each recognition conducted at a certain level of interpretation, namely Pre-Iconographical Recognition (level 1), Iconographical Recognition (level 2), and Iconological Recognition (level 3). See: https://w3id.org/icon/docs/

REFERENCES

Athanasius Alexandrinus. (1974). 1: *Vita di Antonio* / [Atanasio] ; introduzione di Christine Mohrmann; testo critico e commento a cura di G. J. M. Bartelink ; traduzione di Pietro Citati e Salvatore Lilla. Fondazione Lorenzo Valla A. Mondadori.

Bannister, E. (2011). 'From Nitria to Sitria': The construction of Peter Damian's Vita Beati Romualdi , *European Review of History: Revue europeenne d'histoire*, 18:4, 499-522. https://doi.org/10.1080/13507486.2011.590272

Baroncini, S. (2024). *Diving into artworks interpretations through the lenses of semantic data: An application on Panofsky's iconological studies* [Doctoral Thesis, Alma Mater Studiorum - Università di Bologna]. https://doi.org/10.48676/unibo/amsdottorato/11312

Bekiari, C., Bruseker G., Doerr, M., Ore,C.E., Stead, S., Velios, A. (2021). *Definition of the CIDOC Conceptual Reference Model, Version 7.1.1*. CIDOC. https://doi.org/10.26225/FDZH-X261

Carboni, N., & de Luca, L. (2019). An Ontological Approach to the Description of Visual and Iconographical Representations. *Heritage*, 2(2), 1191-1210. https://doi.org/10.3390/heritage2020078

Chiapparini, G., & Rochini, M. (2022). *Manuale di agiografia-Fonti, storia e immagini della santità*. Morcelliana. https://publires.unicatt.it/it/publications/manuale-di-agiografia-fonti-storia-e-immagini-della-santit%C3%A0

Constantinou, S., & Høgel, C. (Eds.). (2020). *Metaphrasis:A Byzantine Concept of Rewriting and Its Hagiographical Products*. Leiden, The Netherlands: Brill. https://doi.org/10.1163/9789004438453

Delcorno, C. (1988). Modelli agiografici e modelli narrativi. Tra Cavalca e Boccaccio. *Lettere Italiane*, 40(4), 486–509.

Delcorno, C. (1991). Le «Vitae Patrum» nella letteratura religiosa medievale (secc. XIII–XV). *Lettere italiane*, 43(2), 187–207.

Degl'Innocenti, A. (1984). *Le vite antiche di Giovanni Gualberto: cronologia e modelli agiografici*. Spoleto Centro italiano di studi sull'Alto Medioevo

Dijkstra, J., & van Dijk, H. (Eds.). (2006). *The Encroaching Desert: Egyptian Hagiography and the Medieval West*. Leiden, The Netherlands: Brill. https://doi.org/10.1163/9789047411628

Impett, L. (2020). Analyzing Gesture in Digital Art History | *33* | *The Routledge Companio*n to Digital Humanities and Art History (pp.386-407). https://www.taylorfrancis.com/chapters/edit/10.4324/9780429505188-33/analyzing-gesture-digital-art-history-leonardo-impett

Kerscher, G. (1993). Hagiographie und Kunst: Der Heiligenkult in Schrift, *Bild und Architektur*. D. Reimer.

Leonardi, C. (1991). Modelli agiografici nel secolo VIII: da Beda a Ugeburga. In *Les fonctions des saints dans le monde occidental (IIIe-XIIIe siècle)* Actes du colloque de Rome (27-29 octobre 1988) Rome: École Française de Rome, 1991. pp. 507-516. (Publications de l'École française de Rome, 149). www.persee.fr/doc/efr_0000-0000_1991_act_149_1_4214

Leonardi, C., & Santi, F., Degl'Innocenti, A. (2011). *Agiografie medievali*. Firenze SISMEL Edizioni del Galluzzo.

Milani, F., & Fraternali, P. (2021). A Dataset and a Convolutional Model for Iconography Classification in Paintings. *J. Comput. Cult. Herit.*, 14(4), 46:1-46:18. https://doi.org/10.1145/3458885

Petrus Damiani. (1957). *Petri Damiani Vita Beati Romualdi* / a cura di Giovanni Tabacco. Istituto storico italiano.

Pratsch, T. (2012). *Der hagiographische Topos: Griechische Heiligenviten in mittelbyzantinischer Zeit*. De Gruyter. https://doi.org/10.1515/9783110912715

Rigo, A., Trizio, M., Despotakis, E. (2018). *Byzantine hagiography: texts, themes & projects*. Turnhout Brepols.

Rufinus, T., Schulz-Flügel, E. (1990). *Historia monachorum, sive De vita sactorum patrum.* Berlin W. de Gruyter.

Sartini, B., Van Erp, M., & Gangemi, A. (2021). Marriage is a Peach and a Chalice: Modelling Cultural Symbolism on the Semantic Web. *Proceedings of the 11th Knowledge Capture Conference*, 201–208. https://doi.org/10.1145/3460210.3493552

Sartini, B., Baroncini, S., Van Erp, M., Tomasi, F., & Gangemi, A. (2023). ICON: An Ontology for Comprehensive Artistic Interpretations. *J. Comput. Cult. Herit.*, 16(3), 59:1-59:38. https://doi.org/10.1145/3594724

Sartini, B. (2024). IICONGRAPH: improved Iconographic and Iconological Statements in Knowledge Graphs. In *The Semantic Web: 21st International Conference*, ESWC 2024, Hersonissos, Crete, Greece, May 26–30, 2024, Proceedings, Part II. Springer-Verlag, Berlin, Heidelberg, 57–74. https://doi.org/10.1007/978-3-031-60635-9_4

Shotton, D. (2010). CiTO, the Citation Typing Ontology. *Journal of Biomedical Semantics*, 1, 1-18. https://doi.org/10.1186/2041-1480-1-S1-S6

Wiśniewski, R. (2020). Relate and Retell. Eastern Monastic Stories and the Beginnings of Latin Hagiography in *Metaphrasis:A Byzantine Concept of Rewriting and Its Hagiographical Products*, 63–82. https://doi.org/10.1163/9789004438453_003

Veninata, C. (2020). Dal Catalogo generale dei beni culturali al knowledge graph del patrimonio culturale italiano: il progetto ArCo. *DigItalia*, 15(2), 43–56. https://doi.org/10.36181/digitalia-00013

Vrandečić, D., & Krötzsch, M. (2014). Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10), 78-85.

A Linguistic Knowledge Graph of Word Borrowings from Portuguese

Anas Fahad Khan[1], Ana Salgado[2]
[1] CNR-ILC, Italy - fahad.khan@ilc.cnr.it
[2] NOVA CLUNL - Centro de Linguística da Universidade Nova de Lisboa, Portugal, Portugal - anacastrosalgado@gmail.com

ABSTRACT (ENGLISH)

We present ongoing work in the creation of a linguistic knowledge graph of word borrowings from Portuguese into South Asian languages. The first version of this resource, soon to be published, consists of lexicons of Hindi and Urdu words deriving from Portuguese. We start by describing our approach and the data collection process. Then we describe how different aspects of the data are represented as linked data.
Keywords: Linked Open Data; Semantic Web; Ontologies; RDF

ABSTRACT (ITALIANO)

Un Grafo di Conoscenza Linguistica sui Prestiti Lessicali dal Portoghese

Presentiamo un lavoro in corso per la creazione di un linguistic knowledge graph sulle influenze lessicali del portoghese nelle lingue dell'Asia meridionale, una risorsa che stiamo per pubblicare in una prima versione costituita da lessici di parole in hindi e urdu derivanti dal portoghese. Iniziamo descrivendo il nostro approccio e il processo di raccolta dei dati. Successivamente, spieghiamo come diversi aspetti dei dati vengono rappresentati come linked data.
Parole chiave: Linked Open Data; web semantico; ontologie; RDF

1. INTRODUCTION

Despite its being over a decade old, and despite being successfully used in important projects like Linking Latin (LiLa) (Passarotti et. al. 2020), linguistic linked open data (LLOD) is still a relatively underdeveloped and/or underutilised means of publishing linguistic data. This is unfortunate as LLOD is very well adapted for publishing numerous different kinds of linguistic dataset thanks to its graph-like structure, something which ultimately allows for the creation of extensive linguistic knowledge graphs[1] as in the case of the aforementioned LiLa project; it also has the advantage of supporting the enrichment of linguistic resources with non-linguistic resources via RDF's native linking mechanism. LLOD can, moreover, facilitate the development of language resources which better exploit LOD's technological (and other) strengths in making data more accessible/machine actionable. In this article, we highlight one kind of linguistic resource that seems particularly adapted to publication as an RDF linguistic knowledge graph: a language resource describing lexical borrowings from a single language into a number of other languages. Note that although we describe such resources as *linguistic* knowledge graphs, the kinds of information they contain are useful to researchers in a number of other disciplines too, including historians, and scholars in cultural studies. In this article we will concentrate on a particular example/use-case of a language resource of borrowings, one which the authors of the paper have been working on over the last few months, and the first version of which we have made available as RDF via Github,[2] CHAMUÇA. In particular, we will look at the different ways that RDF and the Semantic Web facilitate or afford the structured representation of various, salient aspects of our data, aspects which are typical of such data.

The rest of the paper is organised as follows. We begin by giving an overview of the project in Section 2. This is followed by a description of the data sources of our linguistic knowledge graph in Section 3. In Section 4, we describe our linguistic knowledge graph and focus on the affordances offered by the Semantic Web for representing the contents of our data.

[1] By linguistic knowledge graph we mean a knowledge graph whose principal scope is to describe linguistic phenomena.

[2] https://github.com/anasfkhan81/chamuca/tree/main/chamuca_lex_resource

1.1. TRACING THE IMPACT OF PORTUGUESE ON SOUTH ASIAN LANGUAGES: THE CASE OF HINDI AND URDU

The work described here began with a simple question: *how many words in the two South Asian languages Urdu/Hindi derive from Portuguese?* Although Portuguese does not seem, prima facie, to have had a profound impact on both languages -- certainly nowhere near as great as English has had -- there are a number of very popular words (e.g., the most common words in both languages for 'key', 'room' and 'church', all three of which are shared by both languages) which ultimately derive from the language, and which seem to evidence the great influence which Portuguese traders and colonists had had in Asia at the beginning of the Modern era. Indeed the more one looks into this question, the more it becomes clear that, leaving aside the obvious case of English, no other European language has had as great a direct impact on Urdu/Hindi as Portuguese -- something which seems to be common for many other Asian, and especially South Asian languages too (notable exceptions being South-east Asian languages such as Vietnamese which have received greater influence from French and Tagalog which was greatly influenced by Spanish). This led the authors to ask whether there were any modern electronic resources, and in particular any dictionaries, which could be consulted on this topic, and, additionally, whether such resources existed for other (South) Asian languages, in order to compare the levels of influence which Portuguese had had on nearby languages. The answer is that, although much of the relevant information *was* available distributed across different sites/resources/print volumes, in most cases it was **not** available as structured data, and was often not completely curated. There exists one well known scholarly work on this topic, 1913's *Influência do vocabulário português em línguas asiáticas* ('Influence of the Portuguese Vocabulary on Asian Languages'), a lexicon compiled by the early 20th century Goan scholar Sebastião Rodolfo Dalgado and which features Portuguese borrowings in 50 languages. However, this was over 110 years old by now and hadn't been updated.

We decided to fill the existing gap by building our own language resource, **Cultural Heritage and Multilingual Understanding through lexiCal Archives (CHAMUÇA)**, an RDF linguistic knowledge graph consisting of lexicons of Portuguese borrowings in various different Asian languages (beginning with the languages of South Asia). CHAMUÇA will be published (in different versions) with an open license, and will be based, as far as possible, on already existing resources. As of the time of writing, we have compiled the Hindi and Urdu lexicons and an indexical Portuguese lexicon, these are available as RDF graphs on Github[3] and we are planning to publish them as linked data on our triple store, as one of the use-cases of the Italian national project, H2IOSC,[4] shortly. We chose these two languages because of the linguistic competences of the authors as well as their popularity as languages on the subcontinent, and the interesting status of the two in relation to one another (many linguists regard them as two different registers of the same language). However, we are in discussion with experts of languages such as Bengali, and Sinhala for the next updated version of our graph.

2. DATA SOURCES

We derived an initial list of over a hundred entries for Hindi and Urdu on the basis both of Dalgado's monumental study as well as from Wiktionary and Wikipedia; these entries were cross-checked and enriched with respect to a number of several different lexical sources (including those made available by the Digital Dictionaries of South Asia[5] platform). Each entry in our lexicons includes basic lexical information such as *lemma*, *gender, part of speech*, and other morphological variants of the lemma (nouns in Urdu/Hindi have 6 morphological variants, Portuguese has 2; however, in the rare cases of verbs/adjectives we have come across we haven't listed all their forms); there is also link from each entry to its supposed Portuguese etymon (as well as information on other potential etymologies, taken from different sources). Note that Dalgado uses his own Roman alphabet transliterations for all of the languages featured in his lexicon. In many cases, it was easy to find Hindi/Urdu lemmas in their Devanagari/Nastaliq (respectively) forms; in a handful of cases it was much less straightforward and

[3] https://github.com/anasfkhan81/chamuca

[4] https://www.h2iosc.cnr.it/

[5] https://dsal.uchicago.edu/dictionaries/

required consulting and comparing a number of lexicographic sources. One of the authors of the current work is chief editor for the dictionaries of the Academia das Ciências de Lisboa and was able to contribute with resources on Portuguese etymons. The result of this data collection phase was a fairly comprehensive list of around 100 Hindi lemmas, circa 80 Urdu lemmas and around 50 Portuguese lemmas with relevant lexical/linguistic information.

3. BUILDING A LINGUISTIC KNOWLEDGE GRAPH OF PORTUGUESE BORROWINGS

Our idea, to reiterate, was to bring together and curate already existing information, enrich it (especially with reference to more recent scholarship) and make it available as linked data, namely, as an RDF linguistic knowledge graph; and in terms of the latter, the more (suitable/relevant) links there are between the different nodes in our data, the better. In what follows, we look at how we converted our original data (stored as TSV files) into linked data and how we tried to take advantage of the different expressive and technological possibilities offered by the Semantic Web. The modelling of basic linguistic information via the OntoLex vocabulary[6] together with the lexinfo ontology[7] (the latter acting as a Data Category Registry for the former) was fairly straightforward.[8] As an example in Figure 1 we present part of an entry from the CHAMUÇA Hindi lexicon, namely the entry for अलमारी [almari] meaning 'cupboard', a word ultimately derived from the Portuguese *armario*, with some of this basic information shown.

```
:अलमारी_entry a ontolex:LexicalEntry;
      lexinfo:etymologicalRoot pt_lex:armário ;
      lexinfo:gender lexinfo:feminine ;
      lexinfo:partOfSpeech lexinfo:commonNoun ;
      ontolex:canonicalForm :अलमारी_lemma ;
      ontolex:lexicalForm :अलमारियो_vp_form_अलमारी,
         :अलमारियों_op_form_अलमारी,
         :अलमारी_dp_form_अलमारी,
         :अलमारी_os_form_अलमारी,
         :अलमारी_vs_form_अलमारी ;
```

Figure 1. Part of the RDF encoding for the entry अलमारी

Here, one can see reference to the gender, and part of speech of the word, a link to the resource representing the lemma and the word's other morphological variants together with a link to the sense of the word; this is the minimum standard information for all the words in our Portuguese, Hindi and Urdu lexicons. There is also a link to the proposed Portuguese etymon which belongs to the Portuguese lexicon *pt_lex* and a link to the corresponding Urdu entry via the RDFS *seeAlso* property.

Our data focuses on describing a network of relationships between words of different languages; the graph-based nature of RDF makes it a natural fit for this (relationships between words viewed as nodes). The LiLa project, cited above, was an important step in revealing the potential of linked data for creating rich linguistic knowledge graphs (LKG) on the basis of a specific conceptual architecture. In the case of LiLa this is based on the existence of a hub component of the LKG, *the lemma bank* (Mambrini and Passarotti, 2023). In our case we have a slightly different architecture. Our entire dataset is a Lexical Resource,[9] that is a container for a number of lexicons, each one of which contains borrowed words from

[6] OntoLex or OntoLex-Lemon is a popular RDF based lexicon for encoding lexical resources, https://www.w3.org/2016/05/ontolex/

[7] https://lexinfo.net/

[8] However, as we begin to look at other languages, especially non-Indo European ones, it may prove somewhat more challenging, since OntoLex has up till now been used to model a fairly limited selection of languages.

[9] We take the definition of Lexical Resource used in for instance LMF (Francopoulo, 2013).

Portuguese, and an indexical Portuguese lexicon which acts as a hub. We would like to propose this as a sort of basic pattern[10] for similar work in lexicons of borrowings or in language resources for contact linguistics. The use of these and similar patterns would offer an extra level of interoperability in addition to the use of the RDF data model and shared vocabularies and ontologies.

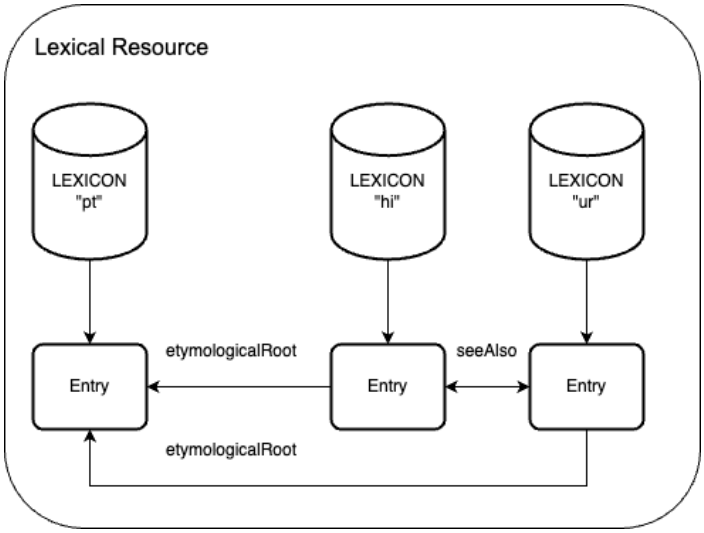


Figure 2. The overall architecture of our resource (we plan to expand this to add more languages)

We decided to focus on exploiting the semantic affordances made available by the Semantic Web to ensure that our LKG would be easier to navigate, and to enable it to answer a wider variety of questions. For instance, we use the popular Semantic Web knowledge organisation vocabulary SKOS[11] to encode a list of basic domains as concepts, and to annotate different entries in our CHAMUÇA lexicons for the domains with which they are associated (using the lexinfo *domain* property). We ended up with 24 domains (this number will undoubtedly expand as we include other languages): *administration, animal husbandry, architecture, botany, calendar, cleaning and hygiene, clothing, culinary, furniture, healthcare, household, materials, metrology, military, music, nationalities, nautics, occupations, proper names, religion, sensory perception, sewing, trade, transportation*. For instance, in our Hindi/Urdu lexicons we have several examples of entries which are related to the Christian religion, these include **गिरजा/**گِرْجا (‘church’), **पादरी**پادری (‘priest’) and **क्रूस**/ (‘cross’); each of these entries is associated with the domain *religion* (the latter represented as a SKOS concept). We followed the best practices suggested by (Khan et al, 2023) for adding domain labels to linked data lexicons here.

Finally, we have enriched our KG with etymological information. Quite a few of the words in our Hindi/Urdu lexicon have alternative etymologies (i.e., without Portuguese etymons), and in some cases the Portuguese origin of the word is quite doubtful. Etymologies are far from simple to represent in RDF, however, since each etymology is effectively an individual word history and there is no agreed upon best practice for how to represent even very simple narratives as RDF graphs. There have been proposals for encoding this in a ‘deeper’ way, in the sense of attempting to explicitly represent as much of the relevant information contained in an original textual version of the etymology as possible as structured data in RDF; we decided, however, that we would give the etymology as a string in a standard form. These etymologies, each of which is associated with an individual entry via the lexinfo:etymology property, are derived from a number of sources, including Dalgado and (McGregor, 1993), provenance information which we mark explicitly in our string etymologies. For instance, in the case of the Urdu word مرمر [marmar] (‘marble’) we have two contrasting etymologies, one from Portuguese that was proposed by Dalgado and another from Persian, suggested both by McGregor and Wiktionary; these alternative etymologies are represented in the string etymology divided by the ‘;’. We use the same standard format for representing all of our etymologies as strings, with each etymology divided from its predecessor by the

[10] In this we are inspired by Ontology Design Patterns (Gangemi & Presutti, 2009), but we would like to apply this concept to LLOD.

[11] https://www.w3.org/2004/02/skos/

semi-colon and the source given in abbreviated form between brackets, i.e., "etymology 1 (source: X1) ; etymology 2 (source: X2, X2')....; etymology k (source: XK,...,XK')".

```
:مرمر_entry a ontolex:LexicalEntry;
    lexinfo:etymologicalRoot pt_lex:mármore ;
    lexinfo:etymology
           "mármore (Source: Dalgado); Persian
     مرمر  (Source: Wiktionary, McGregor)" .
```

Figure 3. The etymological part of a the entry for مرمر

4. CONCLUSIONS

In this article we have looked at some of the features of the linguistic Knowledge Graph we have been creating. As we have mentioned above, the Urdu/Hindi lexicons are almost complete and can be consulted on Github.[12] This first version is shortly due to be made available via a public SPARQL endpoint as one of the use cases of the H2IOSC project, and deposited in a CLARIN repository. Looking ahead, we are working on a next version of CHAMUÇA that will incorporate borrowings from Portuguese into other South Asian languages, including Bengali and Sinhala. We are actively collaborating with linguistic experts to ensure the accuracy and relevance of the newly integrated data. In addition, we aim to refine our methods for representing etymological uncertainty and to improve the alignment of CHAMUÇA with other linguistic knowledge graphs, thereby enhancing its interoperability within the broader LOD ecosystem.

ACKNOWLEDGEMENTS

This work is partly supported by the H2IOSC Project - Humanities and cultural Heritage Italian Open Science Cloud funded by the European Union NextGenerationEU - National Recovery and Resilience Plan (NRRP) - Mission 4 "Education and Research" Component 2 "From research to business" Investment 3.1 "Fund for the realization of an integrated system of research and innovation infrastructures" Action 3.1.1 "Creation of new research infrastructures strengthening of existing ones and their networking for Scientific Excellence under Horizon Europe" - Project code IR0000029 - CUP B63C22000730005. Implementing Entity CNR.

REFERENCES

Dalgado, S. R. (1913). Influência do vocabulário português em línguas asiáticas. Coimbra.

Francopoulo, G. (Ed.). (2013). LMF: Lexical markup framework. Wiley.

Gangemi, A., & Presutti, V. (2009). Ontology design patterns. In S. Staab & R. Studer (Eds.), Handbook on ontologies (pp. 221–243). Springer.

Khan, F., Salgado, A., Costa, R., Ramos, M., Carvalho, S., Silva, R., & Almeida, B. (2023). Encoding domain labels in RDF: Guidelines [GitHub repository]. GitHub. https://github.com/anasfkhan81/EncodingDomainLabelsRDF/blob/main/Guidelines.md

Mambrini, F., & Passarotti, M. C. (2023). The LiLa Lemma Bank: A knowledge base of Latin canonical forms. Journal of Open Humanities Data, 9, 28. https://doi.org/10.5334/johd.145

McGregor, R. S. (1993). The Oxford Hindi-English dictionary. Oxford University Press.

Passarotti, M., Mambrini, F., Franzini, G., Cecchini, F. M., Litta, E., Moretti, G., Ruffolo, P., & Sprugnoli, R. (2020). Interlinking through lemmas: The lexical collection of the LiLa knowledge base of linguistic resources for Latin. Studi e Saggi Linguistici, 58(1). https://doi.org/10.4454/ssl.v58i1.277

[12] https://github.com/anasfkhan81/chamuca/tree/main/chamuca_lex_resource

E.T and Visual Culture Ontology (ETVCO): Perspectives on Extraterrestrial Influence in Visual Heritage

Kaosaier Wusiman[1], Simone Casazza[2]
[1] University of Bologna, Italy – kaosaier.wusiman@studio.unibo.it
[2] University of Bologna, Italy – simone.casazza@studio.unibo.it

ABSTRACT (ENGLISH)
The ETVCO ontology models visual representations related to Ancient Astronaut Hypothesis, integrating cognitive and visual dimensions to analyze interpretations of artifacts and artworks. By leveraging the Cognitive Perspectivisation Pattern (CP) and the ICON ontology, it highlights biases and alternative perspectives. Following the Extreme Design (XD) methodology, the ontology captures diverse interpretations by incorporating Perception Recognition, Iconographic Levels, and Cognitive Perspectivisation. This contribution presents an ontology designed to represent the interactive relationship between visual imagery and cognitive perspectives. The evaluation confirms its effectiveness for the study of visual heritage, support for interdisciplinary research, and the analysis of interpretive biases. Future work will focus on enhancing cognitive modeling and expanding the dataset.
Keywords: Visual Heritage; Ontologies; RDF

ABSTRACT (ITALIANO)
ETVCO – Extraterrestri e Cultura Visiva: un'ontologia delle teorie alternative sul patrimonio artistico
L'ontologia ETVCO modella rappresentazioni visive relative alla teoria degli antichi astronauti, integrando aspetti cognitivi e visivi per analizzare le interpretazioni di artefatti e opere d'arte. Utilizzando il Cognitive Perspectivisation Pattern e l'ontologia ICON, l'ontologia evidenzia bias e prospettive alternative. Seguendo la metodologia Extreme Design (XD), l'ontologia modellizza le varie interpretazioni sfruttando il riconoscimento della percezione, i livelli iconografici e la prospettivizzazione cognitiva. Nel presente contributo si fornisce un'ontologia per la relazione interattiva tra immagini visive e prospettive cognitive. L'evaluation conferma la sua efficacia per lo studio delle fonti iconografiche, il supporto alla ricerca interdisciplinare e l'analisi dei bias interpretativi. In futuro, ci si concentrerà maggiormente sulla modellazione cognitiva e sull'espansione del dataset.
Parole chiave: Patrimonio visivo; Ontologie; RDF

1. INTRODUCTION
The E.T. and Visual Culture Ontology (ETVCO)[1] project aims to develop a structured framework for modeling and analyzing visual representations associated with the Ancient Astronaut Hypothesis. This hypothesis posits that extraterrestrial beings visited earth in the distant past and influenced early human civilizations, leaving traces in their cultural artifacts and artistic expressions (Rayhan, Abu, 2023). The ontology focuses on visual artifacts from ancient, medieval, and early modern periods, examining diverse interpretations by archaeologists, art historians, and conspiracy theorists. It integrates both visual and cognitive dimensions, capturing the formal characteristics of artworks (e.g., shapes, motifs, and compositions) and the cognitive perspectivization processes underlying these interpretations.
To achieve this, the project leverages well-established ontology design patterns like the Cognitive Perspectivisation Pattern (Gangemi & Presutti, 2022: 208-228) and integrates ontologies such as ICON (Sartini, Baroncini, van Erp, Tomasi & Gangemi, 2022: 1-38). In addition, the project adopts the Extreme Design (XD) methodology (Presutti, Daga, Gangemi & Blomqvist, 2009) to ensure a user-focused and iterative approach, incorporating requirement stories and competency questions tailored to stakeholders (ChatGPT has been used as a fictional stakeholder).
The ontology provides a solid foundation for documenting and analyzing visual heritage, focusing on the cognitive mechanisms that influence artifact interpretation from diverse perspectives. By integrating cognitive perspectivization processes, the tool enables the construction of alternative interpretations of visual motifs, shaped by the background knowledge of interpreters. This includes the varying attitudes of different conceptualizers, which result in divergent interpretations of the same visual themes.
Furthermore, the ontology supports semantic queries to explore how these interpretations differ from

[1] https://raw.githubusercontent.com/knowledge-representation-2024/E.T-and-visual-culture-ontology-perspectives-on-Extraterrestrial-influence-in-visual-heritage-ETVCO-/refs/heads/main/etvcontology.ttl#

mainstream perspectives, fostering interdisciplinary research into the cognitive frameworks and cultural contexts shaping these narratives. Ultimately, it reveals the interpretative biases underlying different perspectives.

2. STATE OF THE ART

Several ontologies have been developed to address image-related entities. For example, CIDOC-CRM enables the exchange and integration of information from heterogeneous sources, supporting the reconstruction and interpretation of the past at a human scale, based on all kinds of material evidence (Bekiari, Bruseker, Canning, Doerr, Michon, Ore, Stead, & Velios, 2024). It includes classes such as E22 Human-Made Object and E4 Period.

The Icon ontology, based on Panofsky's theory of iconography (Sartini, Baroncini, van Erp, Tomasi & Gangemi, 2022: 1-38), adopts a layered interpretative model, howbeit it focuses primarily on the pre-iconographic level, and incorporates less cognitive patterns. In contrast, ETVCO make also use of the CP pattern and adds a perception-recognition layer to capture multiple paths of interpretation of the same visual motif (e.g., a circular shape being interpreted as an UFO). This approach enhances the ontology's capacity to represent interpretive diversity and the dynamic interaction between visual imagery and cognitive processing.

3. METHODOLOGY

To ensure a user-focused and iterative approach, we adopted the Extreme Design (XD) methodology, incorporating user stories and competency questions (CQs) to address the needs of stakeholders like archaeologists, art historians, and cognitive scientists.

Firstly, we conducted extensive research on visual representations linked to the Ancient Astronaut Hypothesis, including sculptures, petroglyphs, and paintings. We then synthesized insights from academic papers, conspiracy theorist forums, and alternative narratives to encapsulate both mainstream and non-mainstream perspectives. To conclude the domain study, recurring visual motifs were analyzed, such as humanoid figures and celestial symbols, often interpreted as evidence of extraterrestrial contact.

In alignment with the subsequent steps of the Extreme Design (XD) methodology, ChatGPT was employed as a fictional stakeholder to simulate the needs of different users, including art historians, archeologists, and conspiracy theorists. Through iterative sessions, we formulated user stories that reflected practical use cases, such as the interpretation of artifacts with anomalous features or the exploration of intersections between conspiracy theories and cultural heritage. Additionally, we developed testable competency questions (CQs) to inform the ontology's structure and properties, including: "Which artifacts are interpreted as humanoid figures with exaggerated proportions?" and "How are celestial symbols in medieval manuscripts reinterpreted by different groups?".

Due to constraints in time and resources, ChatGPT was employed as a fictional stakeholder in the ontology development process. Despite this workaround, it proved to be a valuable tool for accelerating the design phase. Large Language Models (LLMs), when prompted appropriately, can effectively simulate fictional end users, offering a promising means of eliciting and understanding user requirements. Nevertheless, this approach presents certain limitations. In particular, the quality of the generated user stories is highly dependent on the structure and specificity of the prompts provided.

In our case study, we instructed the LLM to assume the roles of various domain experts, such as art historians, archaeologists, and cognitive scientists. Subsequently, it was asked to generate user stories and associated competency questions tailored to each role, following the template outlined in the Extreme Design Methodology (Presutti, Daga, Gangemi, & Blomqvist, n.d.). Selected portions of the output were then manually refined to better align with the anticipated needs of professional end users.

While the use of LLMs in this context shows promise, it remains an emerging area that requires further refinement. For instance, we did not implement a systematic prompting strategy, nor was the model fine-tuned with domain-specific documentation, both of which could potentially enhance the relevance and quality of the results.

In line with established integration practices, we reused the ICON ontology to model artistic interpretation and introduced a new *Perception Recognition* layer to represent the initial stages of visual recognition. This layer captures perceptual features, such as shape, size, color, and position, prior to any cultural or symbolic interpretation. Additionally, by incorporating the Cognitive Perspectivisation (CP) Pattern, we modeled agent-specific perspectives and biases, allowing for the analysis of how cultural and cognitive factors shape interpretation processes. The ETVCO core components include:

- Perception Recognition class: Captures basic perceptual elements before interpretive processing occurs.
- Pre-iconographical and Iconographical Levels: Aligned with the first and second levels of the ICON ontology to maintain conceptual continuity (see Fig. 1).
- Cognitive Perspectivisation: Models diverse agent perspectives, integrating concepts such as conceptualizer attitudes, interpretive lenses, and inherent biases.

By integrating these elements, the ontology (see Fig. 2), enables nuanced analyses of how varying agents perceive and interpret visual motifs, revealing underlying cultural influences and interpretative activities.

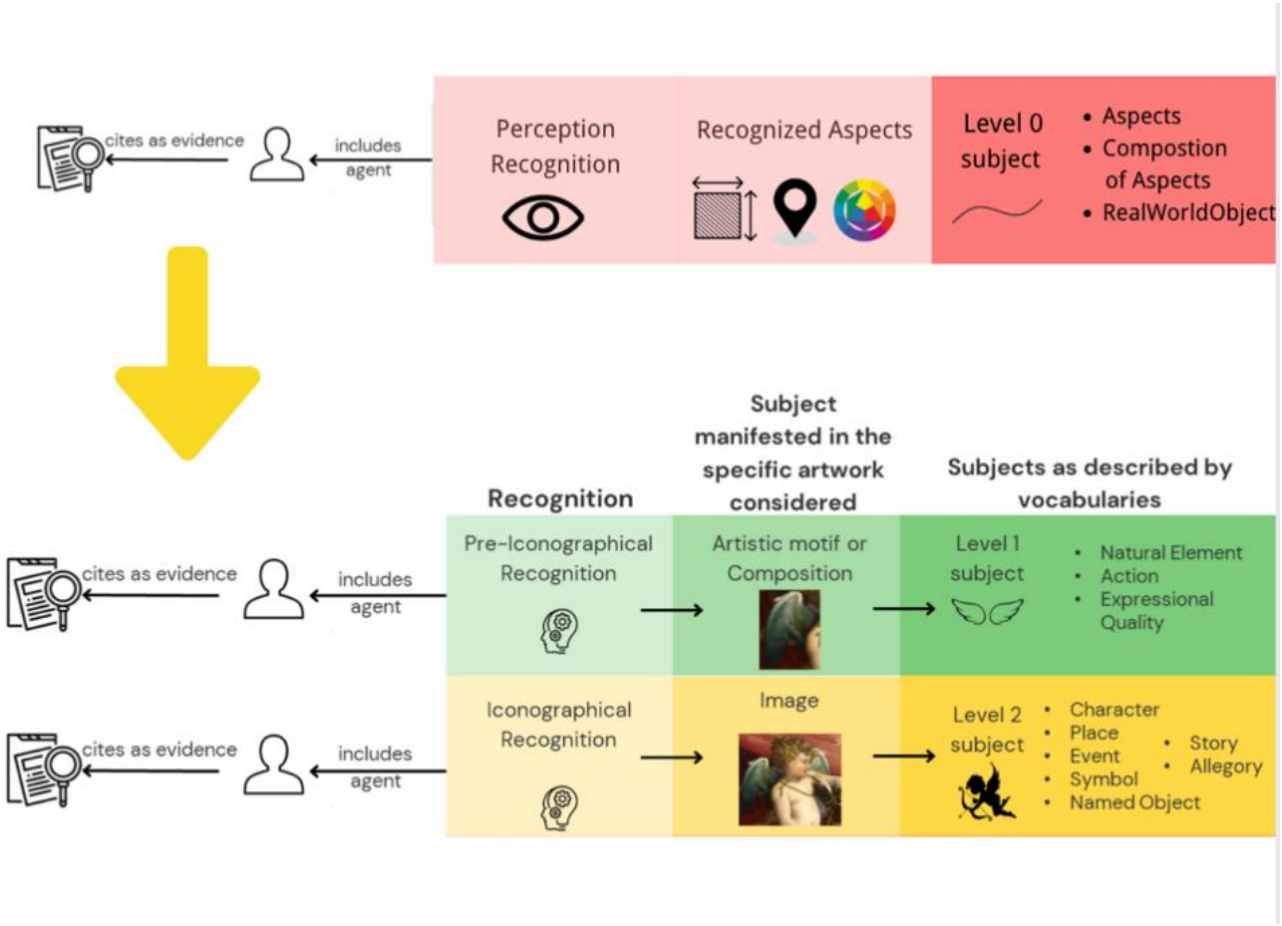


Figure 1. Align Perception-Recognition (in red) with the first two levels of the ICON ontology[2]

[2] Original image sourced from: https://dl.acm.org/doi/10.1145/3594724.

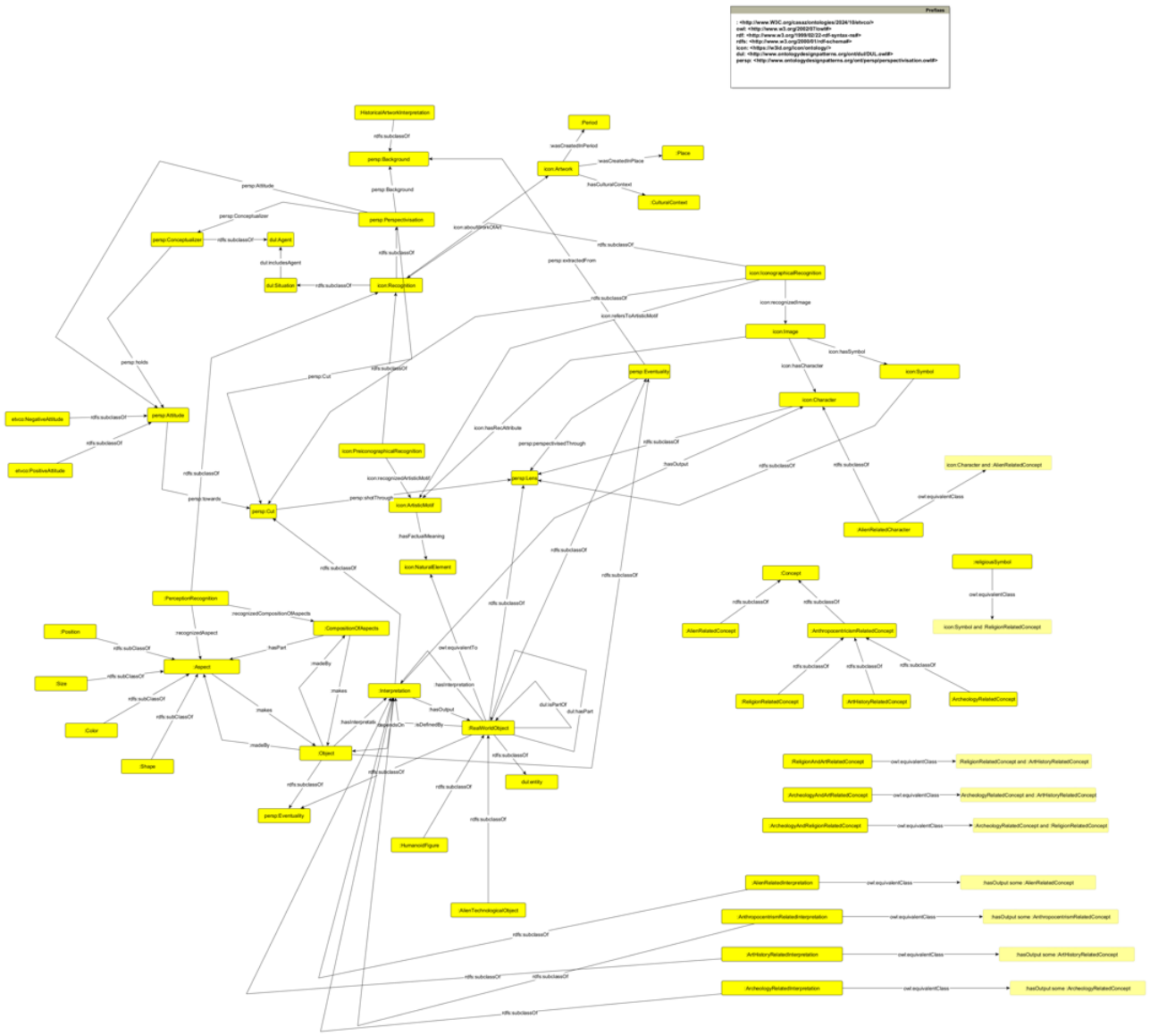


Figure 2. key classes and properties of ETVCO

4. EVALUATION

To assess the effectiveness and applicability of the ontology, we followed an iterative and user-centric evaluation methodology grounded in the Extreme Design (XD) framework. This process involved multiple stages, including data collection, knowledge graph construction, and competency question (CQ) testing, all designed to ensure the ontology meets the nuanced needs of diverse stakeholders.

We began by collecting a diverse dataset from scholarly articles, blogs, and conspiracy theorist websites. These sources typically featured images of cultural artifacts alongside descriptive text, often reflecting contrasting interpretations. The data was manually curated and structured into a CSV format, including fields such as image ID, title, textual description, and source URL. This structured dataset provided a foundation for both ontology development and evaluation, allowing us to simulate real-world usage scenarios.

Next, we utilized the FRED machine reader to automatically extract semantic knowledge from textual descriptions. These extractions were converted into RDF/XML files representing structured knowledge graphs. The resulting graphs were then imported into the ontology using Protégé connect individuals to classes and properties to construct a comprehensive A-box.

A central component of the evaluation involved simulating stakeholder needs through the construction of user stories. These narratives represented diverse perspectives—including those of art historians, archaeologists, conspiracy theorists, educators, and cognitive scientists—and formed the foundation for the development of competency questions (CQs). Each user story was translated into formalized CQs,

which were subsequently tested within Jupyter Notebooks to assess the ontology's expressiveness and reasoning capabilities.
These evaluations aimed to determine whether the ontology could support complex queries related to visual features, interpretive frameworks, agent perspectives, and cultural contexts. For instance, it was observed that conspiracy theorists often emphasize particular shapes or atypical anatomical features when interpreting artworks as evidence for the Ancient Alien Theory. Testing the ontology with the second competency question from the first user story ("What visual features contribute to the alien interpretation?") indicated that characteristics such as large eyes and oval-shaped heads are frequently associated with alien representations (see Fig. 3).
Furthermore, the results from testing CQ3 and CQ4 of the second user story demonstrated that circular or disk-shaped objects—especially those depicted in the sky—are commonly interpreted as unidentified flying objects (UFOs) by conspiracy theorists. A comparative analysis of these same objects revealed divergent interpretations from other agents; for example, art historians often regard such elements as symbolic representations of sacred or divine entities (see Fig. 4 and 5).
To ensure reproducibility and transparency, all CQ tests and results were documented and made publicly available via GitHub[3]. Iterative testing led to continual refinement of the ontology, confirming its capacity to support interdisciplinary research. The final evaluation demonstrated ontology's strength in enabling detailed semantic queries about visual heritage, accommodating both mainstream academic interpretations and alternative viewpoints. Moreover, it facilitated the analysis of cognitive processes such as perspectivization and recognition, reinforcing the ontology's relevance to fields like cognitive science and digital humanities.

CQ2: What visual features(attribute) (e.g., head size, body shape) contribute to the alien interpretation?

```
cq = """
PREFIX etvco:<https://raw.githubusercontent.com/knowledge-representation-2024/E.T-and-visual-culture-ontology-perspectives-
PREFIX persp:<http://www.ontologydesignpatterns.org/ont/persp/perspectivisation.owl#>
PREFIX dul:<http://www.ontologydesignpatterns.org/ont/dul/DUL.owl#>
PREFIX icon:<https://w3id.org/icon/ontology/>

SELECT DISTINCT ?character ?realWorldObject ?visualFeature ?quality
WHERE {
    ?character a etvco:AlienRelatedCharacter;
        icon:isCharacterOf ?image.
          ?image a icon:Image;
        icon:hasRecAttribute ?artisticMotif.
    ?artisticMotif a icon:ArtisticMotif;
        icon:hasFactualMeaning ?realWorldObject.
    ?realWorldObject a etvco:RealWorldObject.
      ?visualFeature dul:isPartOf ?realWorldObject;
        dul:hasQuality ?quality.
}"""

cqres = newg.query(cq)

for row in cqres:
    print(row)
```

```
(rdflib.term.URIRef('http://www.ontologydesignpatterns.org/ont/fred/domain.owl#grey_1'), rdflib.term.URIRef('http://www.ontol
ogydesignpatterns.org/ont/fred/domain.owl#humanoid_1'), rdflib.term.URIRef('http://www.ontologydesignpatterns.org/ont/fred/do
main.owl#eye_1'), rdflib.term.URIRef('http://www.ontologydesignpatterns.org/ont/fred/domain.owl#Large'))
(rdflib.term.URIRef('http://www.ontologydesignpatterns.org/ont/fred/domain.owl#grey_1'), rdflib.term.URIRef('http://www.ontol
ogydesignpatterns.org/ont/fred/domain.owl#humanoid_1'), rdflib.term.URIRef('http://www.ontologydesignpatterns.org/ont/fred/do
main.owl#head_1'), rdflib.term.URIRef('http://www.ontologydesignpatterns.org/ont/fred/domain.owl#Oval'))
```

Figure 3. First example of ETVCO-Ontology evaluation using SPARQL queries and competency questions

[3] https://github.com/knowledge-representation-2024/E.T-and-visual-culture-ontology-perspectives-on-Extraterrestrial-influence-in-visual-heritage-ETVCO-/tree/main

CQ1: "What kind of shapes are interpreted as alien technology?"

```
cq ="""
PREFIX etvco:<https://raw.githubusercontent.com/knowledge-representation-2024/E.T-and-visual-culture-ontology-perspectives-
PREFIX dul:<http://www.ontologydesignpatterns.org/ont/dul/DUL.owl#>

SELECT DISTINCT ?shape ?realworldobject
WHERE {
    ?shape a etvco:Shape.
    {
        ?shape dul:isPartOf ?composition.
        ?composition a etvco:CompositionOfAspects;
        etvco:makes ?object.
    }
    UNION
    {
        ?shape etvco:makes ?object .
    }
    ?object etvco:hasInterpretation ?interpretation.
    ?interpretation etvco:hasOutput ?realworldobject.
    ?realworldobject a etvco:AlienTechnologicalObject.
}
"""

cqres = newg.query(cq)

for row in cqres:
    print(row)
```

```
(rdflib.term.URIRef('http://www.ontologydesignpatterns.org/ont/fred/domain.owl#Disc-shaped'), rdflib.term.URIRef('http://ww
w.ontologydesignpatterns.org/ont/fred/domain.owl#ufo_1'))
(rdflib.term.URIRef('http://www.ontologydesignpatterns.org/ont/fred/domain.owl#Round'), rdflib.term.URIRef('https://raw.githu
busercontent.com/knowledge-representation-2024/E.T-and-visual-culture-ontology-perspectives-on-Extraterrestrial-influence-in-
visual-heritage-ETVCO-/refs/heads/main/etvcontology.ttl#ufo_2'))
```

Figure 4. Second example of ETVCO-Ontology evaluation using SPARQL queries and competency questions

5. RESULT AND FUTURE WORK

The evaluation confirmed the ontology's effectiveness in several key areas. First, it demonstrated the capacity to support semantic queries that explore diverse interpretations of visual motifs. Additionally, it facilitated interdisciplinary research by bridging cognitive, cultural, and historical analyses. It also proved useful in documenting the range of attitudes and biases that influence interpretation, including those associated with conspiracy theories. Collectively, these findings highlight the ontology's potential as a valuable tool for analyzing visual heritage and advancing research related to the Ancient Astronaut Hypothesis.

Future developments will build upon the current version of the ontology in several ways. One major direction involves integrating psychological theories of attention, memory, and stereotyping in order to deepen the understanding of cognitive biases that shape visual interpretation. Another focus will be on enhancing the ontology's capacity to model complex perspective-dependent analyses by incorporating more detailed representations of agent-context interactions. Furthermore, the dataset will be expanded to include a wider range of visual artifacts and interpretive narratives, thereby supporting broader applications across disciplines.

By addressing these objectives, we aim to develop a more comprehensive model that supports the study of both mainstream and alternative interpretations of human artifacts, along with the cognitive processes underlying them. This approach is expected to advance innovative research at the intersection of cognitive science and cultural analysis, particularly in the domain of visual artifact interpretation.

ACKNOWLEDGEMENTS

In this project, ChatGPT has been used as a fictional stakeholder incorporating user stories and competency questions tailored to stakeholders.

REFERENCES

Bekiari, C., Bruseker, G., Canning, E., Doerr, M., Michon, P., Ore, C.-E., Stead, S., & Velios, A. (Eds.). (2024). Volume A: Definition of the CIDOC Conceptual Reference Model (Version 7.3). https://cidoc-crm.org/sites/default/files/cidoc_crm_version_7.3.pdf

Gangemi, A., & Presutti, V. (2022). Formal representation and extraction of perspectives. In P. Vossen & A. Fokkens (Eds.), Creating a more transparent internet: The perspective web (pp. 208–228). Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/9781108641104.016

Presutti, V., Daga, E., Gangemi, A., & Blomqvist, E. (2009). eXtreme design with content ontology design patterns. Semantic Technology Laboratory, ISTC-CNR. https://ceur-ws.org/Vol-516/pap21.pdf

Rayhan, A. (2023). Sapiens: A brief history of humankind: Exploring the alien hypothesis. https://doi.org/10.13140/RG.2.2.25808.69125

Sartini, B., Baroncini, S., van Erp, M., Tomasi, F., & Gangemi, A. (2022). ICON: An ontology for comprehensive artistic interpretations. ACM Journal on Computing and Cultural Heritage, 16(3), Article 59, 1–38. https://doi.org/10.1145/3594724

Automating XML-TEI Encoding of Unpublished Correspondence: A Comparative Analysis of two LLM Approaches

Marco De Cristofaro, Daniel Zilio
[1] Université de Mons; Université de Namur, Belgium - Marco.DECRISTOFARO@umons.ac.be
[2] Università di Padova, Italy – daniel.zilio@unipd.it

ABSTRACT (ENGLISH)

Encoding texts in TEI-XML format is critical for research projects dealing with copyright and open access issues. The PubCiNET project reconstructs the social network of Italian intellectuals in publishing and film between the 1950s and 1980s. During the three decades, exchanges and collaborations between professionals in these creative industries increased, affecting the convergence of literature and film, film professionals' engagement in publishing, and the perception of publishing as a prestigious field for filmmakers. The project utilizes an XML-TEI encoded corpus of archival correspondence to map the intellectual network. However, key challenges arise, including the complex retrieval of data from vertically structured archives, copyright issues due to the contemporary timeframe, and the sustainability of handling vast volumes of documents. This study proposes a first attempt at these challenges by applying automated text encoding through large language models (LLMs). The research explores automated encoding using ChatGPT-4 and Claude 3.5 Sonnet, analyzing their capabilities in enhancing access to archives and automating the labor-intensive encoding process. Initial findings indicate varying success rates: while both LLMs efficiently extract metadata, they differ in their ability to recognize information in the text of the letters. Improving their efficiency in terms of information recognition and the reliability of reference materials could contribute to more efficient and faster encoding, allowing for greater sustainability in research.

Keywords: XML-TEI Encoding; Intellectual Networks; Automated Text Analysis; Publishing and Film Industries; Large Language Models (LLMs)

ABSTRACT (ITALIANO)

Codifica automatica in XML-TEI della corrispondenza inedita: un'analisi comparativa di due approcci basati su LLM

La codifica XML-TEI dei testi fornisce un buon supporto per i progetti di ricerca che affrontano questioni di copyright. Il progetto PubCiNET ricostruisce la rete intellettuale tra professionisti dell'editoria e del cinema tra gli anni '50 e '80. Questo periodo ha segnato un aumento degli scambi e delle collaborazioni tra le due industrie creative, influenzando la convergenza tra letteratura e cinema. Per ricostruire la rete intellettuale, il progetto utilizza un corpus di corrispondenza d'archivio codificato in XML-TEI. Tuttavia, la complessa raccolta di dati da archivi strutturati verticalmente, le criticità relative al copyright del materiale archivistico e la sostenibilità nella gestione di grandi volumi di documenti presentano problematiche relative all'accessibilità e all'utilizzo dei risultati sia da parte dei ricercatori sia da parte degli utenti finali. Questo studio propone un tentativo di codifica automatizzata di un corpus di testi attraverso Large Language Models (LLMs). La ricerca esplora la codifica automatizzata utilizzando ChatGPT-4 e Claude 3.5 Sonnet, analizzando le loro capacità di migliorare il processo intensivo della codifica. I risultati iniziali indicano tassi di successo variabili: mentre entrambi gli LLMs estraggono efficientemente metadati, dimostrano un grado di affidabilità differente per quanto riguarda la capacità di riconoscere determinate informazioni dei testi. Incrementare la loro efficienza in termini di riconoscimento delle informazioni e di affidabilità dei materiali di riferimento potrebbe contribuire a una codifica più efficiente e rapida permettendo una maggiore sostenibilità della ricerca

Parole chiave: XML-TEI; Reti intellettuali; Analisi testuale automatizzata; Editoria e industria cinematografica; Large Language Models (LLMs)

1. INTRODUCTION

Encoding new texts in XML-TEI format plays a central role in research projects that focus on automatically analyzing texts with issues related to copyright, as is the case with the PubCiNET project. In the PubCiNET project, we aim to reconstruct the social network that emerged between the 1950s and 1980s among intellectuals from two creative industries: publishing and film industry. As it has been suggested and demonstrated (Brunetta, 2004; Ivaldi, 2001; Simonetti, 2018), starting from the post-war period,

intellectual exchanges between professionals in the publishing world and those in the film industry intensified and strengthened. The intensification of the exchange of ideas has led to three main consequences. First of all, several authors admit the influence of cinema on their stylistic choices. Secondly, there is a growing interest in the publishing market among film professionals such as directors, screenwriters, and producers. They contribute by making editorial proposals related to their own cinematic works or by publishing books based on their creative experience. Finally, film professionals—particularly directors—regard the publishing world as a space for achieving recognition beyond the boundaries of their specific cultural domain. To make this crucial intellectual network accessible and visible, one can draw on correspondence between the various professionals involved: publishers, directors, publishing house consultants, and producers. This approach seeks to map, for the first time, for what concerns the Italian cultural field, the process of transmitting ideas from the film industry to publishing and vice-versa. The overall goal of PubCiNET is to identify the key nodes through which new forms of cinematic expression circulate within the publishing sphere. To support this objective, the encoding of both published and unpublished letters helps to map the points of contact between the film and publishing industries. Given the large number of letters collected from seven different archives and transcribed manually, automated encoding can provide substantial support for the project's success. This paper focuses on the initial attempts to automatically encode the extensive corpus of collected letters.

The study of correspondence has proven to be a valuable tool for exploring criticism. Correspondence between professionals in the cultural industries makes it possible to investigate the relationship between the private dimension—that is, the initial intentions of the various agents involved—and the public dimension, meaning the final outcome accessible to the public. The role of correspondence in analyzing the relationship between initial intentions and final outcomes has been demonstrated in both the publishing (Cadioli, 2012, 2021; Italia, 2006, 2013; Pischedda, 2022) and film fields (Guerra & Martin, 2019; Mariani & Venturini, 2017; Noto, 2019; Noto et al., 2020; Rigola, 2021). Nevertheless, the structure of publishing and cultural archives makes it difficult to study interactions between multiple figures rather than focusing on just one.

The three main critical issues that emerged during the study of the project are as follows: i) the vertical structure of personal or publishing house archives makes direct data retrieval challenging: while the workflow for collecting materials from paper archives, both in publishing and cinema, has been well defined and outlined(Cesana, 2006; Guerra & Martin, 2019), the involvement of multiple archives with different consultation and publication practices makes data collection significantly more difficult; ii) given the selected time frame, 1950–1980, considerations on the copyright status of each document are necessary: the need to request authorization from all rights holders to first consult and then publish the materials raises questions about the sustainability of projects dealing with contemporary archives; iii) a third critical issue again concerns research sustainability: the volume of documents consulted and collected, which would need to be encoded to enable future consultation and analysis, necessitates the exploration of automated processes for managing data in light of advancements in encoding by LLMs (DeRose, 2024).

Building on an initial manual encoding in XML TEI of the correspondence collected, we will compare automated encoding techniques using two different LLMs: a widely available LLM, ChatGPT4[1], selected based on previous experiments specifically involving the automated encoding of correspondence(Pollin et al., 2023), and another LLM up-and-comer, Claude 3.5 Sonnet[2] chosen to explore the potential for creating a ready-to-use chatbot for educational purposes.

2. The XML-TEI Encoding

The structure of the data model was inspired by four different projects: the edition of *Vespasiano da Bisticci. Lettere* (Tomasi, 2013); the Darwin Correspondence Project (University of Cambridge, 2022); the Van Gogh Letters(Van Gogh Museum & Huygens ING, 2018); the Bellini Digital Correspondence(Del Grosso & Spampinato, 2023).

The proposed markup model aims to translate the core elements of the edition into a formal system. The core elements are: the people mentioned in the letters; the occupation of the people; works; book series; the organizations; the places; the name of the awards and bibliographic references. We represent all the data in a macro.xml file through appropriate elements and attributes. The data encoded in the specific file of each letter are linked to the macro.xml file. This choice will prove particularly useful in the

[1] https://openai.com/index/gpt-4/.
[2] https://www.anthropic.com/news/claude-3-5-sonnet.

phase of encoding automatization. Most of the elements considered are therefore based on a pre-existing approach, and we refer to De Cristofaro(2024) for further details.
The most interesting element, however, is the encoding of the works. They are not always identified by a specific title in the letters, as they often refer to drafts that precede the final published version. In fact, the same book may be mentioned at different stages of its development. In our encoding, we chose to treat each draft as a separate work, since they have distinct characteristics and the final version is not always published. Annotating drafts separately also supports further philological and editorial research. Another goal of the project is to actively describe the publication process. To encode the works, we used a generic <rs> tag to define the references to each work as they appear in the letters. The @type attribute specifies the type of work: bibliographic or cinematic. We encoded each phase of a work as a different work to define the various moments of the publication process. These will be the main challenges in the automatization of the process through LLMs: i) how to recognize a work that is quoted without its specific published name; ii) how to encode a pre-publication step of a specific volume. The prompt and the provided referenced encoded texts would be crucial as well as the macro.xml file.

3. The XML-TEI Encoding automatization: ChatGPT4

Given the vast amount of material underpinning the project, in this work, we are experimenting with an automated encoding process using the LLMs. We will compare automated encoding techniques using two different large language models (LLMs): a widely available LLM, OpenAI ChatGPT-4, selected based on previous experiments specifically involving the automated encoding of correspondence (Pollin et al., 2023); and an LLM developed by the Anthropic Claude 3.5 Sonnet. We decided to use ChatGPT in chat mode because it has demonstrated promising capabilities at the current state of the art. In contrast, we opted for Claude 3.5 via API due to its timing performance compared to the chat mode. For future tests, our planned approach will involve using both models through their API.
Our initial attempt to apply automated XML TEI encoding using LLMs will be grounded in the research by Pollin et al. (2023). Two main findings from their study were: i) the prompt is crucial for success; and ii) human evaluation is becoming increasingly important. However, there is one key difference between our project and the primary goal of Pollin's experiment: they applied the automated LLM strategy solely to the text of the letters and only to a single text. In contrast, given our objective to automatically encode a corpus of hundreds of letters, we have decided to conduct our experiment by also including the <teiHeader> and considering from the outset encoding more than one letter through a single prompt. Building on this difference and considering the crucial role of the prompt, we developed our prompt based on the example provided by Pollin et al. (2023).
The prompt has been divided into two parts. The first part gathers all the information regarding the <TeiHeader>. Since LLMs cannot deduce archival metadata solely from the text, we decided to adopt the following strategy: providing the letters organized by archive in order to automatically encode the section related to each repository. This approach would also facilitate the automation of the metadata section. In this initial part of the prompt, we provide a detailed map of all the information of the <TeiHeader>. Here, we present an example of key information to be included in the TeiHeader, as outlined in the prompt.:

> You will act as a skilled expert automaton that is proficient in transforming unstructured text, specifically multilingual letters, into well-formed TEI XML. Analyze the provided text based on the mapping rules I have shared and then execute the transformation to produce TEI XML, ensuring you adhere to the TEI guidelines and annotate if certain.
> Mapping rules:
> * <teiHeader>: contains metadata;
> fill the <teiHeader> using all the tags that are relevant for the project metadata as follows:[…]
> * The first <correspAction> nest in <correspDesc> has @type attribute "sent"
> * The <persName> nest in the first <correspAction> has @ref attribute that links the sender to the right person who sends the letter in the macro.xml file. You can understand who sends the letter from the title and the signature. This <persName> also has a @role attribute that refers to the role attributed to the person who sends the letter in the macro.xml file

The second section of the prompt has been structured specifically in relation to the content of the letter. Here is an example of how we structured this second part of the prompt:

> * <opener> Opening of the letter

> * <salute> Salutations within the letter
> * <closer> Closing of the letter
> * <signed> Signature section
> * <postscript> Represents a postscript
> * <persName> Person with a @ref attribute that links the person to the corresponding person in the macro.xml file
> * <orgName> Organisation with a @ref attribute that links the organization to the corresponding organization in the macro.xml file

Since we are communicating via ChatGPT4, we have provided the macro.xml file containing all the information about the individuals, organizations, volumes, and films mentioned. We also provided nine sample files of the encoded letters collected from the same archive. The final part of the prompt refers to the files provided as examples:

> I provide you with nine examples of encoded letters collected in the same archive.
> I provide you with the macro.xml file where you can find all the information you need to encode all the data of the letters.

Finally, we provided the plaintexts of the four letters to be encoded.
The first and fundamental issue encountered with encoding the four letters together is that the result is invalid encoding. In particular, validation fails due to three recurring errors in the TeiHeader:

- Invalid characters: the inclusion of unrecognized characters "©", which represent the copyright symbol "©" present in the example letters.
- Incorrect placement of the <langUsage> tag: the <langUsage> tag should be nested within <profileDesc>. However, in all four cases, GPT closes the <profileDesc> tag first and then places the <langUsage> tag, which is incorrect.
- Improper use of the <p> tag: within the <closer> tag, which contains the closing and signature of the letter, GPT inserts a <p> tag that is not allowed.

Since these were recurring errors, we specifically requested corrections via chat. GPT successfully corrected the three recurring issues, subsequently providing valid files.
Once valid files were obtained, a second verification was conducted on ChatGPT's ability to encode the text of the letters within the . In this case, it demonstrated the ability to encode the various parts that characterize the letter: <opener>, <closer>, <signed>, and <postscript> were correctly identified. The letter was divided into paragraphs: our encoding does not require a strict <lb> tag since visualization of the archival document is not the ultimate goal of the research. A division into <p> is sufficient for the project's objectives.
Regarding the ability to identify relevant information within the text, performance recorded an average success rate of 85%. The success rate was calculated based on a comparison between the elements recognized and correctly encoded by the LLM and those identified and manually encoded in the same letter. In this initial phase of experimentation, manual supervision is necessary for the evaluation metrics. Almost all names were correctly identified with the <persName> tag and the @ref attribute linking to the correct person in the macro.xml file (Fig. 1). Once again, a recurring shortcoming in the encoding was noted: the person included in the <signed> tag was never encoded in any of the 4 letters. However, a brief prompt specifying the inclusion of that figure in the encoding allowed for the correction of the omission in all four letters.
As for organizations, publishing houses were correctly identified and encoded using the <orgName> tag. In this case, it does not appear that the macro.xml file was used as a reference for encoding the organization. In the manual encoding, indeed, the publishing houses are encoded according to the following scheme: <orgName ref="macro.xml#cappelli-edizioni">Cappelli</orgName>. The publishing house's name is supplemented with the term "edizioni." ChatGPT's encoding, on the other hand, adopts the following scheme: <orgName ref="macro.xml#longanesi">Longanesi</orgName>.
Two considerations are necessary. On the one hand, the support of the macro.xml file does not always seem relevant and, in some cases, proves counterproductive. On the other hand, when the LLM does not rely on the information from the macro.xml file, it performs much better in recognizing names and organizations. The encoding of books is more complex. As previously specified, the objective here is to trace the publication process. We aim to encode published works and works in the development stages. In

the stages before publication, works may have different or not always explicit titles. In the macro.xml file, we have encoded each stage with a corresponding unique identifier. ChatGPT demonstrates strong performance in encoding published works but fails when encoding the pre-publication stages. This result demonstrates that the macro.xml file was not considered the main reference point, while it relies on its own information for encoding works.

4. The XML-TEI Encoding automatization: Claude 3.5 Sonnet

We developed a Python script to directly interact with the model through the provided API[3]. We decided to use it instead of the dedicated interface, maintaining the same logic as the previous approach. Indeed, this script follows a two-step process to produce the final result. In the first step, the prompt comprises three components: a list of instructions, an XML file containing all directives, and the plain text of the letter. The instructions have been articulated in a manner that aligns more closely with the requirements of the model while still conveying the same essential demands as those outlined for ChatGPT. The output from this first prompt then serves as the input for the second step, where the directives in the macro.xml file are re-evaluated to identify the names of people, movies, and other relevant terms.

The results of the first prompt demonstrate its ability to produce valid files. The metadata is correctly encoded, no invalid characters are present, and the nesting order is properly executed.

At the textual level, within the , two substantial aspects distinguish the automated encoding performed by Claude Sonnet 3.5. The first is that there is perfect adherence to the plaintext file: the encoding of the letter's parts is correct, and if the plaintext, as happens in some cases, lacks the letter's signature, the <closer> is also absent from the encoding. A second substantial element is the greater alignment with the macro.xml file. The information is primarily drawn from the reference file, as confirmed by the encoding of <persName>, all of which include the @ref attribute with the correct identifier from the macro.xml file, and the encoding of <orgName>, which also features the corresponding @ref attribute linking to the correct identifier in the macro.xml file.

Here, the main issues are found in the encoding of works. Published works are once again correctly recognized and linked to the macro.xml file. However, the stages before publication are not encoded.

```
purtroppo negativa. Come lei forse sapeva dal comune amico <persName
  ref="macro.xml#renzo-renzi">Renzo</persName>, io avevo ricevuto lusinghiere e
considerevoli offerte da case editrici con i cui direttori editoriali ero e sono da tempo
in ottimi rapporti: ma per l'antica amicizia che mi lega a <persName
  ref="macro.xml#renzo-renzi">Renzo</persName>, avevo declinato le proposte concrete e
pressanti della <orgName ref="macro.xml#longanesi">Longanesi</orgName>, di <orgName
  ref="macro.xml#mondadori">Mondadori</orgName> e di <orgName ref="macro.xml#rizzoli"
  >Rizzoli</orgName>, per privilegiare, la <orgName ref="macro.xml#cappelli"
  >Cappelli</orgName> alla quale mi sento vicino per il lavoro di tanti anni.</p>
```

Fig. 1 ChatGPT4 encoding of a Letter from Federico Fellini to Mario Musso, 21 January 1990

```
risposta purtroppo negativa. Come lei forse sapeva dal comune amico <persName
    ref="macro.xml#renzo-renzi">Renzi</persName>, io avevo ricevuto lusinghiere
e considerevoli offerte da case editrici con i cui direttori editoriali ero e
sono da tempo in ottimi rapporti: ma per l'antica amicizia che mi lega a
    <persName ref="macro.xml#renzo-renzi">Renzo</persName>, avevo declinato le
proposte concrete e pressanti della <orgName ref="macro.xml#longanesi-edizioni"
    >Longanesi</orgName>, di <orgName ref="macro.xml#mondadori-edizioni"
    >Mondadori </orgName> e di <orgName ref="macro.xml#rizzoli-edizioni"
    >Rizzoli</orgName>, per privilegiare, la <orgName
    ref="macro.xml#cappelli-edizioni">Cappelli </orgName> alla quale mi sento
vicino per il lavoro di tanti anni. </p>
```

Fig. 2 Claude encoding of a Letter from Federico Fellini to Mario Musso, 21 January 1990

5. Conclusions. For a comparison between the two models

We acknowledge the limited scope of our current evaluation, which analyzed only four letters as a preliminary test case. This small sample size was chosen for an initial proof-of-concept to establish baseline methodological approaches and identify key challenges before scaling to larger datasets. Furthermore, at this initial stage, the evaluation of the success rate of the encoding process had to be carried out manually. To assess the accuracy of the encoding, we compared the letters automatically encoded by the two LLMs with the same letters manually encoded in XML-TEI by the project lead. The comparison focused particularly on the ability of the automatic encoding to identify all relevant elements and annotate them correctly, as in the manual encoding. The evaluation metrics were primarily based on a quantitative calculation of the number of elements recognized and correctly annotated automatically compared to those annotated manually. Once a satisfactory success rate is achieved, the automatic encoding will be extended to a larger corpus, organized by archive. From the previous evaluations, we can attempt a comparison between the two models. If ChatGPT initially shows difficulties in producing a valid file, this is not the case with Claude, which immediately encodes the four letters and returns valid files. At the same time, Claude demonstrates greater adherence to the input files provided: both to the examples of letters and to the macro.xml file from which it extracts all the most relevant information to encode. ChatGPT, by deviating from the input information, proves to be more effective in independently

[3] The model used was *anthropic.claude-3-5-sonnet-20241022-v2:0*, running on Amazon Bedrock.

recognizing textual data, showing potential for improvement in encoding elements that are not previously included in the macro.xml file. It is worth noting that encoding must be carried out on a corpus of letters to demonstrate its actual efficiency and its real contribution to the sustainability of the research. The main challenges of the project lie precisely in the encoding and manual standardization of a large corpus of texts originating from different archival locations. Manual encoding and verification would not only be complex, requiring constant review, but also time-consuming, without always ensuring uniformity in the encoding process.

The implementation of a reference database, to which the texts of the letters would link with the respective identifiers for each element to be encoded, could offer additional advantages in automated encoding. In that case, Claude's greater adherence to input files would appear more efficient. On the other hand, to expand the project, we foresee a hybrid approach that also takes into account ChatGPT's superior capabilities in encoding elements not present in the input files and flagging them. We propose to explore a hybrid approach that integrates Large Language Models (LLMs) with complementary techniques in both the pre-processing and post-processing phases. We plan to implement a key initial technique called Named Entity Recognition (NER). This technique will enhance our model's ability to identify and classify specific entities within texts, including technical terms, and domain-specific concepts. We anticipate that this capability will significantly improve the contextual understanding and relationship mapping in our analysis pipeline. In the post-processing phase, we aim to establish an automatic feedback system based on current patterns of encoded letters. This will create a self-improving mechanism that continuously evaluates output quality against established benchmarks. This system will serve as a foundational tool for training the model to refine its prompts effectively. Additionally, our next research stage will focus on comprehensive fine-tuning of the parameters for both models. Our goal is to optimize their performance, particularly in terms of response relevance and contextual precision.

To address the methodological limitations identified in this preliminary phase of our research, we plan to adopt a more quantitative evaluation framework in the next stages. This will involve metrics to assess the performance of LLM models in XML-TEI encoding, including: (1) precision in identifying named entities (such as people, organizations, and works), along with analyses of false positives and false negatives; (2) accuracy in associating the correct references from the macro.xml file; (3) the structural validity of the generated markup; and (4) computational efficiency. This systematic approach will enable a more objective evaluation of the capabilities of different LLMs and offer concrete guidance for optimizing prompts and implementation strategies in similar archival contexts.

While manual verification still appears indispensable, as argued by Pollin (2023) and DeRose(2024), encoding a corpus rather than a single text would still benefit the project's development, resulting in substantial savings in both human and economic resources. From the perspective of research sustainability, involvement of LLMs has already demonstrated their potential(Schmidgall et al., 2025). Considering the archival challenges we have identified, from the plurality of figures involved to the diverse origins of archival materials, the possibility of automated encoding of an extensive corpus of letters would greatly support the achievement of the project's objectives.

ACKNOWLEDGEMENTS

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska Curie grant agreement No 101034383

REFERENCES

Brunetta, G. P. (2004). *Gli intellettuali italiani e il cinema*. Bruno Mondadori.

Cadioli, A. (2012). *Le diverse pagine. Il testo letterario tra scrittore, editore e lettore*. Il saggiatore.

Cadioli, A. (2021). *«La sana critica». Pubblicare i classici italiani nella Milano di primo Ottocento*. Firenze University Press.

Cesana, R. (2006). La memoria bibliografica: Storia dell'editoria e archivi editoriali. *Bibliologia : an International Journal of Bibliography, Library Science, History of Typography ant the Book*, *1*, 175–197.

De Cristofaro, M. (2024, October 10). *From Cinema to Publishing, there and back. Encoding a Corpus of Letters Between Filmmakers and Publishers in XML TEI*. TEI 2024. Texts, languages and communities.

Del Grosso, A. M., & Spampinato, D. (2023). *Bellini Digital Correspondence* (digitale). Cnr Edizioni. https://bellinicorrespondence.cnr.it

DeRose, S. J. (2024). Can LLMs help with XML? *Proceedings of Balisage: The Markup Conference 2024*, *29*. https://doi.org/10.4242/BalisageVol29.DeRose01

Guerra, M., & Martin, S. (2019). La cultura della lettera. La corrispondenza come forma e pratica di critica cinematografica. *Cinergie – Il cinema e le altre arti*, *8*(15), 1–3. https://doi.org/10.6092/issn.2280-9481/9661

Italia, P. (2006). Le «penultime volontà dell'autore». Considerazioni sulle edizioni d'autore del Novecento. *Ecdotica*, *3*, 175–186.

Italia, P. (2013). *Editing Novecento*. Salerno.

Ivaldi, F. (2001). *Effetto rebound. Quando la letteratura imita il cinema*. Felici.

Mariani, A., & Venturini, S. (2017). L'archivio e lo studioso. *Bianco e Nero*, *78*(588–589), 40–51.

Noto, P. (2019). Quale "mestiere del critico"? Un'intrusione nella corrispondenza di Guido Aristarco. *Cinergie – Il Cinema E Le Altre Arti*, *8*(15), 55–67. https://doi.org/10.6092/issn.2280-9481/9357

Noto, P., Malvezzi, J., & Mariani, A. (2020). Spostamenti progressivi nella critica cinematografica tra 1930 e 1970: Spazi, relazioni, movimenti. *Cinergie – Il Cinema E Le Altre Arti, 12*(23), 1–4. https://doi.org/10.6092/issn.2280-9481/17672

Pischedda, B. (2022). *La competizione editoriale. Marchi e collane di vasto pubblico nell'Italia contemporanea (1860-2020)*. Carocci.

Pollin, C., Steiner, C., & Zach, C. (2023). *New Ways of Creating Research Data: Conversion of Unstructured Text to TEI XML using GPT on the Correspondence of Hugo Schuchard with a Web Prototype for Prompt Engineering*.

Rigola, G. (2021). I fondi archivistici personali, la corrispondenza e la ricerca sul cinema. Il caso del carteggio tra Elio Petri e Leonardo Sciascia. In *Scrivere la Storia, costruire l'archivio. Note per una storiografia del cinema e dei media* (pp. 55–67). Meltemi.

Schmidgall, S., Su, Y., Wang, Z., Sun, X., Wu, J., Yu, X., Liu, J., Liu, Z., & Barsoum, E. (2025). *Agent Laboratory: Using LLM Agents as Research Assistants* (arXiv:2501.04227). arXiv. https://doi.org/10.48550/arXiv.2501.04227

Simonetti, G. (2018). «Quasi verità». Scrittori italiani e cinema. In *La letteratura circostante. Narrativa e poesia nell'Italia contemporanea*. il Mulino.

Tomasi, F. (2013). *Vespasiano da Bisticci, Lettere* (Digitale). CRR-MM, Università di Bologna. http://vespasianodabisticciletters.unibo.it/

University of Cambridge. (2022). *Darwin Correspondence Project*. https://www.darwinproject.ac.uk

Van Gogh Museum, & Huygens ING. (2018). *Vincent van Gogh. The Letters* (digitale). https://vangoghletters.org/vg/letters.html

Modelli e tecnologie integrate e innovative per una cittadinanza digitale equa e sostenibile

Cristina Marras[1], Vittoria Fabiani[1], Enrico Pasini[2], Lisa Reggiani[1], Pietro Sichera[1], Paolo Ongaro[3], Martina Rossi[3]
[1] CNR-ILIESI, Italia - cristina.marras@cnr.it, vittoria.fabiani@cnr.it, lisa.reggiani@cnr.it, pietro.scihera@cnr.it
[2] Università di Torino, Italia - enrico.pasini@unito.it
[3] M.E.T.A- Italia - paolo.ongaro@gruppometa.com martina.rossi@grupposcai.it

ABSTRACT (ITALIANO)
Accessibilità è generalmente intesa come la capacità dei sistemi informatici di erogare servizi e fornire informazioni fruibili, senza discriminazioni; il contributo discute questa definizione generale a partire dalle azioni dell'infrastruttura europea OPERAS - *OPen Scholarly Communication in the European Research Area for Social Sciences and Humanities*, in particolare il nodo italiano OPERAS-IT, che ha tra le sue sfide quella di affrontare l'ambivalenza degli ecosistemi digitali attraverso una scienza equa, aperta e accessibile (¶1). Si presentano (¶2) piattaforme/spazi dedicati, come i marketplace, e modelli di lavoro che esplicitano e favoriscono interconnessione, aggregazione, accesso e fruibilità di servizi e contenuti. Definiti gli obiettivi principali (¶3), si evidenzia come alcune delle applicazioni e degli strumenti (¶4) che sono alla loro base, per esempio le API - *Application Programming Interface*, grazie anche alla collaborazione/co-progettazione tra ricercatori e sviluppatori, a partire dalla loro definizione fino alla implementazione, sono elementi costitutivi e integranti del modello di scienza aperta e accessibile. Il paragrafo dedicato alle tecnologie è seguito da un caso d'uso del MarketPlace e del suo sistema di orchestrazione (¶5). In conclusione (¶6), l'accesso aperto ai dati e ai risultati della ricerca contribuisce certo all'inclusione proprio per il principio di condivisione e comunicazione delle conoscenze, ma è importante rimarcare il ruolo integrato delle tecnologie che le sostengono e le veicolano. L'insieme sinergico di questi componenti rafforza la libertà di pensiero e di ricerca verso una cittadinanza digitale consapevole.
Parole chiave: Ricerca aperta; Infrastrutture di Ricerca; MarketPlace; Application Programming Interface; Diamond Open Access

ABSTRACT (ENGLISH)
Integrated and innovative models and technologies for an equitable and sustainable digital citizenship.
Accessibility is generally understood as the ability of information systems to provide services and usable information without discrimination. The paper discusses this definition through the actions of the European infrastructure OPERAS - OPen Scholarly Communication in the European Research Area for Social Sciences and Humanities, in particular the Italian node OPERAS-IT, which has as one of its challenges to address the ambivalence of digital ecosystems through fair, open and accessible science (¶1). We present (¶2) dedicated platforms/spaces, such as marketplaces, and working models that make explicit and promote the interconnection, aggregation, access and usability of services and content. After defining the main objectives (¶3), the aim is to highlight and show how some of the applications and tools (¶4) that underpin them, e.g. APIs - Application Programming Interfaces, are constitutive and integral elements of the Open and Accessible Science model, also thanks to the collaboration/co-design between researchers and developers, from their definition to their implementation. This section describing the technologies is followed by a use case of the MarketPlace and the orchestration system (¶5). In conclusion (¶6), while open access to data and research results (e.g. scholarly publishing) certainly contributes to inclusion precisely because of the principle of sharing and communicating knowledge, it is also important to emphasise the integrated role of the technologies that support and communicate them. The synergetic combination of these components strengthens the freedom of thought and research towards a conscious digital citizenship.
Keywords: Open Research; Research Infrastructures; Marketplace; Application Programming Interface; Diamond Open Access

1. INTRODUZIONE
Accessibilità è generalmente intesa come "la capacità dei sistemi informatici di erogare servizi e fornire informazioni fruibili, senza discriminazioni" (Council of the European Union, 2023). Il contributo parte da questa definizione e prende in considerazione le attività svolte dal nodo italiano OPERAS-IT *OPen Scholarly Communication in the European Research Area for Social Sciences and Humanities* dell'infrastruttura

europea OPERAS [1] nell'ambito del progetto H2IOSC, che riunisce un cluster collaborativo di infrastrutture di ricerca italiane nei settori delle scienze umane e del patrimonio culturale.[2] OPERAS ha nelle sue sfide quella di rendere la scienza pubblica in forma equa, aperta e accessibile. Possiamo identificare quattro assi principali di intervento sulle quali al momento OPERAS concentra la sua attività: 1. Diversità linguistica nelle pubblicazioni; 2. Diversità nell'accesso ai contenuti; 3. Libero accesso e forme specifiche della produzione scientifica, con particolare accento sull'Accesso aperto "Diamond"; 4. Far crescere la fiducia (*fostering trust*) all'interno della sfera digitale.
Il tema della accessibilità equa e non discriminatoria va declinato (e affrontato) nelle diverse dimensioni che caratterizzano l'ecosistema digitale, dalle interfacce,[3] ai software e dati aperti,[4] agli algoritmi generativi,[5] alla valorizzazione della diversità linguistica. Il settore pubblico della ricerca e dell'innovazione sta perseguendo attivamente un ecosistema di comunicazione accademica senza scopo di lucro. Science Europe, cOAlition S, OPERAS e l'Agenzia Nazionale della Ricerca francese (ANR) hanno redatto nel 2022 un Piano d'Azione per il *Diamond Open access* (OA) (Ancion et al., 2022), un modello equo di pubblicazione accademica che non impone alcun costo agli autori o ai lettori e in cui gli elementi relativi al contenuto della pubblicazione sono di proprietà e controllati dalle comunità accademiche. Il piano propone di allineare e sviluppare risorse comuni per l'intero ecosistema (comprese riviste e piattaforme) nel rispetto della diversità culturale, multilingue e disciplinare. Il modello poggia su quattro elementi centrali: efficienza, standard di qualità, sviluppo delle capacità e sostenibilità, dando seguito alle raccomandazioni dell' "Open Access Diamond Journals Study" (Bosmann et al., 2021; Becerril et al., 2021). Il progetto DIAMAS ha sviluppato il DOAS-Diamond Open Access Standard (Consortium, 2024) che permette di valutare la rispondenza delle attività pubblicistiche a un insieme multidimensionale di criteri. Recentemente è stato lanciato l'EDCH-European Diamond Capacity Hub,[6] un'iniziativa collettiva pionieristica per far progredire la pubblicazione ad accesso aperto in Europa, nella quale OPERAS ha un ruolo centrale.
Il progetto H2IOSC sopramenzionato sviluppa una piattaforma aperta e multidisciplinare che permetterà di condurre ricerche innovative e computazionalmente intensive su dati e oggetti digitali complessi e integrare culturalmente e operativamente risorse digitali di tipi e ambiti diversi. Nel progetto, OPERAS-IT ha come principale attività la realizzazione del MarketPlace,[7] un luogo unico e di facile accesso per le esigenze specifiche della ricerca di dominio e interdominio. Lo sviluppo avviene secondo un modello ispirato agli esempi di EOSC - *The European Open Science Cloud*[8] e SSHOC -*Social Sciences & Humanities Open Cloud*,[9] che al tempo stesso introduce alcune importanti innovazioni. Nel seguito discutiamo le peculiarità del MarketPlace in rapporto all'accesso aperto e in particolare alle esigenze di innovazione equa e non discriminatoria nell'accesso a risorse e servizi.

2. IL MARKETPLACE

Nel contesto più ampio di creazione di ecosistemi di ricerca digitale aperti e interoperabili, volti a promuovere l'accessibilità e l'approccio FAIR[10] ai dati (Wilkinson et al., 2016), il MarketPlace di H2IOSC non è semplicemente un elenco di servizi: si presenta piuttosto come uno spazio unificato dove dati,

[1] Cfr. https://www.operas-eu.org/ (cons. 25/01/2025). OPERAS sostiene tre progetti finanziati dall'UE che supportano percorsi guidati dalla comunità per un'equa pubblicazione accademica aperta. OPERAS partecipa a CRAFT-OA e coordina DIAMAS e PALOMERA. Nonostante le loro aree di interesse siano distinte, lavorano insieme verso una visione ampia e comune per un ecosistema di editoria accademica più aperto ed equo.
[2] H2IOSC - Humanities and Heritage Italian Open Science Cloud**,** è un ecosistema di ricerca, finanziato dall'Unione europea NextGenerationEU – PNRR M4C2 - Codice progetto IR0000029, https://www.h2iosc.cnr.it/) (cons. 12/01/2025)
[3] Accessibilità e inclusività del web è tra gli obiettivi del World Wide Web Consortium (W3C) che ha definito fin dal 2008 delle Linee Guida, WCAG, per l'accessibilità nella progettazione di interfacce web: https://www.w3.org/TR/WCAG20/ (cons. 12/01/2025).
[4] Cfr. l'Open Knowledge Foundation che ha lavorato all'intersezione tra strumenti digitali all'avanguardia e una rete distribuita di comunità e movimenti per la conoscenza aperta: https://okfn.org/en/who-we-are/ (cons. 24/01/2025); si veda anche: AISA, Associazione italiana per la promozione della scienza aperta https://aisa.sp.unipi.it (cons. 06/10/2024).
[5] Cfr. Special Committee on Artificial Intelligence in a Digital Age, 2021 (cons. 24/01/2025)
[6] Cfr. https://diamas.org/ (cons. 22/01/2025).
[7] Distingueremo con la M e la P maiuscole il MarketPlace sviluppato nell'ambito del progetto H2IOSC (vd. nota 2) (https://www.operas-eu.org/, https://www.operas-it.org/).
[8] Cfr. https://eoscfuture.eu/results/ (cons. 25/01/2025).
[9] Cfr. https://sshopencloud.eu/ssh-open-marketplace (cons. 25/01/2025).
[10] Cfr. https://force11.org/info/the-fair-data-principles/ (cons. 25/01/2025).

servizi, workflow, software, strumenti, dataset, materiale didattico, progetti pilota sono resi disponibili con lo scopo di supportare le molteplici esigenze della ricerca di dominio e trasversale. Promuovendo l'adozione di paradigmi di scienza aperta, il MarketPlace si pone l'obiettivo di diventare un punto di accesso universale per molteplici tipologie di utenti e diversi modelli di utilizzo delle comunità di ricerca coinvolte (Cristofaro et al., 2024).
Per rendere tutto questo possibile, sono stati inizialmente identificati attori e scenari di utilizzo della piattaforma, sono stati poi svolti diversi studi preliminari con lo scopo di realizzare un prodotto tecnologico che rispondesse esattamente alle necessità definite. Nella Tabella 1 riportiamo gli esiti di un benchmark realizzato in fase di progettazione per valutare le funzioni e le sezioni principali del front-end del MarketPlace,con lo scopo di seguire le linee guida di accessibilità e le migliori pratiche più diffuse non solo nel settore delle *digital humanities*, ma prendendo anche in considerazione i leader del mercato tecnologico e della comunicazione.

	A	St	alphabetica	a	BASE	Behance	clarin:el	CORE	coursera		JSTOR	Medium	NIH	ProQuest	Semantic Scholar		Unsplash	WorldCat
Home	-	-	⊗	-	⊗	-	8/10	6/10	-	6/10	⊗	⊗	6/10	8/10	9/10	⊗	-	7/10
Search	8/10	9/10	7/10	6/10	⊗	8/10	⊗	⊗	7/10	⊗	8/10	7/10	8/10	8/10	8/10	6/10	9/10	**10/10**
List Page	9/10	9/10	⊗	6/10	6/10	9/10	7/10	7/10	9/10	9/10	7/10	7/10	**10/10**	6/10	7/10	7/10	-	7/10
List Page card	7/10	-	⊗	-	9/10	-	7/10	7/10	-	8/10	9/10	8/10	7/10	⊗	8/10	6/10	-	9/10
Product page	-	-	6/10	-	6/10	-	8/10	8/10	**10/10**	9/10	8/10	-	8/10	6/10	6/10	9/10	-	9/10
Bookmarks P.	8/10	6/10	-	8/10	9/10	7/10	-	-	-	-	9/10	9/10	8/10	-	**10/10**	-	7/10	9/10
Single Sing-on	-	-	-	-	-	-	8/10	-	-	-	-	-	-	8/10	8/10	9/10	-	-

Tabella 1 Rating di applicazioni

Il confronto con i marketplace dello spazio europeo della ricerca, effettuato per definire progettualmente l'esperienza che avrebbe offerto il MarketPlace, ha indicato come elemento cruciale la possibilità di fornire agli utenti un accesso rapido e intuitivo a strumenti, dataset e applicazioni e metterli nelle condizioni di scegliere i servizi più adatti alle loro esigenze, richiedendone l'esecuzione o combinandoli in workflow personalizzati ed effettivamente eseguibili in forma assistita. Questo rappresenta uno degli aspetti tecnologici più avanzati dell'intero progetto: nel MarketPlace è stato progettato un sistema di orchestrazione (basato su API e servizi di integrazione) per coordinare in modo armonioso risorse provenienti da diverse infrastrutture, trasformando il MarketPlace stesso in un punto di snodo centrale per la ricerca interdisciplinare. Grazie all'utilizzo e all'integrazione di strumenti della suite WSO2,[11] il MarketPlace offre un ambiente dove è possibile progettare flussi di lavoro complessi senza dover gestire direttamente l'esecuzione dei singoli componenti. In questo modo, gli aspetti infrastrutturali (es. autenticazione, passaggio di dati e conversioni di formato) vengono gestiti dal sistema in maniera automatica e trasparente per l'utente, riducendo le barriere per chi non possieda conoscenze specifiche di sviluppo o programmazione. Questo modello riduce il carico computazionale, e garantisce anche una maggiore scalabilità, flessibilità e sicurezza nella gestione delle risorse. Viene così a crearsi una infrastruttura solida in grado di soddisfare le esigenze di utenti con competenze e bisogni diversi.

3. OBIETTIVI

Il MarketPlace è un'infrastruttura tecnologica avanzata, progettata con l'obiettivo di assicurare una migliore esperienza nell'uso dei servizi, favorire la collaborazione, costruire un ecosistema di ricerca più aperto e inclusivo per sostenere l'interdisciplinarità e promuovere la cittadinanza digitale, in linea con le esigenze delle digital humanities e delle scienze sociali. Come mostreremo (¶4), la sinergia tra *WSO2 API Manager, WSO2 Identity Server* e *WSO2 Micro Integrator*, integrati nel MarketPlace, in combinazione con il cluster di infrastrutture del progetto H2IOSC, costituisce anche un test per la messa a punto e lo sviluppo di un modello virtuoso per il futuro delle piattaforme digitali per la ricerca. Ciò favorisce anche una gestione più efficiente delle risorse, una maggiore trasparenza, al fine di costruire un rapporto di fiducia con la comunità scientifica e con gli utenti in generale.

[11] *WSO2* fornisce tecnologia open-source e soluzioni software e cloud per lo sviluppo di applicazioni e gestione di identità e accesso (IAM), https://wso2.com/ (cons. 25/01/2025).

4. TECNOLOGIE

Il principale elemento di innovazione introdotto nel progetto del MarketPlace consiste nell'offerta agli utenti di strumenti di **aggregazione operativa** tra i servizi individuali (primi tra tutti i servizi pilota sviluppati nell'ambito del WP7 del progetto), disponibili a diversi livelli nella piattaforma, in **servizi complessi** (flussi di lavoro attivabili dal MarketPlace ed eseguiti trasversalmente nelle piattaforme delle diverse infrastrutture). Grazie all'**orchestrazione dei servizi digitali**, il MarketPlace consente di integrare risorse distribuite e di facilitare la collaborazione tra ricercatori, istituzioni (pubbliche e private) e cittadini. Questo risultato si ottiene attraverso un'infrastruttura basata su strumenti di gestione delle API e di integrazione[12] come **WSO2 API Manager** e **WSO2 Micro Integrator**, che insieme garantiscono un sistema scalabile, interoperabile e sicuro (Sichera et al., 2024).

Il ruolo delle API e dei relativi livelli per far interagire e dialogare proficuamente sistemi anche assai disomogenei è dunque decisivo. La tecnologia garantisce pertanto l'interoperabilità, l'integrazione, il dialogo e lo scambio fra sistemi, servizi e risorse molto diversificati. È utile ribadire che l'adozione di un protocollo leggero ed efficiente come OAI-PMH,[13] per l'harvesting, l'estrazione e lo scambio di metadati tra sistemi differenti, permette la creazione di un ecosistema di ricerca nativamente interconnesso, aperto e inclusivo. Nel MarketPlace si integrano e si aggregano con facilità servizi, sistemi e componenti applicativi assai eterogenei, che tuttavia mantengono, nella pluralità del contesto, la loro identità peculiare, grazie ad un modello strategico imperniato su ampliamenti graduali.

4.1 WSO2 API Manager

Il MarketPlace sfrutta l'orchestrazione delle API per coordinare servizi provenienti da fonti diverse, armonizzandoli in un'unica interfaccia accessibile. In questo contesto, **WSO2 API Manager** svolge un ruolo centrale, fungendo da hub per la gestione e l'esposizione delle API. Grazie alla possibilità di creare API Proxy, WSO2 API Manager semplifica l'interazione con servizi sottostanti, nascondendo complessità infrastrutturali e garantendo standardizzazione. WSO2 API Manager permette di configurare flussi di mediazione avanzati, in cui i dati vengono trasformati e adattati alle necessità specifiche dei vari utenti. Questo è particolarmente rilevante in ambiti interdisciplinari, dove ricercatori provenienti da diverse discipline possono utilizzare gli stessi servizi per scopi differenti, senza la necessità di adattamenti manuali.

4.2 WSO2 Identity Server

L'aspetto inclusivo del MarketPlace è rafforzato dall'utilizzo di **WSO2 Identity Server,** che garantisce un accesso sicuro e personalizzato. Attraverso protocolli di autenticazione standard come OAuth2, gli utenti possono accedere ai servizi utilizzando le proprie credenziali istituzionali o di terze parti. Questa soluzione non solo riduce le barriere tecniche, ma promuove la partecipazione di una comunità eterogenea, rendendo il MarketPlace uno strumento realmente inclusivo per ricercatori, studenti e cittadini. La gestione dei diritti d'accesso basata sui ruoli assicura che le risorse siano protette e accessibili in modo controllato, favorendo un ambiente di collaborazione sicuro e rispettoso della privacy.

4.3 WSO2 Micro Integrator

L'integrazione dei dati e l'automazione dei processi sono supportate dal **WSO2 Micro Integrator**, che consente di orchestrare flussi di lavoro complessi e di collegare servizi eterogenei. Questo componente offre strumenti avanzati per semplificare la creazione di pipeline di integrazione, sfruttando un wizard online oppure un plugin dedicato per **Visual Studio Code** che permette di progettare e gestire flussi direttamente in un ambiente di sviluppo familiare (anche con l'aiuto dell'**AI Copilot**) come esemplificato nella Figura 1; questi strumenti aiutano gli utenti nella configurazione di integrazioni complesse attraverso un'interfaccia intuitiva, facilitando la creazione di pipeline per la trasformazione e il trasferimento dei dati tra diverse fonti.

[12] Per una discussione sul ruolo delle API in ambienti *multi-cloud* scalabili cfr. Putta et al., 2024 e Kondam, 2024.

[13] OAI-PMH *Open Archives Initiative Protocol for Metadata Harvesting* https://www.openarchives.org/pmh/

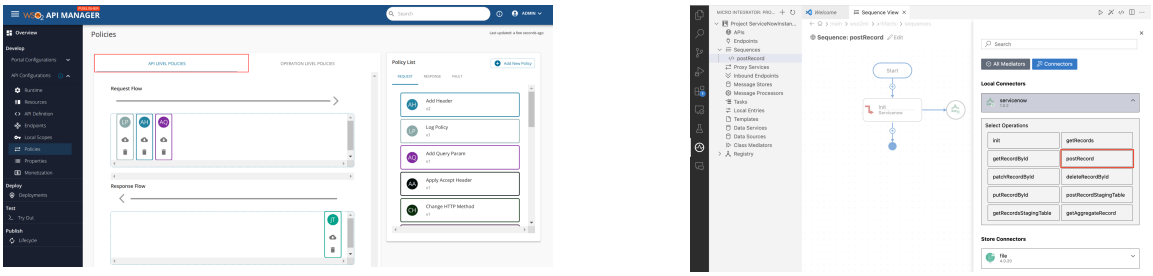


Figura 1. Wizard per orchestrazione online e su Visual Studio Code

Grazie a queste funzionalità, WSO2 Micro Integrator consente di combinare dati provenienti da fonti eterogenee, come archivi bibliografici, strumenti di analisi semantica e database storici, per restituire un output unificato e coerente. Questo lo rende uno strumento particolarmente adatto a contesti interdisciplinari, dove la necessità di aggregare informazioni da risorse diverse, su cui compiere operazioni anche di carattere eterogeneo, è fondamentale. WSO2 Micro Integrator si distingue per la gestione di flussi asincroni, che permettono di elaborare grandi volumi di dati o di sincronizzare risorse distribuite in contesti collaborativi internazionali, garantendo elevata scalabilità e affidabilità.

5. CASI D'USO DEL MARKETPLACE E DEL SISTEMA DI ORCHESTRAZIONE

Un esempio concreto può aiutare a comprendere il funzionamento inclusivo del MarketPlace. Si consideri il caso di un/una ricercatore/trice che lavora a un'edizione digitale di manoscritti storici. Dopo aver fairificato (cioè reso conformi agli standard FAIR) i propri dati – ad es. immagini ad alta risoluzione e file XML-TEI[14] con le trascrizioni – l'utente carica o seleziona tali materiali all'interno del MarketPlace. Qui può inviarli automaticamente a una piattaforma per la creazione di edizioni digitali.[15] Tutte le operazioni di autenticazione, trasferimento e orchestrazione – compresi eventuali passaggi di formati o chiamate asincrone a servizi esterni – sono gestite in modo trasparente dal MarketPlace. Il risultato è che lo studioso riceve, in un unico flusso di lavoro, i testi trascritti o la bozza dell'edizione digitale, senza doversi preoccupare di configurare gli strumenti o di capire come scambiare i dati tra diverse piattaforme. Questo approccio, oltre a valorizzare il potenziale dei singoli servizi, abbatte barriere di natura tecnica, rendendo la ricerca più agevole e inclusiva.

6. CONCLUSIONI

Gli elementi innovativi del MarketPlace che abbiamo descritto servono ad assicurare una migliore esperienza nell'uso della piattaforma e offrire un maggiore supporto alla cittadinanza digitale. La combinazione da parte degli utenti del MarketPlace di servizi preesistenti in nuovi servizi (in certi casi anche a valore aggiunto) sempre pubblicati sul MarketPlace, porta a creare nuovi flussi di lavoro e quindi nuovi metodi di ricerca. L'orchestrazione delle API consente, in tal modo, non solo di rendere più lineari attività complesse di ricerca, ma anche di monitorare l'utilizzo dei servizi e le loro concatenazioni in tempo reale, raccogliendo dati preziosi per comprendere come le risorse digitali vengono sfruttate e per identificare i margini di miglioramento e le esigenze della comunità scientifica.[16] Questo approccio, se combinato con l'apertura e il libero accesso ai dati del monitoraggio, favorisce una gestione più efficiente delle risorse e una maggiore trasparenza e responsabilità, elementi fondamentali per costruire un rapporto di fiducia con la comunità scientifica e con il pubblico.

Il progetto del MarketPlace, definito e affinato da OPERAS in stretto dialogo con le altre infrastrutture del cluster di H2IOSC, intende dimostrare come un'infrastruttura tecnologica possa sostenere in modo trasparente, equo e non discriminatorio, l'interdisciplinarità e la collaborazione, contribuendo a costruire un

[14] XML: Extensible Markup Language (cfr. https://www.w3schools.com/xml/) - TEI: Text Encoding Initiative (cfr. https://tei-c.org/).

[15] Il servizio di fairificazione dei dati e la piattaforma per la creazione di edizioni digitali di testi a stampa sono oggetto di sviluppo di due pilot del WP7 Community pilots: innovative cross domain services and environment di H2IOSC: 7.10a Design, development and implementation of a fairification workflow for the contents of the H2IOSC project and realisation; 7.10b Development of a web application for the management of multi-tenant and multi-user digital libraries.

[16] Questa attività si rifà anche a dati e indicazioni tratte dal lavoro di *Landscaping and building communities* portato avanti nel WP2 del progetto.

ecosistema di ricerca più aperto e inclusivo. L'apporto prioritario di OPERAS corrisponde all'impegno per la scienza aperta che qualifica l'infrastruttura ed è pienamente in linea con il suo caratteristico sostegno alla pubblicazione scientifica aperta nelle scienze umane e sociali. Ciò rappresenta, dal nostro punto di vista, un passo in avanti per promuovere l'interoperabilità tra le risorse digitali e la loro fruibilità scientifica e culturale, rendendole un fattore di crescita della cittadinanza digitale. La sinergia tra *WSO2 API Manager*, *WSO2 Identity Server* e *WSO2 Micro Integrator*, integrata in maniera tecnologicamente corretta e avanzata nel MarketPlace, è una soluzione di eccellenza tecnologica e, in combinazione con il cluster di infrastrutture, un modello virtuoso per il futuro delle piattaforme digitali per la ricerca, dimostrando che l'innovazione non risiede solo nelle tecnologie utilizzate, ma nella loro combinazione per un progetto ambizioso, che abbia il potenziale di mettere le persone, i gruppi di ricerca e il pubblico al centro della conoscenza e della partecipazione attiva.

BIBLIOGRAFIA

Ancion, Z., Borrell-Damián, L., Mounier, P., Rooryck, J., & Saenen, B. (2022). *Action Plan for Diamond Open Access*. Zenodo. https://doi.org/10.5281/zenodo.6282403 (cons. 25/01/2025)

Becerril, A., Bosman, J., Bjørnshauge, L., Frantsvåg, J. E., Kramer, B., Langlais, P.-C., Mounier, P., Proudman, V., Redhead, C., & Torny, D. (2021). *OA Diamond Journals Study. Part 2: Recommendations*. Zenodo. https://doi.org/10.5281/zenodo.4562790 (cons. 25/01/2025)

Bosman, J., Frantsvåg, J. E., Kramer, B., Langlais, P.-C., & Proudman, V. (2021). *OA Diamond Journals Study. Part 1: Findings*. Zenodo. https://doi.org/10.5281/zenodo.4558704 (cons. 25/01/2025)

Consortium of the DIAMAS project. (2024). *The Diamond OA Standard (DOAS) - version 1.1*. Zenodo. https://doi.org/10.5281/zenodo.12179620 (cons. 25/01/2025)

Council of the European Union, General Secretariat. (2023, May). *Draft Council conclusions on high-quality, transparent, open, trustworthy and equitable scholarly publishing*. https://data.consilium.europa.eu/doc/document/ST-8827-2023-INIT/en/pdf (cons. 25/01/2025)

Cristofaro, S., Fabiani, V., Marras, C., Pasini, E., Sichera P., & Yu, M. (2024). Infrastrutture di ricerca come strumenti di "interculturalità digitale". In A. Di Silvestro & D. Spampinato (a cura di), *Me.Te. Digitali. Mediterraneo in rete tra testi e contesti, Proceedings del XIII Convegno Annuale AIUCD2024. Quaderni di Umanistica Digitale* (pp. 572-576). AIUCD. DOI 10.6092/unibo/amsacta/7927 (cons. 25/01/2025)

Kondam, A. (2024, October 10). Api Gateway Orchestration In Multi-Cloud Environments: Challenges And Solutions. *International Journal of Creative Research Thoughts*, 12, 596-602. https://www.researchgate.net/publication/385345794_Api_Gateway_Orchestration_In_Multi-Cloud_Environments_Challenges_And_Solutions (cons. 25/01/2025)

Mittal, D., Mease, R., Kuner, T., Flor, H., Kuner, R., & Andoh, J. (2023). Data management strategy for a collaborative research center. *GigaScience*, 12, giad049. https://doi.org/10.1093/gigascience/giad049 (cons. 25/01/2025)

Putta, N., Dave, A., Balasubramaniam, V., Prasad, M., & Kumar, S. (2024). Optimizing Enterprise API Development for Scalable Cloud Environments. *Journal of Quantum Science and Technology (JQST),1(*3), Special Issue July-Sept 2024, 229-246. https://www.researchgate.net/publication/387523300_Optimizing_Enterprise_API_Development_for_Scalable_Cloud_Environments (cons. 25/01/2025)

Sichera, P., Marras, C., & Pasini, E. (2024, 19 novembre). *Orchestrazione API per workflow applicativi nell'ambito delle Digital Humanities, Scientific-technical Report.* Zenodo. https://doi.org/10.5281/zenodo.14187534 (cons. 25/01/2025)

Special Committee on Artificial Intelligence in a Digital Age. (2021). *AIDA Working Paper on AI and Bias*. https://www.europarl.europa.eu/cmsdata/244798/Working%20Paper%20-%20AIDA%20Hearing%20on%20AI%20and%20Bias_.pdf (cons. 25/01/2025)

Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.* (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data,* 3, 160018. https://doi.org/10.1038/sdata.2016.18

A Case Study in Cultural Heritage: A System Linking Three Open Data Tools – Digital Philology for Dummies (DPhD), Edition Visualization Technology (EVT), and a Relational Database

Renato Caenaro[1], Chantal Pivetta[2]
[1] Silentwave SRLS, Italy – renato.caenaro@silentwave.eu
[2] Lund University, Sweden – chantal.pivetta@rom.lu.se

ABSTRACT (ENGLISH)

This contribution presents a practical case study focused on Cultural Heritage, specifically the study of Cassinese monasteries in Italy. The emphasis is on the methodologies and technologies used within an inclusive and open science framework. The project is inherently multidisciplinary, involving art history, architectural history, computer science, engineering, and philology, and aims to connect data across multiple platforms, including text visualizers, maps, relational databases, and engineering-architectural models.

The methodology is designed to accommodate the diverse expertise of the team, which includes members without technical backgrounds and from various academic disciplines. The presentation focuses on two key phases of the workflow: semantic annotation and encoding of documentary sources. The focus is on two platforms: Digital Philology for Dummies (DPhD) and Edition Visualization Tecnhology (EVT). Finally, complex analysis and interconnection of data.

The project establishes a robust data-sharing framework that promotes the early adoption of open data and open access principles. This ensures that data can be published immediately in alignment with open data and open access philosophies. By implementing structured open APIs, the data becomes accessible not only as complete database downloads but also as interconnected, linked data, thereby enhancing accessibility and application potential.

This methodology demonstrates how collaborative tools, interdisciplinary workflows, and linked data technologies can advance research in Cultural Heritage while ensuring inclusivity, accessibility, and open knowledge sharing.

Keywords: Open Data; XML/TEI; Interdisciplinary Research; Relational Database

ABSTRACT (ITALIANO)

Un caso studio per il patrimonio culturale: un sistema che collega tre strumenti Open Data - Digital Philology for Dummies (DPhD), Edition Visualization Technology (EVT), e un Database Relazionale

Questo contributo presenta un caso studio pratico, inerente al campo del Cultural Heritage, che ha come obbiettivo lo studio dei monasteri cassinesi in Italia. Si vuole illustrare il sistema e le tecnologie utilizzate, in un contesto inclusivo e orientato ai principi dell'open science. Il progetto è multidisciplinare e coinvolge esperti di storia dell'arte, storia dell'architettura, informatica, ingegneria e filologia, con lo scopo di connettere dati attraverso strumenti diversi, tra cui visualizzatori di testi, mappe, database relazionali e modelli ingegneristico-architettonici.

Il metodo è stato progettato considerando la provenienza multidisciplinare dei membri del team – costituito anche da risorse senza conoscenze informatiche – creando quindi un approccio strutturato che risponda alle diverse esigenze del gruppo. Il contributo si concentra su due fasi principali del flusso di lavoro: prima la codifica e annotazione delle fonti documentarie. Si evidenzierà il ruolo di Digital Philology for Dummies (DPhD) e Edition Visualization Technology (EVT). Segue l'analisi dei dati e loro connessione con altri sistemi.

Il progetto dispone di dati immediatamente pubblicabili e accessibili secondo i principi openData e openAccess. Grazie all'implementazione di openAPIs strutturate, i dati sono accessibili non solo come download completi, ma anche come dati collegati e interconnessi, ampliandone le possibilità di utilizzo e applicazione.

Questo approccio dimostra come l'uso di strumenti collaborativi, metodologie interdisciplinari e tecnologie di linked data possa migliorare la ricerca nei Beni Culturali, garantendo inclusività, accessibilità e una condivisione aperta della conoscenza.

Parole chiave: Open Data; XML/TEI; Ricerca Multidisciplinare; Database Relazionale

1. INTRODUCTION

The study of Cultural Heritage, particularly the historical and architectural significance of the Cassinese monasteries in Italy, has been increasingly enriched by the integration of digital methodologies and open science frameworks. This contribution presents a practical case study- a pivotal project- focusing on the methodologies and technologies employed to facilitate access to and the interlinking of Digital Humanities data resources, including text visualizations, geospatial mapping tools, relational databases, and engineering-architectural models. The outcomes enhance the inclusivity and efficiency of digitization, annotation, and data integration processes. Moreover, the study underscores how this approach supports the principles of open data, promoting long-term accessibility, sustainability, and interdisciplinary collaboration.

The significance of this pilot project and the resulting system lies in two key aspects: 1. The system is designed to be reusable in multidisciplinary projects involving large teams with diverse backgrounds, where contributions must be monitored and approved. 2. The system demonstrated how research interests may evolve after the initial data collection phase, offering a valuable perspective when confronted with a large volume of information to be processed. In the case study, the initial focus was on monasteries and the artworks they contained. However, following the preliminary results, the project's emphasis shifted toward documentary sources for the definition of the Entity-Relationship (ER) model, with a particular focus on Named Entities and their interrelations.

2. CASE STUDY

The PRIN project *"CoenoBIuM"* (www.digitalcoenobium.eu) – coordinated by Gianmario Guidarelli (University of Padua), in collaboration with Sonia Cavicchioli (University of Bologna) and Paolo Borin (University of Brescia) – investigates the architectural and artistic culture of the Cassinese Benedictine Congregation from its foundation through the late 18th century. It adopts a comparative perspective while experimenting with the innovative application of Building Information Modeling (BIM). The reform initiated by Ludovico Barbo in 1419 profoundly reshaped monastic life, redirecting its focus toward prayer and contemplation, and leaving a significant impact on architectural spaces and artistic expression within churches and monasteries.

Scholars have long hypothesized the existence of a shared architectural and visual culture across the Congregation, fostered by the common regulations issued by the Capitoli Generali. To test this hypothesis, the project aims to reconstruct the original relationships between decorative programs and built environments, integrating a wide range of heterogeneous sources (archival, iconographic, material, and spatial).

The pilot project presented here focuses on the development of a system that enables the interconnected digitization of information across multiple disciplines. The system includes the design of an Entity-Relationship (ER) schema, the implementation of a relational database, the development of a data entry portal, the publication of open APIs, the integration of authoring software and text visualizator.

REQUIREMENTS AND CHALLENGES

This paper addresses only a small portion of the complex system that was developed. The main requirements and challenges encountered during its development were:

1. The need to account for practical constraints, such as the cost and time availability of the participants involved in the project.
2. The varying levels of digital literacy among participants necessitated the use of user-friendly technological tools to ensure satisfactory task completion.
3. The imperative to guarantee interoperability and long-term sustainability of the data produced through the various technological tools adopted.

METHODOLOGIES EMPLOYED

The methodology is structured around two key phases: 1. Semantic Annotation and Encoding: The project use TEI guidelines to transcribe, annotate, and encode documentary sources collaboratively. A user-friendly, shared environment allows participants from diverse fields and technical skill levels to contribute effectively. The resulting structured data is stored in a relational database, enabling systematic organization and query-based retrieval. 2. Complex Data Analysis and Integration: Structured data is used to build tools that enable relational and geospatial analysis, linking encoded sources to external systems such as mapping platforms. This integration supports the creation of new insights, facilitating connections between textual, architectural, and geographical data. Throughout both phases, a semi-structured data

environment is employed, which supports the semi-automatic generation of relationships between datasets from various disciplines. OpenAPIs ensure that the data can be accessed as both complete database dumps and interconnected linked data.

3. FIRST STEP: ENCODING AND TRANSCRIBING IN XML/TEI

The workflow's first phase focuses on transcribing and encoding documentary sources using *Digital Philology for Dummies* (DPhD). This user-friendly and open-data software is designed to simplify and accelerate the transcription and encoding of texts following TEI guidelines. It is specifically created to assist users without technical expertise or knowledge of XML/TEI while offering advanced functionality for experienced users.

In the project's initial phase, *Digital Philology for Dummies* (DPhD) was chosen for two primary reasons. First, the group of transcribers was highly heterogeneous, composed of participants with no prior experience in XML/TEI encoding and representing a range of disciplinary backgrounds. As illustrated in the figure below, the interface is intuitive and accessible even for users unfamiliar with TEI. It allows the image and transcription to be displayed side by side, with the corresponding XML/TEI encoding automatically generated in real time. Semantic annotation is performed by selecting the relevant portion of text and applying predefined tags. These tags are standardized across all users and curated by more experienced contributors.

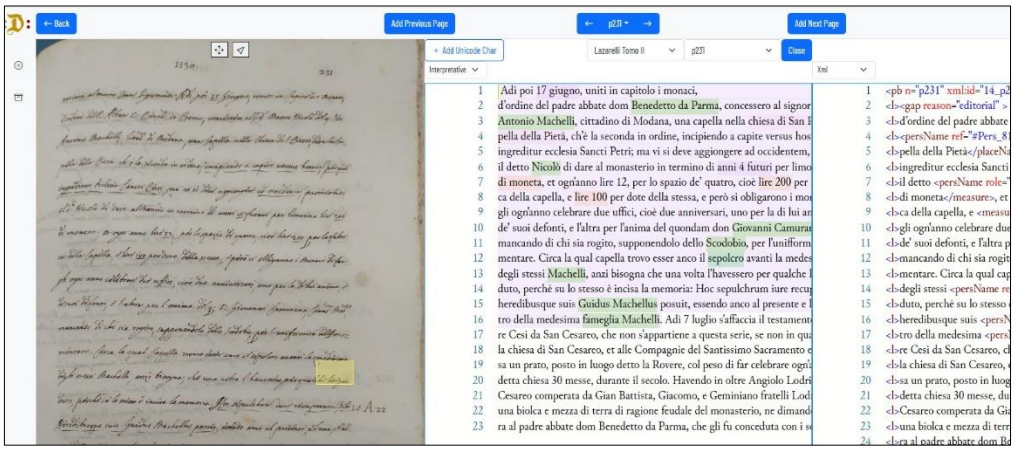


Figure 1 DPhD Interface

Second, a shared workspace was needed, in which all transcribers could collaborate following a unified set of rules to ensure consistency and standardization. Within the workspace, each transcriber is assigned a specific role by the project coordinator. The activities carried out by the transcribers are summarized in the following figure:

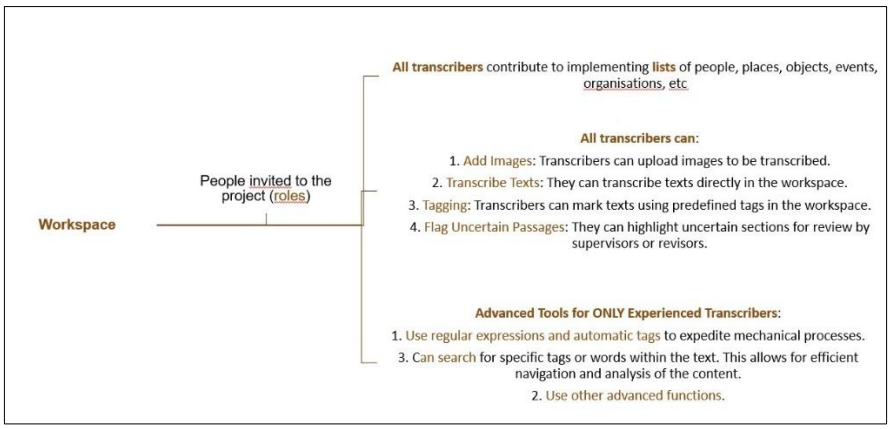


Figure 2 Transcription workspace

A critical step in the process, successfully managed, is the creation of unique identifiers (@xml:id), which is managed by the relational database. The database generates these identifiers through the dynamic creation of APIs, which continuously update the system. This workflow is highly interactive and iterative: transcribers input textual data into the platform and, in return, receive @xml:id values, which are

automatically integrated into the annotated texts. One example: among the main research questions, one focused on the study of artworks preserved within the monasteries—specifically, identifying where within the monastery these works were conserved, whether they were relocated to other monasteries over time, and how their context evolved. For this purpose, the artworks were identified and annotated using XML/TEI encoding. When a transcriber encountered a reference to a work of art in the text, they would input its metadata into a dedicated data entry form linked to the relational database (see figure 3).

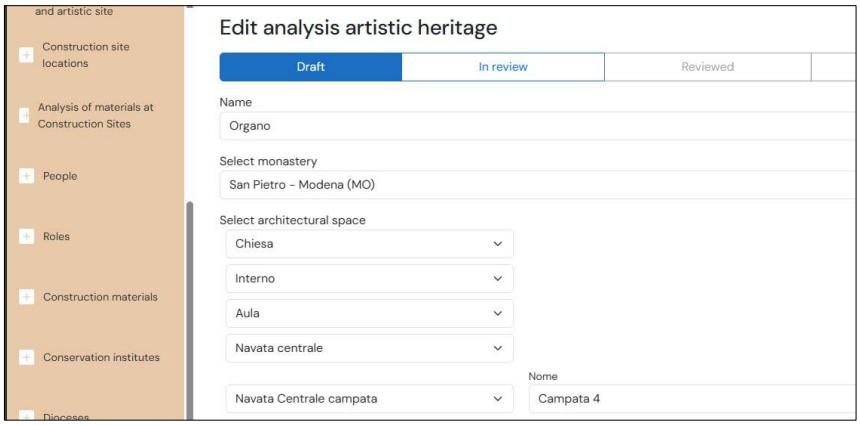


Figure 3 Coenobium Data entry for an organ

The database would then generate a unique @xml:id, which the transcriber would insert into the corresponding markup within the annotated text (see figure 4).

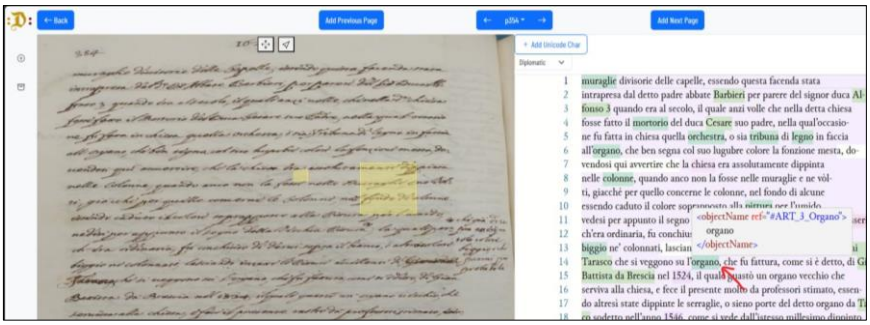


Figure 4 DPhD transcription of an organ

The database, in turn, links each artwork to additional elements, such as the location where it is preserved, the dates on which it is mentioned in documentary sources, its commissioner, cost, materials used, and other relevant attributes. This information is accessible also from the 3D model selecting the image:



Figure 5 3D model of an organ

Finally, regarding the visualization of these texts (in progress), another advantage of using DPhD is that the software is directly integrated with Edition Visualization Technology (EVT), enabling a smooth workflow that combines transcription, encoding, and visualization in a seamless process. DPhD uses a new feature

called 'EVT as a service,' enabling the real-time display of XML/TEI-encoded documents directly within EVT, without requiring installation or configuration. This feature simplifies its use for viewing general textual data, even beyond the specific context of digital scholarly editions Edition Visualization Technology (EVT) is an open data and open source platform specifically designed for visualizing and navigating texts encoded in XML/TEI. The software has reached its third version, which focuses on synoptic editions—a feature that has proven to be particularly effective for this project. The synoptic view allows the visualization of a diverse collection of texts in multiple languages, enabling users to see how they relate to one another. This capability is essential for addressing the various queries posed by the study of Cassinese monasteries, where the ability to navigate and analyze relationships between texts is critical to uncovering insights. This makes EVT an ideal tool for handling the complexity and heterogeneity of the project's textual data.

4. DATA ENTRY: WHY NOT ONLY RDF?

Concerning data entry, a hybrid system was adopted. While RDF excels at representing data with strong semantic structure, its use can be excessive in application domains that require simpler modeling or where semantic complexity is only marginal. In such cases, more traditional approaches—such as relational databases—may prove more appropriate and efficient.

Although RDF is designed to promote interoperability and data sharing, integrating data entry tools developed across different disciplinary domains into an RDF-based framework would have required significant customization. We also considered the perspective of the end users of the platform—namely, the data entry operators—who might not feel comfortable with the input modalities imposed by an RDF system, particularly when interfaces must conform to complex metadata structures. For this reason, we prioritized offering a more guided and user-friendly experience.

Finally, from a cost perspective, developing and maintaining a system based purely on RDF would have been significantly more expensive due to the need for specialized tools, advanced technical expertise, greater oversight of data input, and highly customized integration processes.

5. SYSTEM INTEROPERABILITY

The *Coenobium* project, as mentioned before, involved a range of disciplinary fields, which can be broadly categorized into four main areas: architectural history, engineering, art history, and philology. Regardless of the number or type of disciplines involved, each field brings its specificities, distinct research objectives, and, within the context of the Digital Humanities, specialized digital tools. It was therefore essential to develop a system capable of interconnecting and integrating the information produced by each discipline. To this end, an Entity-Relationship (ER) schema was designed to define the core entities relevant to the project, along with their relationships and constraints. Based on this schema, a relational database (RDB) was implemented to enable the structured mapping of entities and their interrelations.

A data entry portal was also developed, accessible to all project participants, allowing for the guided registration of entities through a structured interface. The system includes attribution of entered data and a workflow comprising review and approval phases.

In parallel, a set of open APIs was created to deliver the data stored in the RDB in various formats (XML, XML-TEI, JSON). Each software tool used within the different disciplinary domains can query these APIs to retrieve the information needed for generating project outputs, which may take diverse forms such as XML-TEI files, PDFs, images, 3D models (IFC, OBJ-GLTF), point clouds, and others.

Finally, a shared data environment was established, where each output is collected and becomes an extension of the RDB. This infrastructure allows users, within their respective software environments, to access information produced by other disciplines—even without direct knowledge of their tools or datasets—since the content is made available through structured open APIs based on a common schema.

6. EXAMPLE OF ENTITY RELATIONSHIP SCHEMA

As an example, we present below a subset of the Entity-Relationship (ER) schema from the DBR database, which illustrates the connections between the Named Entities <people> and those identified as <object type="MonasteryComplex"> and <event type="ConstructionSite"> within the textual sources.

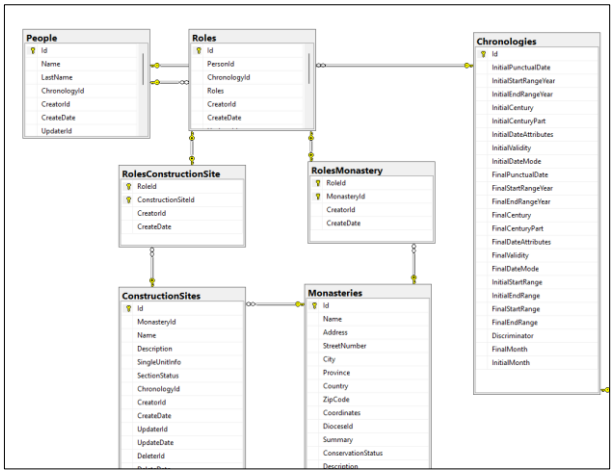


Figure 6 Portion of ER Schema

As can be observed, the relationships between people, construction sites, and monasteries are conceptually structured as RDF-like triples. However, in practice, the attribute corresponding to the "role" within each triple has been externalized into a dedicated table, as it is linked to a temporal dimension. This chronological component allows for the specification of start and end dates, which can be expressed as precise dates, intervals, or centuries, and includes the option to assign a degree of certainty to each temporal statement.

This schema plays a crucial role in the data entry phase, as it enables not only the normalization of data but also the validation of input. Furthermore, a public HTTP-based openAPI interface is provided, both for project participants and for the software systems that utilize the data. This API also serves external users, allowing for the mapping of these complex entities into a more canonical form that is compatible with SPARQL-based querying environments.

7. RESULTS

The outcome is a sustainable and interoperable system, capable of integrating large volumes of data from multiple disciplinary domains. Importantly, it also allows individuals without advanced digital expertise to contribute meaningfully to the enrichment of the project. Moreover, the system is replicable and adaptable to other initiatives involving architectural heritage and documentary sources, whether historical or contemporary.

More specifically, the methodology has already demonstrated significant potential in:

1. Supporting advanced textual and relational analysis through the use of structured data encoded in TEI and managed within relational databases.
2. Improving the accessibility and usability of historical data through visualization tools such as EVT and geospatial mapping platforms.
3. Facilitating collaborative workflows that include participants with varying levels of technical expertise.
4. Promoting the principles of open science and data interoperability by publishing data in alignment with open data and open access frameworks.

REFERENCES

Andrews, T. L. (2013). The Third Way: Philology and Critical Edition in the Digital Age. *The Journal of the European Society for Textual Scholarship*, *10*, 61–76. https://doi.org/10.1163/9789401209021_006

Baker C., Cheung K.-H. (eds) (2007). Semantic Web—Revolutionizing Knowledge Discovery in the Life Sciences. Berlin/Heidelberg: Springer.

Bleier, R., Bürgermeister, M., Klug, H. W., Neuber, F., & Schneider, G. (Eds.). (2018). *Digital Scholarly Editions as Interfaces* (Vol. 12). BoD. https://kups.ub.uni-koeln.de/9085/

Burghart, M. (2017). *Creating a Digital Scholarly Edition with the Text Encoding Initiative*. https://halshs.archives-ouvertes.fr/halshs-01984319

Cheng G., Shao F., Qu Y. (2017). An empirical evaluation of techniques for ranking semantic associations. IEEE Transactions on Knowledge and Data Engineering, 29(11): 1.

Ciotti, F., Corradini, E., Cugliana, E., D'Agostino, G., Ferroni, L., Fischer, F., Lana, M., Monella, P., Roeder, T., Turco, R. R. D., & Sahle, P. (2022). Manifesto per le edizioni scientifiche digitali. *Umanistica Digitale*, *12*, Article 12. https://doi.org/10.6092/issn.2532-8816/14814

Di Pietro, C., & Rosselli Del Turco, R. (2018). Between Innovation and Conservation: The Narrow Path of User Interface Design for Digital Scholarly Editions. In *Digital Scholarly Editions as Interfaces* (Vol. 12, pp. 133–163). BoD. https://kups.ub.uni-koeln.de/9115/

Eero, H. Heikki, R. (2021). Knowledge-based relational search in cultural heritage linked data, Digital Scholarship in the Humanities, Volume 36, Issue Supplement_2, Pages ii155–ii164, https://doi.org/10.1093/llc/fqab042

Favretto, A. (2022). Relational Database, GIS Layers, and Geodatabase for Cultural Heritage Management In: D'Amico, S., Venuti, V. (eds) Handbook of Cultural Heritage Analysis. Springer, Cham. https://doi.org/10.1007/978-3-030-60016-7_46.

Hyvönen, E. (2020). Using the Semantic Web in Digital Humanities: shift from data publishing to dataanalysis and serendipitous knowledge discovery. Semantic Web, 11(1):187–93.

Hyvönen E. (2012). Publishing and Using Cultural Heritage Linked Data on the Semantic Web. Palo Alto, CA: Morgan & Claypool.

Indrawan-Santiago, M. (2012). Database research: Are we at a crossroad? Reflection on NoSQL. In 2012 15th International Conference on Network-Based Information Systems (pp. 45-51). IEEE.

Monella, P., & Rosselli Del Turco, R. (2020). Extending the DSE: LOD Support and TEI/IIIF Integration in EVT. In C. Marras, M. Passarotti, G. Franzini, & E. Litta (Eds.), *Atti del IX Convegno Annuale dell'Associazione per l'Informatica Umanistica e la Cultura Digitale (AIUCD). La svolta inevitabile: Sfide e prospettive per l'Informatica Umanistica* (pp. 148–155). Available online as a supplement of Umanistica Digitale: https://umanisticadigitale.unibo.it.
https://doi.org/10.6092/UNIBO/AMSACTA/6316

Robinson, P. M. W. (2022). An approach to complex texts in multiple documents. *Digital Scholarship in the Humanities*, fqab108. https://doi.org/10.1093/llc/fqab108

Rosselli Del Turco, R. (2019). Designing an advanced software tool for Digital Scholarly Editions: The inception and development of EVT (Edition Visualization Technology). *Textual Cultures*, *12*(2), 91–111. https://doi.org/10.14434/textual.v12i2.27690

Rosselli Del Turco, R., Buomprisco, G., Di Pietro, C., Kenny, J., Masotti,R., Pugliese, J. (2015). Edition Visualization Technology: A Simple Tool to Visualize TEI-based Digital Editions. *Journal of the Text Encoding Initiative*, *Issue 8*, 1–21. https://doi.org/10.4000/jtei.1077

Schreibman, S., & Papadopoulos, C. (2019). Textuality in 3D: Three-dimensional (re)constructions as digital scholarly editions. *International Journal of Digital Humanities*, *1*(2), 221–233.
https://doi.org/10.1007/s42803-019-00024-6

Tamper M., Leskinen P., Apajalahti K., Hyvönen E. (2018). Using biographical texts as linked data for prosopographical research and applications. In Digital Heritage. 7th International Conference on Progress in Cultural Heritage: Documentation, Preservation, and Protection, EuroMed 2018, Springer, Nicosia, Cyprus.

Modeling Intermediality and Interpretations in Contemporary Combinatory Literature: Revealing *Il Giuoco dell'Oca* by Edoardo Sanguineti

Enrica Bruno[1], Maria Francesca Bocchi[2], Francesca Tomasi[3]
[1] University of Bologna, Italy - enrica.bruno2@unibo.it
[2] University of Bologna, Italy - maria.bocchi4@unibo.it
[3] University of Bologna, Italy - francesca.tomasi@unibo.it

ABSTRACT (ENGLISH)
This paper investigates the application of Semantic Web technologies to the analysis of Edoardo Sanguineti's *Il Giuoco dell'Oca*, a significant example of combinatory literature. The study employs computational literary methods to model and analyze the intermedial relations within the novel.
The research introduces MIrO (**M**ultimedia **I**nter-**r**elation **O**ntology), a model that integrates established ontologies to represent intermedial relations, intratextual connections, and interpretative acts within the text. By making explicit and formalizing these relations, the study shows how computational strategies of text enrichment, resulting from a critical interpretation, can enhance the reading experience and understanding of literary texts.
Keywords: Semantic Web; Edoardo Sanguineti; Computational Literary Studies; Intermedial Relations.

ABSTRACT (ITALIANO)
Modellare l'Intermedialità e le Interpretazioni nella Letteratura Combinatoria Contemporanea: una Lettura de "Il Giuoco dell'Oca" di Edoardo Sanguineti
Il presente contributo indaga l'applicazione delle tecnologie del Web Semantico all'analisi de *Il Giuoco dell'Oca* di Edoardo Sanguineti, un esempio significativo di letteratura combinatoria. Lo studio impiega metodi letterari computazionali per modellare e analizzare le relazioni intermediali all'interno del romanzo.
La ricerca introduce MIrO (**M**ultimedia **I**nter-**r**elation **O**ntology), un modello che integra ontologie consolidate per rappresentare le relazioni intermediali, le connessioni intratestuali e gli atti interpretativi all'interno del testo. Esplicitando e formalizzando queste relazioni, lo studio mostra come strategie computazionali di arricchimento del testo, derivanti da un'interpretazione critica, possano migliorare l'esperienza di lettura e la comprensione dei testi letterari.
Parole chiave: Web Semantico; Edoardo Sanguineti; Letteratura Computazionale; Relazioni Intermediali.

1. INTRODUCTION
This paper explores the use of Semantic Web technologies in the Digital Humanities (DH) field to study and analyze contemporary literary texts. It particularly focuses on combinatory literature, which is well-suited for description and modeling through computational methods. These methods enable a deeper understanding of a text by focusing on information extraction, knowledge organization, and data reconciliation. Furthermore, using Linked Open Data (LOD) as a publishing approach offers new pathways to challenge copyright restrictions.
For our case study, we selected the novel *Il Giuoco dell'Oca* by Edoardo Sanguineti. Originally published in 1967 (Sanguineti, 1967), this work was born in the context of the neo-avantgarde linked to the *Gruppo '63* (Curi, 2006) and it is considered an example of combinatory literature (Portesine, 2021: 22-26) for its unique structure and content. The title itself suggests that the author draws inspiration from the traditional game of the goose, where each chapter is like a box of the board game, conveying a ludic dimension within the literary text typical of combinatorial and Oulipian[1] literature (Portesine, 2021: 22; Corosaniti, 2024: 3; Albani, 2018). This structure allows the reader to approach the novel sequentially, moving from the first chapter to the last, or to jump between chapters. In this way, the reader is invited to interpret and collaborate in constructing the narrative (Eco, 1967).

[1] The OuLiPo (Ouvroir de Littérature Potentielle) was founded in 1960 by the engineer and mathematician François Le Lionnais and the writer Raymond Queneau. Its Italian counterpart, OpLePo (Opificio di Letteratura Potenziale), was established in 1990. Sanguineti joined OpLePo in 1998, becoming its president. In this context, *Il Giuoco dell'Oca* is considered an *Oplepian* text *ante litteram*, anticipating the group's principles before its actual foundation (Oplepo, 2005). For an overview of both groups, please refer to the following websites: https://oulipo.net/fr and https://www.oplepo.com/.

A key element of the novel is the use of the rhetorical strategy of the ekphrasis, as each chapter corresponds to a description of a cultural object. These objects primarily relate to the culture of the 1960s, the period in which Sanguineti lived and worked. The use of ekphrasis, however, challenges traditional forms of description by juxtaposing images and emphasizing specific aspects of a variety of items such as works of art, advertisements, films, and songs, collectively evoking the image of a "Biennale" or art gallery (Portesine, 2021). Scholars identified the majority of such items, enriching the comprehension of Sanguineti's literary and intellectual work (Portesine, 2021; Risso, 2023; Sotgiu, 2016).
Consequently, an important aspect of the text is the presence of different kinds of relations that can be articulated into a multi-relational model: (i) the relation connecting a chapter's ekphrasis and the cultural object described within; (ii) the intratextual relations between chapters sharing the description of the same cultural object; (iii) the connection between objects of different media types.
In this talk, we introduce MIrO (**M**ultimedia **I**nter-**r**elation **O**ntology), a formal model designed to integrate various ontologies used in the field of cultural heritage (i.e. LRMoo, RDA, CIDOC-CRM, SKOS, HiCO, and CiTO). MIrO aims to represent the domain of intermedial relations (Rajewsky, 2011) alongside the description of cultural objects related to the novel, including *Il Giuoco dell'Oca* itself, and the interpretation acts of various scholars who have identified and analyzed the different types of relations within Sanguineti's work.

2. STATE OF THE ART

The scholar (Risso, 2023) highlights that the distinctive structure of the novel fosters an immersive virtual environment. In this context, traditional linear reading is transformed into a set of narrative possibilities, creating a mosaic of potential combinations (Calvino, 2015: 212-217). This aspect of Sanguineti's work invites an examination of computational methods' role in literature and their significance in narrative construction. Nevertheless, to the authors' knowledge, *Il Giuoco dell'Oca* has received little attention in the field of Digital Humanities, except for the recent project *Inside the Black Box* (Corosaniti, 2024), a multimedia installation that exploits Artificial Intelligence for a ludic-immersive experience of the gamebook. However, analyzing the structure of the work can enhance our understanding of its combinatorial intratextual aspects and the nature of verbal description concerning the modern concept of ekphrasis (Cometa, 2012; Risso, 2023).
With the rise of Semantic Web technologies as a standard for organizing, representing, and sharing cultural heritage, two main approaches have emerged in literary studies. The first approach focuses on creating semantic editions that facilitate textual analysis and navigation, emphasizing inter- and intratextual connections within literary works. Some examples include the digital edition of Paolo Bufalini's Notebooks[2] (Daquino, 2019), the digital edition of Vespasiano da Bisticci's letters[3] (Tomasi, 2020a) and OntoPavese (D'Agata, 2024). The second approach seeks to move beyond traditional methods of digital scholarly editing by employing ontological models and linked data (Meghini, 2021). Applied to literary works, this approach allows for a deeper understanding of the unique characteristics of texts, facilitates the generation of new semantic data, and proposes semantic narratives helping to overcome copyright limitations in contemporary literature (Bruno, 2024). Regarding the formalization of intertextuality and intermediality within the cultural heritage, the OWL2 ontology INTRO represents a promising attempt to represent media features and their interconnections (Oberreither, 2024). However, while INTRO represents a valuable initial step, it remains at a general level of formalization of the possible interrelations and does not yet offer the granularity needed to capture the specific types of complex semantic relationships, such as those present within *Il Giuoco*.

3. METHODOLOGY

To develop our methodology, we follow the approach for knowledge organization and representation in DH projects (Tomasi, 2020b). We began by selecting a domain of interest (Edoardo Sanguineti's novel) and conducting a selection of cultural objects (step 1). We then created various models, from a conceptual map based on real data to a theoretical model (step 2). We retrieved metadata for the selected objects from the web, addressing the gaps in contemporary cultural heritage (step 3). Finally, we aligned proprietary schema metadata with international standards to enhance data descriptions and reusability (step 4). The final step involved creating a conceptual model based on the theoretical map. This model

[2] https://projects.dharc.unibo.it/bufalini-notebook/.
[3] https://projects.dharc.unibo.it/vespasiano/.

integrates standard ontologies and extends it with new classes and properties. We managed the domain and all items using the WEMI[4] structure (Work, Expression, Manifestation, Item) from LRMoo (Library Reference Model[5]) to detail the lifecycle of objects.
We produced a dataset based on the conceptual model, which includes a selection of 14 cultural objects[6], including *Il Giuoco dell'Oca* and two other cultural objects related to the novel, even though they are not mentioned in the chapters. These include the board game *The Game of the Goose* and Gianfranco Baruchello's *Il Giuoco dell'Oca di Edoardo Sanguineti*, which reinterprets the traditional board game in the light of Sanguineti's work. Each cultural object is described alongside its metadata, and the dataset also includes claims made by scholars regarding the relations among the objects.
The dataset has been converted into an RDF graph. Then, a few competency questions were defined to test the ontology's representativeness and the data's correctness (see section 4).

3. FORMAL MODEL

The model formalizes the requirements mentioned in section 1: (i) the description of the cultural objects related to the novel, including *Il Giuoco dell'Oca*; (ii) the intermedial relations, both the ones concerning each item and the chapter where it is described (emphasizing the intratextual aspects related to the description of the same item across multiple chapters) and the inter-relations between different items related to the novel; (iii) the interpretation acts of various scholars. To achieve these requirements, various existing schemas, vocabularies, and ontologies were reused.
Concerning the first requirement, the WEMI structure of LRMoo and SKOS (Simple Knowledge Organization System[7]) was chosen to model *Il Giuoco dell'Oca* and all the selected items. Additionally, the RDA standard (Resource Description and Access[8]) was reused for the associated metadata[9].
Each item is described in its possible implemented entity in WEMI (`lrmoo:F1_Work`, `lrmoo:F2_Expression`, `lrmoo:F3_Manifestation`, `lrmoo:F5_Item`) to which the relative alignment with the RDA classes corresponds (`rdac:C10001`, `rdac:C10006`, `rdac:C10007`, `rdac:C10003`). Metadata were chosen based on the different WEMI levels, representing some of the categories identified during the scouting phase (see section 3). Additionally, all the agents responsible for the creation of (`rdaw:P10065`, `rdam:P30329`) or contribution to (`rdae:P20328`) the cultural object were specified with the corresponding role through the possible sub-properties of the main RDA properties.
Concerning the modelization of the novel, we considered each chapter to be an instance of the class `lrmoo:F2_Expression` with a fixed text content type (`rdaco:1020` "`text`"). Considering the text content at the Expression level reflects how accurately the author describes cultural objects (Portesine, 2021: 19). This approach allows us to engage with the cultural objects on a higher level of abstraction.

[4] WEMI is a conceptual structure originally introduced in FRBR (Functional Requirements for Bibliographic Records) to describe the relationships between different levels of bibliographic information. FRBR was later integrated into LRMoo, the object-oriented implementation of the IFLA Library Reference Model (LRM), which refined and extended the conceptual model. For the full documentation of FRBR see: https://repository.ifla.org/items/54925d49-b08d-4aeb-807c-1b509ec40b55.
[5] https://repository.ifla.org/items/94aedb49-2d6e-4a6d-9974-f33abb7e3c0e.
[6] A complete overview of the dataset is available at the following link: https://github.com/enricabruno/miro.
[7] https://www.w3.org/TR/skos-reference/.
[8] https://www.rdaregistry.info/.
[9] The formal model representing the subdivision of the domain into WEMI levels and its integration with LRMoo, RDA, and SKOS can be seen at the following link: https://github.com/enricabruno/miro/tree/main/images.

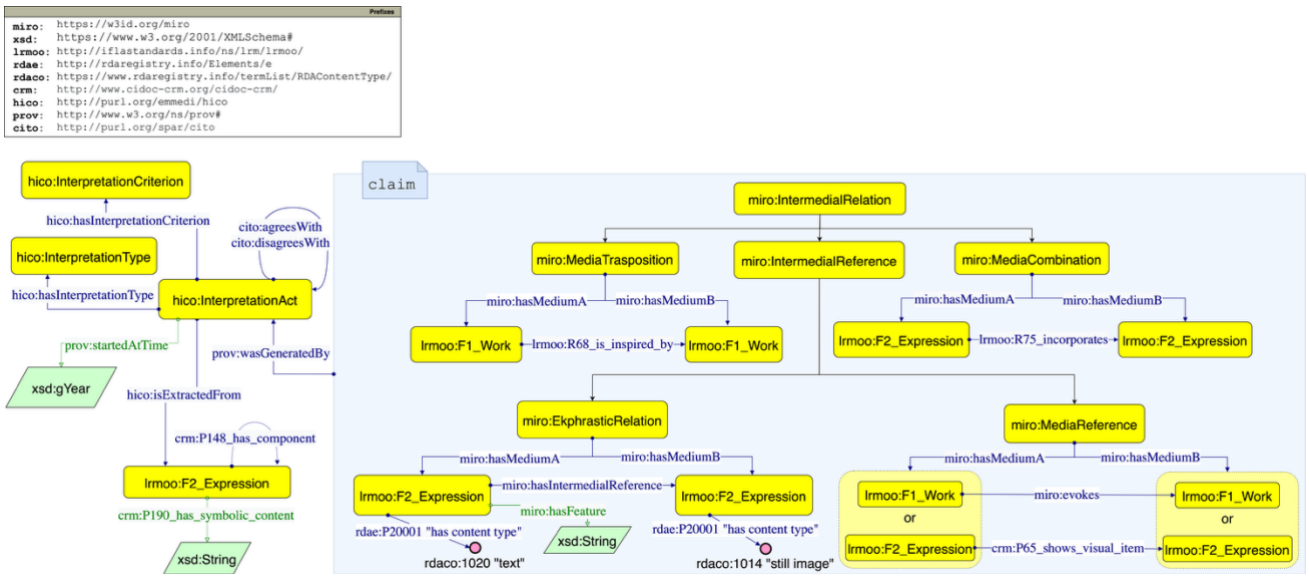


Figure 1. The formalization of the three kinds of relations.

Regarding the second requirement, all the relations were modeled based on the classification proposed by (Rajewsky, 2011), integrating LRMoo, CIDOC-CRM, and RDA to achieve greater detail, but also adding classes and properties created from scratch. This choice was driven by the study of the standards considered from which some limitations emerged, namely concerning the semantics of CIDOC-CRM and LRMoo. Both models offer properties to express relationships between entities and some of them are well-aligned with our domain of interest (e.g. **`lrmoo:R75_incorporates`**). In some cases, the availability of more general properties proved advantageous (e.g. **`lrmoo:R68_is_inspired_by`**), allowing for the expression of semantic nuances more closely aligned with the phenomena found in the texts. Nonetheless, more complex semantic relations remain insufficiently covered. For this purpose, the MIrO data model was created (see Fig. 1). The main class, i.e. `miro:IntermedialRelation`, expresses all those phenomena that occur between different media (Rajewsky, 2011: 46). These phenomena can be divided into three subcategories based on the type of intermediality (Rajewsky, 2011: 51): media transposition (`miro:MediaTransposition`), media combination (`miro:MediaCombination`) and media reference (`miro:IntermedialReference`). Each of these categories is composed of two media (`miro:hasMediumA`, `miro:hasMediumB`) that are linked to one another by a meaningful relation, expressed by an object property. In the first category, we opted to maintain a higher level of modeling by using the symmetric property `lrmoo:R68_is_inspired_by` instead of its asymmetric superproperty `lrmoo:R2_is_derivative_of`. While media transpositions involve the transformation of a cultural product from one medium to another, our case study does not involve real transformations but more an influence between different media. The two related media in this class are described at their `lrmoo:F1_Work` level, for which the mentioned property is already specified by the LRMoo standard. The second category concerns the process of combining at least two distinct media: we decided to reuse the property `lrmoo:R75_incorporates` to express the inclusion of an `lrmoo:F2_Expression` in another `lrmoo:F2_Expression`, both regarding the realization of different works. The third category encompasses two types of media references which depend on the nature of the reference and the specific content type of the cultural objects involved in the relationship. The first type of intermedial reference (`miro:MediaReference`) involves an evocation (`miro:evokes`) or depiction (`crm:P65_shows_visual_item`) of some kind of medium within another one of different nature. The second type of media reference refers instead to a more specific relation, the one between the chapters of Sanguineti's work and the items described. This relation is called ekphrastic relation (`miro:EkphrasticRelation`) and both entities involved must share a specific content type (`rdae:P20001` "`has content type`"), namely a text and an image. In this context, the text serves as the referencing medium (Rajewsky, 2011: 53), and it is inherently connected to the object it refers to through a specific relation (`miro:hasIntermedialReference`). Moreover, from each ekphrasis, a set of features can be extracted (`miro:hasFeature`) to compare the same object described within different chapters, providing information about the intratextual relations.

The model also integrates the part concerning the scholar's hermeneutical analysis. The individuation of the kinds of intermedial relations is supported by a claim connected to a `hico:InterpretationAct` by the property `prov:wasGeneratedBy`. Besides HiCO (The Historical Context Ontology[10]) and PROV-O (Provenance Ontology[11]) ontologies, the part concerning the scholar's interpretations integrates CiTO (Citation Typing Ontology[12]), mainly to express if scholars agree or not with one another (`cito:agreesWith`, `cito:disagreesWith`).

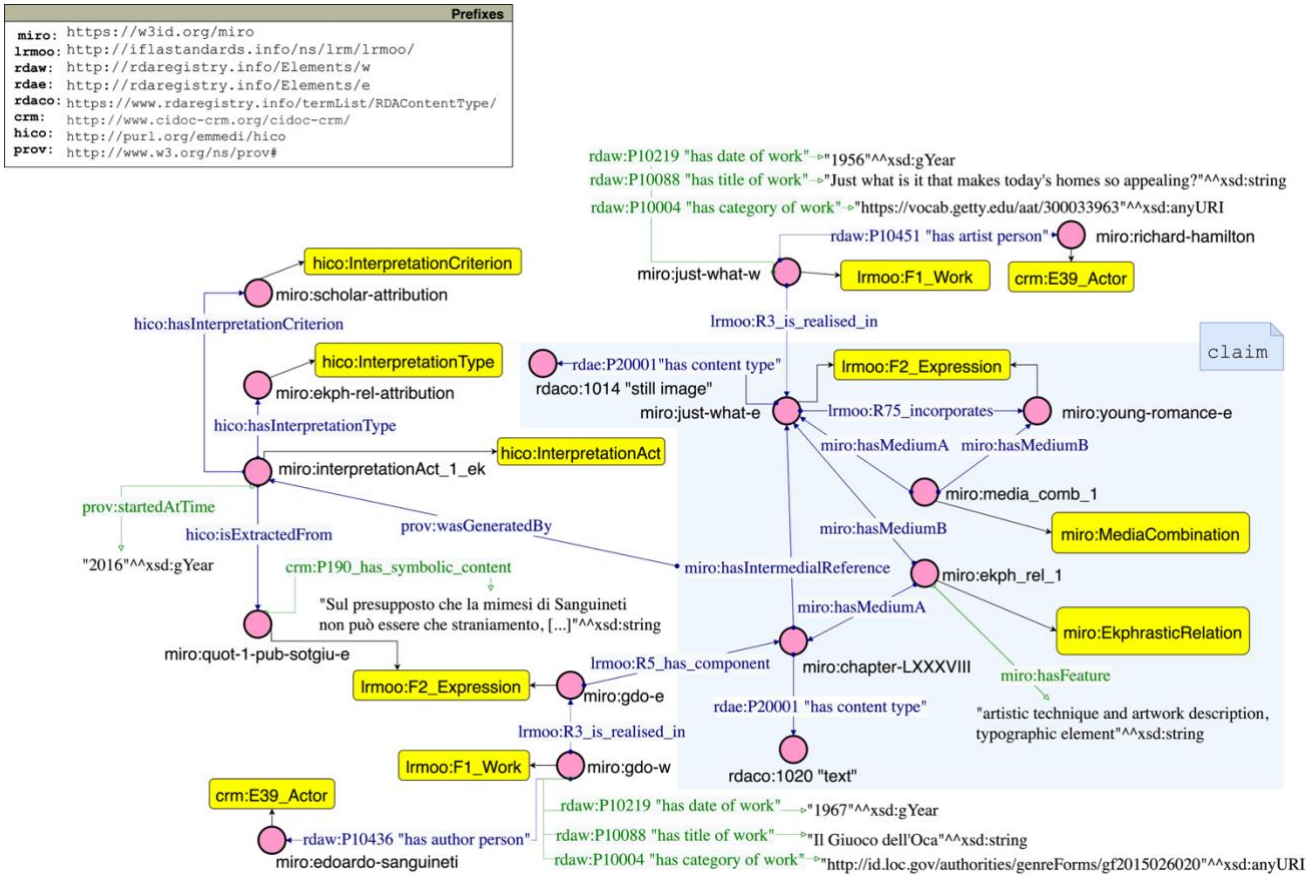


Figure 2. Example of integrated modelization: WEMI levels, relations, and interpretation act.

In Figure 2, we briefly present an example of modelization based on the formal representation of chapter LXXXVIIII and two types of intermedial relations identified within. The first relation is the ekphrastic relation (`miro:hasIntermedialReference`), which connects the chapter to a specific cultural object (i.e. Richard Hamilton's collage *Just What Is It That Makes Today's Homes So Different, So Appealing?*). Particularly, the ekphrasis is described through specific features (`miro:hasFeature`). The second relation links the collage to another cultural object, the comic magazine *Young Romance*, which is incorporated within the collage (`lrmoo:R75_incorporates`). Each cultural object, including *Il Giuoco dell'Oca* of which the chapter is part (`lrmoo:R5_has_component`), is described alongside its metadata respecting the WEMI levels at which the specific intermedial relations occur. Furthermore, each relationship is associated with a specific interpretation act through the property `prov:wasGeneratedBy`.
Part of the scenario presented in Figure 2, concerning the ekphrasis of the cultural object *Just What Is It That Makes Today's Homes So Different, So Appealing?* and its metadata, is formalized in Turtle syntax as follows[13]:

```
<https://w3id.org/miro/work/just-what-w> a lrmoo:F1 ;
```

[10] https://marilenadaquino.github.io/hico/.
[11] https://www.w3.org/TR/prov-o/.
[12] https://sparontologies.github.io/cito/current/cito.html.
[13] The complete RDF dataset in Turtle syntax is available at the following link: https://github.com/enricabruno/miro/tree/main/data.

```
    lrmoo:R3 <https://w3id.org/miro/expression/just-what-e> ;
    rdaw:P10004 "https://vocab.getty.edu/aat/300033963"^^xsd:anyURI ;
    rdaw:P10088 "Just what is it that makes today's homes so appealing?"^^xsd:string ;
    rdaw:P10219 "1956"^^xsd:gYear ;
    rdaw:P10451 <https://w3id.org/miro/agent/richard-hamilton> ;
    owl:sameAs "https://www.wikidata.org/wiki/Q3190307"^^xsd:anyURI .

<https://w3id.org/miro/expression/chapter-LXXXVIII> a lrmoo:F2 ;
    miro:hasIntermedialReference <https://w3id.org/miro/expression/just-what-e> .

<https://w3id.org/miro/ekphrasticRelation/ekph_rel_1> a miro:EkphrasticRelation ;
    prov:wasGeneratedBy <https://w3id.org/miro/interpretationAct/interpretationAct_1_ek> ;
    miro:hasFeature "artistic technique and artwork description"^^xsd:string,
        "typographic element"^^xsd:string ;
    miro:hasMediumA <https://w3id.org/miro/expression/chapter-LXXXVIII> ;
    miro:hasMediumB <https://w3id.org/miro/expression/just-what-e> .
```

4. RESULTS, CONCLUSION AND FUTURE WORK

This presentation suggests a method for analyzing contemporary texts using Semantic Web technologies to enhance reading and comprehension. The chosen case study, *Il Giuoco dell'Oca* by Edoardo Sanguineti, was examined to explore its unique structure and content. The focus was on intermedial relations, intratextual connections, and interpretative actions within the text.

From the SPARQL competency questions conducted on the dataset, initial observations reveal previously hidden aspects of how the author constructed the text. Building on these insights, the formalization of textual features through computational methodologies emerges as a promising approach for deepening textual interpretation. Such an approach could open up innovative research directions and encourage the development of new readings and scholarly perspectives. In particular, the initial insights into Sanguineti's use of ekphrasis across different chapters highlight the author's role in constructing a combinatorial textual structure. Specifically, some cultural objects are described in more than one chapter and, when it happens, the number of chapters sharing the same object is always three. Additionally, the perspective from which the first cultural object (i.e. *Cassa sistina* by Mario Ceroli) is described shifts throughout the novel, suggesting a cinematographic portrayal that begins with a view from inside the box and concludes with a broader environmental overview. Another important result concerns the interpretation acts, particularly the comparison of agreements and disagreements. Specifically, there is only one instance of disagreement, while there are nine instances of agreement, signaling a shared agreement between scholars' identification of the cultural objects connected to chapters[14].

In future works, MIrO ontology could be further extended by integrating the Scholarly Observations ontology in OWL (Sanfilippo, 2024), enhancing its capacity to represent scholars' interpretative acts on literary texts. As part of future developments, we also plan to implement an interactive semantic narrative that leverages the principles of Linked Open Data as a publishing method. In this way, an annotation on the digitized collage of Baruchello's *Il Giuoco dell'Oca di Edoardo Sanguineti*, linking each box to its corresponding chapter, could clarify the novel's structure as a gamebook and provide an alternative means of accessing the text. Additionally, we aim to integrate annotations for each chapter, within copyright limits, referencing both the described cultural objects and related items, as well as the corresponding interpretation acts. Finally, we will test the MIrO model on other case studies, focusing on general texts that exhibit a certain degree of relational classification.

ACKNOWLEDGMENTS

We would like to thank Francesca Massarenti for her valuable contribution to this work, which was developed as part of the *Information Science and Cultural Heritage* course within the Master's program in *Digital Humanities and Digital Knowledge* at the University of Bologna.

This work was partially funded under the National Recovery and Resilience Plan (PNRR), Project PE 0000020 CHANGES - CUP J33C22002850006, Mission 4 Component 2 Investment 1.3, funded by the European Union - NextGenerationEU.

[14] A complete overview of the competency questions tested through SPARQL is available at the following link: https://github.com/enricabruno/miro/tree/main/SPARQL_queries.

REFERENCES

Albani, P. (2018). La letteratura come gioco combinatorio. Griselda. Il portale di Letteratura. https://site.unibo.it/griseldaonline/it/approfondimenti/paolo-albani-letteratura-gioco-combinatorio.

Bruno, E., Pasqual, V., & Tomasi, F. (2024). Italo Calvino's 'destini incrociati'. An experiment of semantic narrative modelling and visualisation. Umanistica Digitale, (17), 47-69. https://doi.org/10.6092/issn.2532-8816/19013.

Calvino, I. (2015). Una pietra sopra. Edizioni Mondadori.

Cometa, M. (2012). La scrittura delle immagini: Letteratura e cultura visuale. Raffaello Cortina.

Corosaniti, V. (2024). Inside the Black Box: Il Giuoco dell'Oca di Edoardo Sanguineti. Sinestesieonline, 24.

Curi, F. (2006). Una poetica della 'contrainte'. Sanguineti, l'avanguardia, l' 'Oulipo'. Poetiche, 3, 369-436.

Eco, U. (1979). Lector in fabula: la cooperazione interpretativa nei testi narrativi. Bompiani.

D'Agata, C., Del Grosso, A. M., Nay, L., Palazzolo, G., Sichera, A., & Spampinato, D. (2024). PAVES-e: Per una Hyperedizione dell'opera di Cesare Pavese. In Di Silvestro, A., Spampinato, D. (Eds.), Me.Te. Digitali. Mediterraneo in rete tra testi e contesti. Proceedings of the XIII Convegno Annuale AIUCD2024 (pp. 191-196). https://doi.org/10.6092/unibo/amsacta/7927.

Daquino, M., Giovannetti, F., & Tomasi, F. (2019). Linked Data per le edizioni scientifiche digitali. Il workflow di pubblicazione dell'edizione semantica del quaderno di appunti di Paolo Bufalini. Umanistica Digitale, 7, 49-75. https://doi.org/10.6092/issn.2532-8816/9091.

Meghini, C., Bartalesi, V., & Metilli, D. (2021). Representing narratives in digital libraries: The narrative ontology. Semantic Web, 12(2), 241-264. https://doi.org/10.3233/SW-200421.

Oberreither, B. (2024, 19–20 November). Text and Image Are the Same, or Close Enough at Least. Reflections on Expanding an Ontology for Intertextual Relations Towards Intermediality. [Conference presentation]. Linked Open Data and Literary Studies, Berlin, Germany. https://www.temporal-communities.de/research/digital-communities/resources/lod-resources/Bernhard_Oberreither_Text_and_Image.pdf.

Oplepo (2005). La Biblioteca Oplepiana. Zanichelli, 21-28.

Portesine, C. (2021). «Una specie di Biennale allargata»: il giuoco dell'ecfrasi nel secondo romanzo di Edoardo Sanguineti. Fabrizio Serra Editore.

Rajewsky, I. O. (2011). Intermediality, Intertextuality, and Remediation: A Literary Perspective on Intermediality. Intermédialités, 6, 43–64. https://doi.org/10.7202/1005505ar.

Risso, E. (2023). La regola del giuoco. Universale Economica, 9-37.

Sanfilippo, E. M., Masolo, C., Mosca, A., & Tomazzoli, G. (2024). Operationalizing Scholarly Observations in OWL. In Bikakis, A., Ferrario, R., Jean, S., Markhoff, B., Mosca, A., Nicolosi Asmundo, M. (Eds.), SWODCH'24: International Workshop on Semantic Web and Ontology Design for Cultural Heritage. CEUR Workshop Proceedings. https://ceur-ws.org/Vol-3809/paper3.pdf.

Sanguineti, E. (1967). Il Giuoco dell'Oca. Feltrinelli.

Sotgiu, E. (2016). Il Giuoco dell'Oca nella trilogia di Edoardo Sanguineti. Italianistica: rivista di letteratura italiana, 45(2), 141-156. https://www.jstor.org/stable/43997194 .

Tomasi, F. (2020a). Vespasiano da Bisticci, Lettere. Knowledge Site 3.0. https://dx.doi.org/10.6092/unibo/amsacta/6852.

Tomasi, F. (2020b). Digital humanities e organizzazione della conoscenza: una pratica di insegnamento nel LODLAM. AIB Studi, 60(2). https://dx.doi.org/10.2426/aibstudi-12068.

From Metadata to Storytelling: A Framework For 3D Cultural Heritage Visualization on RDF Data

Sebastian Barzaghi[1], Simona Colitti[2], Arianna Moretti[3], Giulia Renda[4]
[1] Department of Cultural Heritage, University of Bologna, Italy - sebastian.barzaghi2@unibo.it
[2] Department of Architecture, University of Bologna, Italy – simona.colitti2@unibo.it
[3] Department of Classical Philology and Italian Studies, University of Bologna, Italy – arianna.moretti4@unibo.it
[4] Department of Classical Philology and Italian Studies, University of Bologna, Italy – giulia.renda3@unibo.it

ABSTRACT (ENGLISH)
This paper introduces a pipeline for integrating semantic metadata, 3D models, and storytelling, enhancing cultural heritage digitization. Using the Aldrovandi Digital Twin case study, it outlines a reusable workflow combining RDF-driven narratives and data visualization for creating interactive experiences to facilitate access to cultural heritage.
Keywords: Digital Twin; Storytelling; User Experience Design; Data Visualization Design; RDF

ABSTRACT (ITALIANO)
Dai Metadati allo Storytelling: Un Framework per la Visualizzazione del Patrimonio Culturale 3D su Dati RDF. Questo articolo presenta una pipeline per integrare metadati semantici, modelli 3D e narrazione, migliorando la digitalizzazione del patrimonio culturale. Utilizzando il caso di studio del gemello digitale di Aldrovandi, introduciamo un flusso di lavoro riutilizzabile per combinare narrazioni e visualizzazioni su dati RDF per rendere il patrimonio culturale più accessibile attraverso esperienze interattive.
Parole chiave: Gemelli Digitali; Storytelling; Esperienza dell'Utente; Visualizzazione Dati; RDF

1. INTRODUCTION
The CHANGES project[1] ("Cultural Heritage Active Innovation For Next-Gen Sustainable Society") is an EU-funded initiative to promote, preserve, and make accessible Italian cultural heritage (CH). Its Spoke 4 focuses on using virtual technologies to enhance CH in museums and art collections. Engaging cultural institutions in this effort improves artefacts' knowledge, curation, and management, while also expanding public involvement through innovative communication strategies (Balzani et al., 2024). While RDF is a standard for managing cultural heritage metadata, significant gaps persist in its application to 3D objects, particularly in leveraging metadata to enhance user experiences and engagement. The limitation is critical in museums, where integrating metadata with narrative techniques has shown great potential to enrich visitors' engagement (Grasso et al., 2024). Both physical and digital storytelling approaches combined with data visualization have proven effective in conveying knowledge and fostering interaction with CH collections (Fernie et al., 2012). Nonetheless, no comprehensive framework already bridges the gap between RDF metadata and 3D objects through narrative techniques. This paper explores this gap by reviewing relevant works (Section 2), proposes an interdisciplinary and reusable workflow to address it (Section 3), and demonstrates its application in the Aldrovandi case study (Section 4).

2. STATE OF THE ART
2.1 SEMANTIC METADATA FOR 3D OBJECTS
In CH digitization, capturing metadata about the cultural heritage object (CHO) and the process is required for management, enrichment, and presentation in a virtual environment. In recent years, several projects have explored semantic approaches for 3D CH metadata. Early efforts, like 3D-ICONS (D'Andrea & Fernie, 2013), focused on integrating 3D models into Europeana using CIDOC-CRM and CRMdig extensions. In the following years, widespread ontologies were largely adopted for modeling RDF metadata of CH objects, in projects such as the one by Messaoudi et al. (2018), who created a pipeline to integrate various information about built heritage into an ontological model aligned with CIDOC-CRM and other related modules. Catalano et al. (2020) later developed a methodology to include quantitative documentation data, expressed according to a schema based on CIDOC-CRM and CRMdig, in the description of 3D CH artifacts. Recently, the trend of reusing and extending existing models to describe 3D CH data re-emerged with a focus on reproducibility. Amato et al. (2023) illustrated the metadata management process used to

[1] https://www.fondazionechanges.org/ (cons. 24/01/25)

select elements from existing models and optimize their reuse, while Hermon et al. (2024) proposed a framework to integrate a CHO's data from several sources into a digital twin, including its data, documentation, and digital model.
To make real use of metadata across individual projects, a specific effort should be devoted to ensuring their FAIRness, focusing on making data interoperable. To this end, an extensive crosswalk mapping (European Commission. Directorate General for Research and Innovation. & EOSC Executive Board, 2021) was proposed to facilitate data exchanges. A key practice for interoperability is publishing datasets as Linked Open Data (LOD), using code-oriented tools for semantically structuring data, such as the Python library RDFLib, although inaccessible to users who can't code. Alternative approaches are based on mapping tools, such as the RDF Mapping Language (RML) (Dimou et al. 2014), a Java-based engine limiting the interaction to configuration files only, without the necessity of programming, unless specific conversion necessities emerge. Based on RML, Morph-KGC (Arenas-Guerrero, Chaves-Fraga, et al., 2024) was developed to allow Python programmers to extend RML functionalities with user-defined functions (Arenas-Guerrero, Espinoza-Arias, et al., 2024).

2.2 STORYTELLING, DATA VIZ DESIGN, USER EXPERIENCE DESIGN

Design-driven approaches in the digitization of CH allow for transforming collected data into accessible and meaningful resources by integrating contributions from both design and humanities disciplines (Lupo et al., 2023; Celi et al., 2016). Storytelling through data-informed approaches offers benefits in interpreting and engaging with CH collections, while information visualization enhances access and understanding with representations that simplify complex datasets. In addition to that, integrating RDF metadata allows for exploiting semantic connections in the narratives. This methodology not only increases accessibility in compliance with inclusive standards but also fosters data reuse through open-access licenses, benefiting both general audiences and experts. A systemic design approach ensures swift access, effective technology transfer, and cultural relevance. Combining innovative technologies with open access principles, the design fosters broad, sustainable, and interdisciplinary engagement, enriching CH through community and stakeholder interaction with advanced digital tools. Designers play a crucial role in facilitating communication among heritage experts, technologists, and end users by translating diverse needs and expectations into practical, user-friendly solutions that enhance the accessibility and meaningfulness of cultural heritage (Trocchianesi & Bollini, 2023). The process relies on advanced metadata modeling, prototyping, and accessible outputs such as data stories. These narratives make complex datasets comprehensible, engaging, and reusable for diverse audiences. Data storytelling has been shown to bridge the gap between data complexity and user experience, enabling new forms of engagement and contributing to collective digital memory (Masud et al., 2010). By leveraging tools like RawGraph[2] and Figma[3], pre-prototyping phases facilitate the exploration of visualization strategies that reinforce the project's visual identity (Mauri & Ciuccarelli, 2012).

2.3 SUCCESSFUL STORYTELLING STRATEGIES ADOPTED FOR OTHER DIGITAL RESOURCES

The exploration of data-driven design methodologies for CH has been the key focus of the Polifonia project[4], a European H2020 initiative dedicated to preserving and disseminating music heritage through innovative digital tools. As part of its contributions to the field, Polifonia developed a novel methodology that combines eXtreme Design (Carriero et al., 2021; Presutti et al., 2009) and Design Thinking (Chasanidou et al., 2015; Rowe, 1991) to bridge the gap between ontology-driven approaches and user-centered interface design (Grasso et al., 2024). This hybrid framework ensures that domain-specific requirements, as well as user needs, are seamlessly integrated into the development of tools for information retrieval, exploration, and dissemination (Renda, Grasso, et al., 2023). By integrating these structured insights into design workflows, Polifonia has established a replicable model for creating applications that cater to diverse audiences while leveraging the rich, structured data offered by LOD. Importantly, the tools developed through this methodology have undergone evaluation and are already being used in real-world scenarios. This highlights the practical reliability and effectiveness of the approach, reinforcing the potential for reuse and adaptation.

[2] https://www.rawgraphs.io/ (cons. 24/01/25)

[3] https://www.figma.com/ (cons. 23/01/25)

[4] https://polifonia-project.eu/ (cons. 26/01/25)

One of the key products of this methodology is MELODY[5], an open-source platform that empowers users to author content in the form of *data stories* (Renda, Daquino, et al., 2023). Designed to promote accessibility and engagement with LOD, MELODY enables domain experts and general users to combine data-driven visualizations with curated narratives. It provides a user-friendly *What You See Is What You Get* (WYSIWYG) interface for creating, editing, and publishing articles that integrate SPARQL query results as visual components such as charts, maps, and tables. MELODY's narrative-driven approach addresses significant challenges in presenting LOD. The tool supports the creation of content that combines interpretive storytelling with data visualization, making complex datasets accessible and engaging to a broader audience. Its flexible, open-source architecture ensures that it can be reused with any SPARQL endpoint. Additionally, its open-source nature allows users to extend or modify the platform according to their specific needs, making it an adaptable solution for a wide range of applications in the CH domain.

3. METHODOLOGY

Our research adopts a user experience-oriented, iterative, and multidisciplinary approach to tackle the challenge of enhancing access to CH resources through digital means. This approach is characterized by multidisciplinary collaboration among researchers from diverse fields, ensuring a holistic solution. Open data principles guide the entire process, making all outputs reusable and scalable within the LOD paradigm. A key focus is on narrative techniques, which enhance user engagement through interactive paths within the digital environment. Systematic metadata integration bridges gaps in 3D heritage model accessibility, developing a comprehensive pipeline from data collection to visualization. As shown in Figure 1, this process consists of parallel, iterative activities.

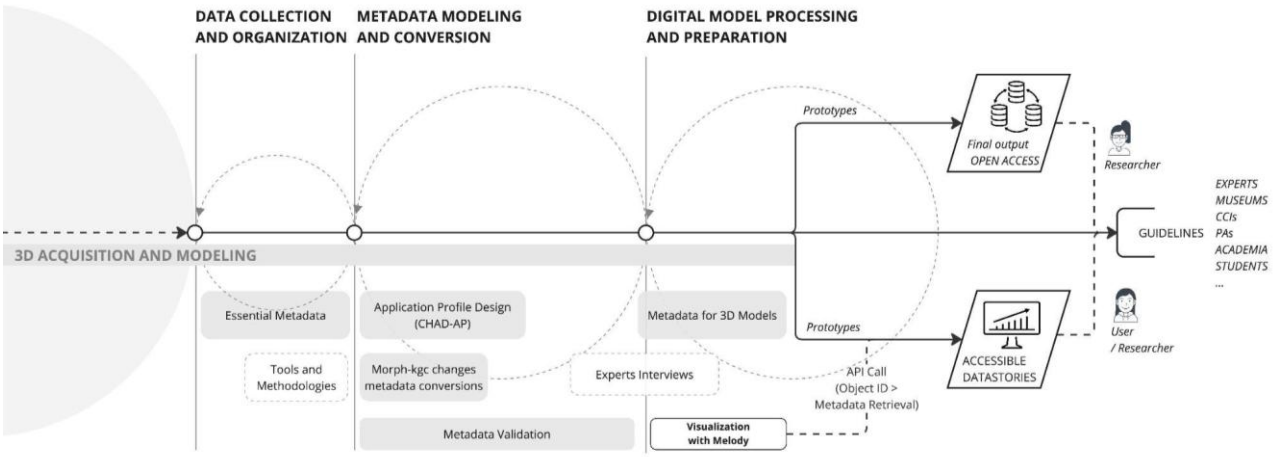


Figure 1. Iterative process for CH digital accessibility. The figure showcases the methodological framework encompassing iterative UX design, metadata integration, and storytelling techniques aimed at improving the accessibility and engagement of CH resources in digital environments.

In the **data collection and organization** phase, metadata about CHOs and digitization steps were organized into two Google Sheets[6] tables, chosen for their intuitive interface, real-time collaboration, and versioning. Each column represents a metadata field, and each row corresponds to a unique CHO.
In the **metadata modeling and conversion** step, the spreadsheet-collected tabular metadata was transformed into RDF triples through a two-step process: 1) using an ontology development methodology such as SAMOD (Peroni, 2017) to create an application profile of CIDOC-CRM for interoperability, and 2) converting tabular data into RDF triples using an ad-hoc extended version of Morph-KGC[7], which leverages RML for semantic conversion.

[5] https://projects.dharc.unibo.it/melody/ (cons. 26/01/25)

[6] https://workspace.google.com/products/sheets/ (cons. 26/01/25)

[7] https://github.com/dharc-org/morph-kgc-changes-metadata (cons. 10/04/2026)

The **prototyping** phase employs iterative user experience and service design principles. Using Figma, prototypes were refined through stakeholder feedback, combining collection-level narratives with object-specific stories to explore connections between CHOs. This approach enhances accessibility, engagement, and adaptability through intuitive interfaces linking metadata insights with storytelling.
For **digital model processing and preparation**, MELODY was used to integrate LOD, data visualization, and narrative techniques to engage users. While robust, MELODY's reliance on single SPARQL endpoints and manual processes limits scalability for large-scale reuse. To address these issues and integrate seamlessly with ATON (Fanini et al., 2021)—the 3D visualization framework used in Spoke 4—we developed an API that reuses MELODY's core components to dynamically generate customized data stories for 3D objects, improving scalability and interoperability.

4. USE CASE

In this section, we present the Aldrovandi temporary exhibition Digital Twin (Balzani et al., 2024) as an example to illustrate the application of our approach. In 2022-2023, the exhibition "*The Other Renaissance - Ulisse Aldrovandi and the Wonders of the World*" was held at Palazzo Poggi in Bologna, gathering a varied selection of objects from different sources. The exhibition, which is no longer accessible today, has been entirely digitized, including reproductions of interiors and 3D models of each object with accurate metadata, both on objects and the entire digitization process. As a result, the Aldrovandi Digital Twin is a 3D representation of the exhibition, showcasing how our methodology integrates RDF metadata-driven storytelling with virtual technologies to enrich user interaction with CH.

4.1 DATA COLLECTION, MODELING AND CONVERSION

The process of digitizing the exhibition's CHOs (Barzaghi, Bordignon, Collina, et al., 2024) started by capturing 3D data through structured-light scanning and photogrammetry, according to the object's physical characteristics and accessibility. Then, the raw data (RAW) was processed and refined into detailed 3D models. More specifically, three versions of each 3D model were retained for practical and scientific purposes: the raw processed model (RAWp), the digital CH object (DCHO), and the optimized version for real-time use (DCHOo). Next, after the models were optimized for performance, they were exported in appropriate formats for real-time web visualization.
In the meantime, two tables were created on Google Sheets: a table was used to collect the descriptive metadata of the exhibition objects, to capture their historical and contextual information, while the other one was used to collect the data of the whole digitization process, recording contextual information for each of the aforementioned phases. While the models were finalized, their metadata in tabular form were exported in CSV format and converted into RDF triples for structured representation. The semantic model used to describe the objects and processes involved in digitizing CHOs in a structured way is an application profile aligned with the CIDOC-CRM framework and called *Cultural Heritage Acquisition and Digitisation - Application Profile* (CHAD-AP)[8] (Barzaghi, Heibi et al., 2024). CHAD-AP has been implemented as an OWL ontology organized into two main modules: the *Object Module* (OM), dedicated to describing the CHO's characteristics and context, and the *Process Module* (PM), for describing the process of acquiring and digitizing it. The conversion of the data contained in both OM and PM into a RDF graph structured according to CHAD-AP specifications was realized by using a custom extension of the Morph-KGC software, which included mapping files, configuration files, and user-defined functions for handling the complex semantic constructs hidden in the tabular format, together with a Python launch script that orchestrated data preprocessing, mapping, and post-processing tasks.
Finally, the ATON framework was used to manage and present the models interactively on the Web, while a SPARQL endpoint was made available for querying and providing structured data about the original CHOs, their digital models, and the whole process of their acquisition and digitization.

4.2 PROTOTYPING

As part of an ongoing design research effort, user needs were continuously gathered and analyzed through questionnaires and expert consultations conducted alongside the development process. This iterative engagement identified specific requirements and highlighted the narrative potential of the collections' data, ensuring that the data stories aligned with user expectations and heritage valorization goals. Feedback from stakeholders shaped key elements for structuring meaningful data stories and refining the metadata

[8] https://w3id.org/dharc/ontology/chad-ap (cons. 24/01/25)

framework, fostering a dynamic exchange of insights. This process ensured that the data-driven storytelling approach remained relevant and adaptable to the evolving project needs.
The creation of data-informed storytelling for CH collections requires a methodological framework that incorporates various visualization techniques and design strategies. This involves several key phases. (1) Metadata modelling and conversion ensure metadata interoperability and quality through structuring, validation techniques, and expert interviews; (2) the selection of visualization and navigation design defines appropriate visualization types (e.g., bar charts, maps, interactive diagrams) and supports the creation of intuitive, immersive interfaces using tools like Figma, with content categorized by themes and visualizations to enhance user guidance; (3) accessibility and usability testing involves rigorous compliance verification using tools like WebAIM[9] for contrast testing and integrating assistive technologies for users with disabilities; (4) standardization and usage guidelines are developed to provide adaptable heritage digitization protocols, including a white paper detailing workflows and data integration methods; (5) monitoring and feedback mechanisms are established to refine data stories and improve project outcomes; (6) impact assessment evaluates accessibility and audience engagement, focusing on integrating digitized heritage into national repositories and expanding the model to additional collections. This iterative, participatory approach ensures that CH is not only digitized but also actively explored and enriched. Visualization tools such as bar charts and interactive diagrams provide an adaptive storytelling experience, facilitating user exploration and navigation through complex heritage data. These tools add value beyond physical exhibitions, fostering deeper engagement and uncovering new connections between collection pieces.

4.3 PROCESSING AND PREPARATION

Within the ATON framework, each DCHOo is visualized as a 3D digital twin. When a user selects an object, ATON sends an HTTP request to the API. This request includes the unique global persistent identifier for the Aldrovandi artifact and the location of a configuration file. Mirroring design research and decisions, the JSON configuration file specifies the details to be displayed in the data story, including the object typology, SPARQL queries for metadata retrieval, textual descriptions, visual elements, and interactive components. The API processes the configuration file dynamically executing the SPARQL queries to retrieve metadata from the specified endpoint. For the Aldrovandi artifact, these queries extract information such as its type, materials, dimensions, historical context, place of conservation, and details about its digitization process. This ensures that all relevant metadata is retrieved and contextualized for the data story. Using the metadata retrieved via SPARQL queries, the API generates a customized HTML page that serves as the data story for the 3D object the user is interacting with. This page is designed to include dynamic, artifact-specific sections such as descriptions of the object's materials, historical period, and the conservation site, along with visualizations like charts or images. The design ensures that the data story complements the 3D representation of the object by providing an additional layer of information, allowing users to explore the artifact in depth while engaging with its historical and cultural context.

5. CONCLUSIONS

In this contribution, we presented an iterative, reusable, and cross-disciplinary workflow for integrating RDF-driven data stories with 3D virtual representations of CHOs. Following a review of the state of the art, we outlined a methodology that includes data collection and organization, metadata modeling and conversion, prototyping, and model processing and preparation. The workflow was applied to the Aldrovandi Digital Twin use case to illustrate how semantically rich metadata can be used to enhance access and usability of virtual representations of CHOs. By integrating storytelling techniques with visualization strategies, the approach seeks to transform raw data into actionable knowledge while preserving the integrity and authenticity of the heritage materials. The effectiveness of this workflow in producing engaging data stories relies heavily on the quality and richness of the underlying data. Insufficient metadata or lack of contextual detail may limit the ability to craft a meaningful narrative or identify compelling connections between objects. The manual metadata preparation and narrative design aspects may also hinder efficiency and reproducibility in some contexts. Moreover, this work has not yet undergone extensive evaluation as an exploratory effort.
Further developments will also address user experience across different audience profiles. Researchers and domain experts are expected to prioritize access to structured metadata, semantic interoperability, and

[9] https://webaim.org/ (cons. 24/01/25)

opportunities for data reuse. In contrast, general users may benefit more from accessible narrative pathways and intuitive, exploratory interfaces. To support both audiences, the integration of panoramic views and synthetic data representations is intended to enable stepwise exploration, from general overviews to detailed insights, aligning with design principles that promote cognitive scalability and content adaptivity. These design choices aim to ensure that the interface not only conveys the richness of the underlying RDF data but also remains accessible and engaging for diverse user groups.
Despite the experimental stage of the process, having been developed within the framework of the Project CHANGES, the introduced workflow is meant to be readopted systematically in other case studies, and eventually become the standard for visualizing RDF data in this context. In the short to medium term, this will provide the opportunity to collect and analyze concrete data to evaluate the workflow's efficacy with different input material rather than the dataset it was developed for and tested on until now. The collected feedback will be exploited to iteratively improve the pipeline, to deliver a research tool as useful and easily reusable as possible. Consequently, future efforts will focus on refining the workflow and testing implementation for the Aldrovandi use case with advanced navigation and search capabilities, prioritizing accessibility and usability testing to ensure an inclusive user experience. Additionally, we aim to evaluate the project's impact on future studies, involving additional case studies (such as other collections and museums) and even situations characterized by sparse or low-quality data. Other interesting venues for further experimentation include testing automatic or semi-automatic approaches to facilitate the data modelling phase and developing quantitative methods to be integrated in the evaluation process that bring together (meta)data quality, service and narrative design heuristics, and the assessment of technical aspects such as data conversion and API processes. Finally, we also plan to work on actionable guidelines that standardize our approach and facilitate its adoption across other cultural heritage projects and institutions.

ACKNOWLEDGEMENTS
This work was partially funded by Project PE 0000020 CHANGES - CUP B53C22003780006, CUP J33C22002850006, NRP Mission 4 Component 2 Investment 1.3, Funded by the European Union - NextGenerationEU.

REFERENCES

Amato, A., Branco, D., Venticinque, S., Renda, G., & Mataluna, S. (2023). Metadata and Semantic Annotation of Digital Heritage Assets: A case study. *2023 IEEE International Conference on Cyber Security and Resilience (CSR)*, 516–522. https://doi.org/10.1109/CSR57506.2023.10224935

Arenas-Guerrero, J., Chaves-Fraga, D., Toledo, J., Pérez, M. S., & Corcho, O. (2024). Morph-KGC: Scalable knowledge graph materialization with mapping partitions. *Semantic Web*, *15*(1), 1–20. https://doi.org/10.3233/sw-223135

Arenas-Guerrero, J., Espinoza-Arias, P., Bernabé-Diaz, J. A., Deshmukh, P., Sánchez-Fernández, J. L., & Corcho, O. (2024). An RML-FNML module for Python user-defined functions in Morph-KGC. *SoftwareX*, *26*. https://doi.org/10.1016/j.softx.2024.101709

Balzani, R., Barzaghi, S., Bitelli, G., Bonifazi, F., Bordignon, A., Cipriani, L., Colitti, S., Collina, F., Daquino, M., Fabbri, F., Fanini, B., Fantini, F., Ferdani, D., Fiorini, G., Formia, E., Forte, A., Giacomini, F., Girelli, V. A., Gualandi, B., … Vittuari, L. (2024). Saving temporary exhibitions in virtual environments: The Digital Renaissance of Ulisse Aldrovandi – Acquisition and digitisation of cultural heritage objects. *Digital Applications in Archaeology and Cultural Heritage*, *32*, e00309. https://doi.org/10.1016/j.daach.2023.e00309

Barzaghi, S., Bordignon, A., Collina, F., Fabbri, F., Fanini, B., Ferdani, D., Gualandi, B., Heibi, I., Mariniello, N., Massari, A., Massidda, M., Moretti, A., Peroni, S., Pescarin, S., Rega, M. F., Renda, G., & Sullini, M. (2024). *A reproducible workflow for the creation of digital twins in the cultural heritage domain*. Zenodo. https://doi.org/10.5281/zenodo.14170838

Barzaghi, S., Bordignon, A., Gualandi, B., Heibi, I., Massari, A., Moretti, A., Peroni, S., & Renda, G. (2024). A Proposal for a FAIR Management of 3D Data in Cultural Heritage: The Aldrovandi Digital Twin Case. *Data Intelligence*. https://doi.org/10.3724/2096-7004.di.2024.0061

Barzaghi, S., Collina, F., Fabbri, F., Giacomini, F., Bordignon, A., Balzani, R., Bitelli, G., Bonifazi, F., Cipriani, L., Colitti, S., Daquino, M., Fanini, B., Fantini, F., Ferdani, D., Fiorini, G., Formia, E., Forte, A., Girelli, V. A., Gualandi, B., … Vittuari, L. (2023). Digitisation of Temporary Exhibitions: The

Aldrovandi Case. *Eurographics Workshop on Graphics and Cultural Heritage*, 181–183. https://doi.org/10.2312/gch.20231176

Barzaghi, S., Heibi, I., Moretti, A., & Peroni, S. (2025). Developing Application Profiles for Enhancing Data and Workflows in Cultural Heritage Digitisation Processes. In G. Demartini, K. Hose, M. Acosta, M. Palmonari, G. Cheng, H. Skaf-Molli, N. Ferranti, D. Hernández, & A. Hogan (A c. Di), *The Semantic Web – ISWC 2024* (pp. 197–217). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-77847-6_11

Bordignon, A., Barzaghi, S., Collina, F., Fabbri, F., Fanini, B., Ferdani, D., Marinello, N., Moretti, A., Rega, M. F., & Sullini, M. (2024). *Guidelines for the digitisation of museum and art collections*. Zenodo. https://doi.org/10.5281/zenodo.14249936

Carriero, V. A., Gangemi, A., Mancinelli, M. L., Nuzzolese, A. G., Presutti, V., & Veninata, C. (2021). Pattern-based design applied to cultural heritage knowledge graphs. *Semantic Web*, *12*(2), 313–357. https://doi.org/10.3233/SW-200422

Catalano, C. E., Vassallo, V., Hermon, S., & Spagnuolo, M. (2020). Representing quantitative documentation of 3D cultural heritage artefacts with CIDOC CRMdig. *International Journal on Digital Libraries*, *21*(4), 359–373. https://doi.org/10.1007/s00799-020-00287-3

Celi, M., Formia, E., Zingale, S., Celaschi, F., Trocchianesi, R., Penati, A., Lupo, E., Colombi, C., & Rizzo, F. (2016). *Humanities Design Lab. Le culture del progetto e le scienze umane e sociali*.

Chasanidou, D., Gasparini, A. A., & Lee, E. (2015). Design Thinking Methods and Tools for Innovation. In A. Marcus (A c. Di), *Design, User Experience, and Usability: Design Discourse* (pp. 12–23). Springer International Publishing. https://doi.org/10.1007/978-3-319-20886-2_2

D'Andrea, A., & Fernie, K. (2013). CARARE 2.0: A metadata schema for 3D cultural objects. *2013 Digital Heritage International Congress (DigitalHeritage)*, *2*, 137–143. https://doi.org/10.1109/DigitalHeritage.2013.6744745

Dimou, A., Sande, M. V., Colpaert, P., Verborgh, R., Mannens, E., & Walle, R. V. de. (2014). RML: A Generic Language for Integrated RDF Mappings of Heterogeneous Data. In C. Bizer, T. Heath, S. Auer, & T. Berners-Lee (A c. Di), *Proceedings of the Workshop on Linked Data on the Web co-located with the 23rd International World Wide Web Conference (WWW 2014), Seoul, Korea, April 8, 2014* (Vol. 1184). CEUR-WS.org. https://ceur-ws.org/Vol-1184/

European Commission. Directorate General for Research and Innovation. & EOSC Executive Board. (2021). *EOSC interoperability framework: Report from the EOSC Executive Board Working Groups FAIR and Architecture.* Publications Office. https://data.europa.eu/doi/10.2777/620649

Fanini, B., Ferdani, D., Demetrescu, E., Berto, S., & d'Annibale, E. (2021). ATON: An Open-Source Framework for Creating Immersive, Collaborative and Liquid Web-Apps for Cultural Heritage. *Applied Sciences*, *11*(22), Articolo 22. https://doi.org/10.3390/app112211062

Fernie, K., Griffiths, J., Stevenson, M., Clough, P., Goodale, P., Hall, M., Archer, P., Chandrinos, K., Agirre, E., de Lacalle, O. L., de Polo, A., & Bergheim, R. (2012). PATHS: Personalising access to cultural heritage spaces. *2012 18th International Conference on Virtual Systems and Multimedia*, 469–474. https://doi.org/10.1109/VSMM.2012.6365960

Grasso, M., Daquino, M., & Renda, G. (2024). From Ontology Design to User Experience. A Methodology to Design Interfaces for Information Seeking Purposes. *Umanistica Digitale*, *18*, Articolo 18. https://doi.org/10.6092/issn.2532-8816/19038

Hermon, S., Niccolucci, F., Bakirtzis, N., & Gasanova, S. (2024). A Heritage Digital Twin ontology-based description of Giovanni Baronzio's "Crucifixion of Christ" analytical investigation. *Journal of Cultural Heritage*, *66*, 48–58. https://doi.org/10.1016/j.culher.2023.11.004

Lupo, E., Camosino, G., Gobbo, B., Motta, M., Mauri, M., & Parente, M. (2023, ottobre 9). Digital for Heritage and Museums: Design-Driven Changes and Challenges. *IASDR 2023: Life-Changing Design*. IASDR 2023: Life-Changing Design. https://doi.org/10.21606/iasdr.2023.397

Masud, L., Valsecchi, F., Ciuccarelli, P., Ricci, D., & Caviglia, G. (2010, luglio). From Data to Knowledge—Visualizations as Transformation Processes within the Data-Information-Knowledge Continuum. *2010 14th International Conference Information Visualisation*. 2010 14th International Conference Information Visualisation. https://doi.org/10.1109/IV.2010.68

Mauri, M., & Ciuccarelli, P. (2012). *Il ruolo dell'Information Visualization nella progettazione di interfacce per archivi digitali eterogenei*. Primo Convegno Annuale dell'Associazione Informatica Umanistica e Cultura Digitale, Firenze. https://doi.org/10.13140/2.1.3470.9446

Messaoudi, T., Véron, P., Halin, G., & De Luca, L. (2018). An ontological model for the reality-based 3D annotation of heritage building conservation state. *Journal of Cultural Heritage*, *29*, 100–112. https://doi.org/10.1016/j.culher.2017.05.017

Moretti, A., Heibi, I., & Peroni, S. (2024). *A Workflow for GLAM Metadata Crosswalk* (arXiv:2405.02113). arXiv. https://doi.org/10.48550/arXiv.2405.02113

Peroni, S. (2017). A Simplified Agile Methodology for Ontology Development. In M. Dragoni, M. Poveda-Villalón, & E. Jimenez-Ruiz (A c. Di), *OWL: Experiences and Directions – Reasoner Evaluation* (pp. 55–69). Springer International Publishing. https://doi.org/10.1007/978-3-319-54627-8_5

Presutti, V., Daga, E., Gangemi, A., & Blomqvist, E. (2009). eXtreme design with content ontology design patterns. *Proceedings of the 2009 International Conference on Ontology Patterns*, *516*, 83–97.

Renda, G., Daquino, M., & Presutti, V. (2023). Melody: A Platform for Linked Open Data Visualisation and Curated Storytelling. *Proceedings of the 34th ACM Conference on Hypertext and Social Media*, 1–8. https://doi.org/10.1145/3603163.3609035

Renda, G., Grasso, M., & Daquino, M. (2023). From ontology design to user-centred interfaces for music heritage. In E. Carbé, G. Lo Piccolo, A. Valenti, & F. Stella (A c. Di), *La memoria digitale: Forme del testo e organizzazione della conoscenza* (pp. 168–172). AIUCD. https://doi.org/10.48550/arXiv.2306.12973

Rowe, P. G. (1991). *Design thinking*. MIT press.

Trocchianesi, R., & Bollini, L. (2023). Design, Digital Humanities, and Information Visualization for Cultural Heritage. *Multimodal Technologies and Interaction*, *7*(11), Articolo 11. https://doi.org/10.3390/mti7110102

Between Text and Icon: Towards A Representational Model for Ekphrastic Relations

Maria Francesca Bocchi[1], Carlo Teo Pedretti[2], Fabio Vitali[3]
[1] University of Bologna, Italy - maria.bocchi4@unibo.it
[2] University La Sapienza of Rome, Italy - carloteo.pedretti@uniroma1.it
[3] University of Bologna, Italy - fabio.vitali@unibo.it

ABSTRACT

Especially before the current, markedly visual-dominated society, ekphrasis, i.e. the textual description of art objects as a formal analysis of them, was regarded as a vital component in discourses about art history, aesthetics and transmediality. Beyond its classical roots, modern perspectives on ekphrasis explore its role in intermediality and visual studies, emphasizing the dynamic relationships between text and image. Supported by the increasing diffusion of Semantic Web technologies, several interdisciplinary research fields leverage theoretical frameworks close to iconology to pursue quantitative art history with computational methods. Nevertheless, while iconographic studies have been recently addressed in ontologies, a complete description of aspects relevant to ekphrasis is still missing. This paper provides a structured formalization of ekphrasis, enabling scholars to annotate, compare, and visualize the interpretive layers of ekphrastic descriptions across multiple works. The presentation, therefore, aims to: (1) survey the theoretical background regarding ekphrasis to represent a formal model by extending and reusing existing ontologies; (2) apply the proposed ontology to the peculiar case study of Giuseppe Raimondi's writings; and (3) create an annotator prototype to leverage the model.

Keywords: ekphrasis; annotations; ontologies; IIIF; Web Annotation Data Model.

ABSTRACT (ITALIANO)

Tra Testo e Icona: verso un Modello Formale per le Relazioni Ecfrastiche

L'ecfrasi, ossia la descrizione testuale di oggetti artistici intesa come analisi formale degli stessi, è considerata un elemento fondamentale nei discorsi sull'estetica, la storia dell'arte e la transmedialità. Oltre alle sue radici classiche, le prospettive moderne sull'ecfrasi ne esplorano il ruolo nell'intermedialità e negli studi visivi, ponendo l'accento sulle relazioni dinamiche tra testo e immagine. Con il crescente utilizzo delle tecnologie del Web Semantico, prospettive teoriche vicine all'iconologia hanno sfruttato metodi computazionali nel campo della storia dell'arte. Tuttavia, sebbene siano stati proposti modelli formali relativamente agli studi iconografici, manca ancora una descrizione completa degli aspetti rilevanti per l'ecfrasi. Il contributo propone una formalizzazione strutturata dell'ecfrasi, permettendo agli studiosi di annotare, confrontare e visualizzare i livelli interpretativi delle descrizioni ecfrastiche. La presentazione si propone quindi di: (1) analizzare il panorama teorico relativo all'ecfrasi al fine di rappresentare un modello formale, estendendo e riutilizzando ontologie esistenti; (2) applicare l'ontologia proposta al caso di studio degli scritti di Giuseppe Raimondi; infine, (3) sviluppare un prototipo di annotatore basato sul modello proposto.

Parole chiave: ecfrasi; annotazioni; ontologie; IIIF; Web Annotation Data Model.

1. INTRODUCTION

This paper explores ekphrasis through the lens of Semantic Web technologies, proposing a formal model that leverages existing ontologies while addressing the limitations of previous approaches.

Ekphrasis traditionally refers to a vivid, often dramatic verbal description of a visual work of art. While in its classical form ekphrasis specifically involved verbal representations of paintings, sculptures, or other visual artworks, taking shape as a specific literary genre with distinct rhetorical techniques, contemporary usage of the term can be broader, encompassing the description of any non-verbal object or scene through the medium of language.

In particular, mimetic (or real) ekphrasis describes an actual, physically existing work of art. For example, J.J. Winckelmann's description of the *Laocoon* (Winckelmann, 2001), or D.G. Rossetti's sonnets on his paintings (Rossetti, 2010). Notional (or imagined) ekphrasis instead describes an artwork that does not exist outside the text itself—sometimes called "fictional ekphrasis" (Hollander, 1995; Krieger, 1992). A classic example is the shield of Achilles in the Iliad, which may be more a product of Homer's imagination than an actual shield. Examples also include ekphrasis as an essential mechanism for the remediation and

description of lost artworks and objects of uncertain attribution. More complex cases, such as Max Aub's ekphrasis of fictional paintings, later on physically created by the author himself, can be seen as a unique hybrid or "self-fulfilling" form of ekphrasis, collapsing the boundary between "fictional" and "actual" art (Aub, 1992).

In addition to the ontological relationship between its textual and iconic referents, the ekphrasis itself is often employed to produce a particular rhetorical effect on the reader. At its simplest level, ekphrasis acts as a denotation, exemplified by the explicit mention of a painting and its attributes in a text (Cometa, 2012). These forms of denotation—which resemble straightforward descriptive modes (as in hypotyposis) (Eco, 2002)—can classically be introduced by deictic expressions (such as "this," "that," etc.) and are typically accompanied by simple iconographic cues that literally point out the meanings of what is depicted (Cometa, 2012: 86; Voilloux, 2005). They could be used by the writer as a means not only to describe but also to dramatize or enact dynamization, possibly imagining events unfolding in the described scene. Homer's description of the shield, for instance, depicts city scenes in motion, almost as if they were living dioramas. Finally, the ekphrasis could also integrate the image with reference to the recipient's personal and cultural experiences, e.g. by supplying historical or stylistic context and adding hermeneutical interpretation of the image.

In the following sections, we provide a brief overview of the domain of ekphrasis and annotation, along with a discussion of related works. We then introduce a formal model for describing ekphrasis and apply it to a case study based on the writings of Giuseppe Raimondi. Finally, an annotator prototype designed to leverage the proposed ekphrastic ontology is presented.

2. RELATED WORKS AND DOMAIN

Given its dyadic structure, linking textual and iconic referents, ekphrasis operates as and can be effectively modeled through annotations. However, it's important to stress that ekphrasis itself is not an annotation; rather, annotations provide a structured and interpretative digital representation of ekphrastic discourse. From a general point of view, annotating is the act of associating different pieces of information together (Sanderson et al., 2013). Indeed, annotations establish a 'dialogic structure' (Boot, 2009) between two entities by explicitly linking a passage of text or an image to its translation, explanation, or another artifact, thus situating and providing insights into the use of the annotated object in a specific context. Furthermore, an annotation always consists of three components, namely: (1) an *annotatum*, something being annotated, the target; (2) an *annotans*, something predicated on the target, the annotation content; and (3) an annotator, who applies the *annotans* to the *annotatum*. Finally, the relationship between the *annotatum* and the *annotans* must be meaningful (Boot, 2009).

As discussed, ekphrasis conveys complex meanings and can establish multiple forms of interaction with its iconic referent. These forms may vary in nature—sometimes focusing on iconographic details, other times on deeper iconological themes. For this reason, ekphrasis should be viewed not as a single annotation but rather as a network of annotations referencing distinct facets or interpretive levels of one or more works of art, and potentially drawing comparisons between them. More importantly, an ekphrasis is a natural-language description created by an author. Conversely, an annotation is a formal, structured representation derived from the ekphrasis, produced by a different interpreter for analytical purposes. Thus, annotations cannot be conflated with the ekphrasis proper, for they arise from a separate hermeneutic event—one that is second-order in nature and possibly driven by a different agent. Therefore, the process of dividing an ekphrasis into annotations results from a subsequent interpretive act performed by a reader other than the original author.

Formally, given an ekphrasis E, we can represent it through a set of structured annotations A:

$$E\ (modeled\ as)\ \rightarrow\ A \subseteq T\ x\ I$$

Where T is the set of textual referents (the *annotans*, or "body"), and I is the set of iconic referents (the *annotatum*, or "target"). Each annotation $a \in A$ explicitly encodes an interpretative link between specific text fragments and image details. Moreover, this separation enables the model to be deployed in scholarly tools for analyzing and interpreting literary phenomena since it distinguishes the original text from the subsequent analytical process.

Practically, we anticipate ekphrasis to be represented by multiple annotations, each capturing distinct interpretative aspects or rhetorical strategies.

From a technical standpoint, a scholarly semantic annotation platform tailored for large-scale visual heritage was introduced, leveraging linked open data and ontologies like CIDOC-CRM to facilitate detailed

annotations and provenance tracking (Chau et al., 2024). The ICON ontology, aligning with Panofsky's three levels of iconological interpretation, offers a granular representation of artistic subjects and symbolic meanings (Sartini et al., 2023), while the MythLOD ontology interlinks ancient mythological motifs with visual representations (Pasqual & Tomasi, 2022). While these approaches underscore the potential of semantic frameworks for visual heritage and iconological studies, ekphrasis as a rhetorical and intermedial phenomenon still lacks consideration.

3. GENERAL EKPHRASTIC ANNOTATION ONTOLOGY

To leverage the theoretical framework discussed above, the present work introduces the General Ekphrastic Ontology (hereafter GEkO), a conceptual model aiming at representing ekphrasis within the digital domain. It reuses standard ontologies such as the Library Reference Model Object-Oriented[1] (LRMoo), the CIDOC Conceptual Reference Model[2] (CIDOC-CRM), the Web Annotation Data Model standard[3] (WADM), the Historical Context Ontology[4] (HiCO), and the Provenance Ontology[5] (PROV-O).
As a case study, we compiled a dataset of ekphrases written by Giuseppe Raimondi, an Italian writer and art critic active until the late 1970s[6]. His archive, housed at the University of Bologna, offers a rich collection of heterogeneous materials, including handwritten notebooks with notes and drafts on artistic phenomena. The ekphrases were manually transcribed from these notebooks and encoded in XML for analysis.
The model consists of two main components: the ekphrasis layer and the annotation layer. The ekphrasis layer focuses on the class `geko:Ekphrasis`, which defines the relationship between two key elements: the textual component (`geko:hasTextualReferent`), represented as an `lrmoo:F2_Expression`, and the visual component (`geko:hasIconicReferent`), represented as an `lrmoo:F1_Work`. Additionally, this layer specifies the form of ekphrasis using the property `geko:hasForm`, distinguishing between mimetic ekphrasis (`geko:Mimetic`), where the visual referent exists, and notional ekphrasis (`geko:Notional`), where the visual referent is fictional or lost. Specifically, `lrmoo:F1_Work` has been selected to represent cases of notional ekphrasis, where the referenced visual item does not exist as a material object. The ontology also includes the property `dct:creator`, identifying the author of the ekphrastic text. Therefore, an instance of `geko:Ekphrasis` represents the original, holistic textual description of an artwork, as authored by the scholar (in this case, Giuseppe Raimondi). It is a piece of natural language discourse existing independently of any digital modeling activity.

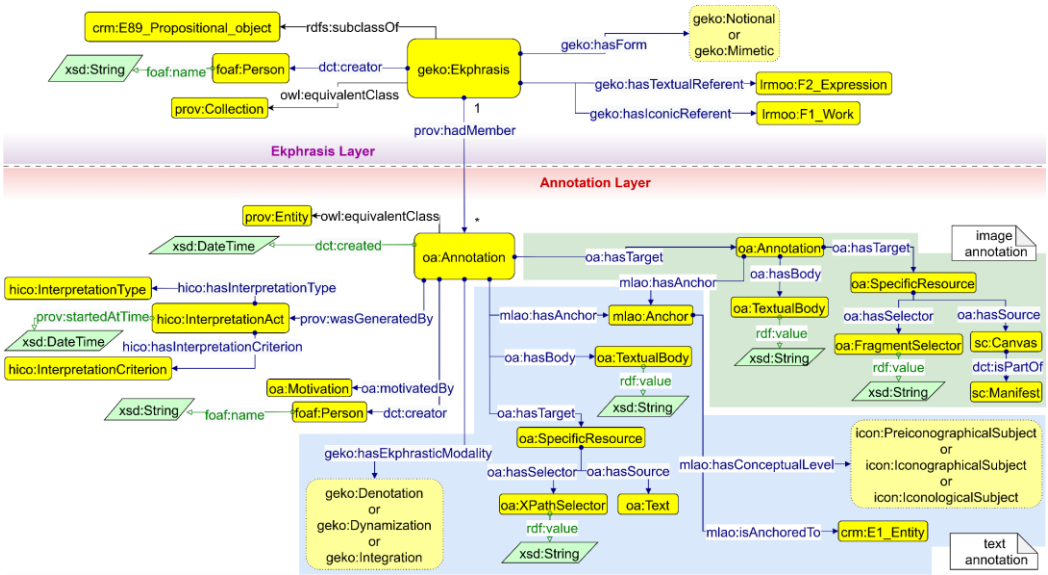


Figure 1. The formal model for General Ekphrastic Ontology.

[1] https://repository.ifla.org/items/94aedb49-2d6e-4a6d-9974-f33abb7e3c0e.
[2] https://cidoc-crm.org/.
[3] https://www.w3.org/ns/oa.
[4] http://purl.org/emmedi/hico.
[5] https://www.w3.org/TR/prov-o/.
[6] For an overview on Giuseppe Raimondi see (Tortora, 2021).

The second layer of the model considers ekphrasis as being composed of multiple parts (`prov:Collection`, `prov:hadMember`), each represented by an individual annotation. To address the specificity of ekphrasis rhetorical modalities, the property `geko:hasEkphrasticModality` was introduced, linking annotations to one of three rhetorical modes (Cometa, 2012): `geko:Denotation`, `geko:Dynamization`, and `geko:Integration`. These categories reflect distinct discursive strategies within ekphrasis: `geko:Denotation` captures the referential naming or listing of visual elements (e.g., "a red cloak," "three figures"), `geko:Dynamization` refers to the textual animation of the image, where static elements are narrated as actions or processes (e.g., "the goddess strides forward"), and `geko:Integration` denotes the blending of the visual content with the observer's interpretive or emotional response, embedding the artwork into a broader semantic or narrative frame (e.g., "a divine presence invades the scene"). This provides annotators with a granular framework for tracing rhetorical phenomena within the text. As such, an instance of `oa:Annotation` represents a formalized portion of this textual description, explicitly identified, isolated, and digitally annotated by a modeler for a particular analytical goal. Thus, the annotation captures the interpretive perspective, intent, and methodological framework of the annotator. Annotations, following the WADM specification, are atomic entities and thus cannot contain other annotations. Collections or grouping constructs (e.g., `prov:Collection`) should instead be used explicitly to group annotations that share common interpretative or rhetorical features.

Moreover, the annotation layer integrates the Multi-Level Annotation Ontology[7] (MLAO) (Pedretti et al., 2024). MLAO introduces the concept of the `mlao:Anchor`, which specifies the intentional object of the annotation (i.e., the specific conceptual entity an annotator considers during the annotation process), potentially distinct from the annotation target itself (Bradley, 2012). For instance, in the annotation of "a radiant halo encircles the saint's head", the textual span is the annotation target, while the `mlao:Anchor` might refer to the conceptual notion of holiness or divine election invoked by the halo. The anchor allows further refinement by linking to a conceptual level (`mlao:hasConceptualLevel`), chosen among iconographic stages (`icon:PreiconographicalSubject`, `icon:IconographicalSubject`, `icon:IconologicalSubject`). Additionally, the anchor can be linked via `mlao:isAnchoredTo` to the corresponding "real-world entity" (`crm:E1_Entity`) referenced by the annotation, such as a specific character in the ekphrastic text—e.g., St. Peter, identified through an authority file or controlled vocabulary. This entity could overlap with the iconic referent of the ekphrasis, maintaining a coherent connection between the model's levels. The anchor itself is shared by the ekphrastic annotation and the image annotation produced on the image, maintaining a connection between text and image at the data level. Each ekphrastic annotation has at least two targets (`oa:hasTarget`), both linking to informational structures. One target connects to a digital reproduction of the image via a canvas in a IIIF manifest (`sc:Canvas`, `sc:Manifest`), while the other references a digital reproduction of the text (`oa:Text`), which corresponds to the textual referent of the general ekphrasis. Additionally, the model incorporates the HiCO ontology to align annotations with acts of interpretation (`hico:InterpretationAct`), categorized into specific types (`hico:InterpretationType`) and further defined by interpretation criteria (`hico:InterpretationCriterion`). This alignment ensures a structured representation of the interpretive process.

[7] http://w3id.org/mlao.

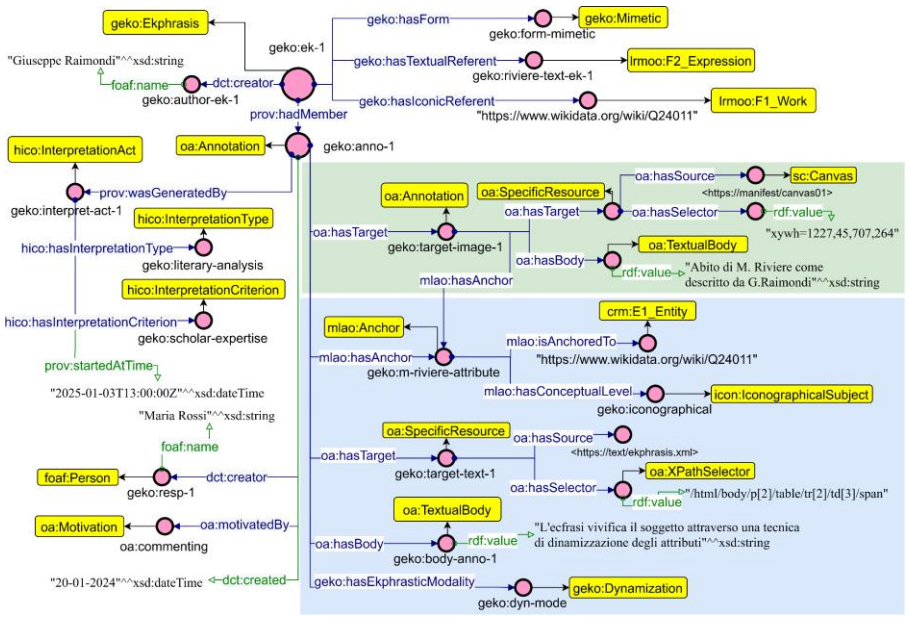


Figure 2. An example of an annotated ekphrasis: G.Raimondi's text on J.A.D. Ingres' painting *Mademoiselle Caroline Rivière*.

An example from Raimondi's case study helps illustrate the theoretical framework discussed. Figure 2 describes Raimondi's ekphrasis on *Mademoiselle Caroline Riviére* (1806), a painting by Jean-Auguste-Dominique Ingres, alongside an annotation example. The ekphrasis adopts a mimetic form, as it describes an existing artwork. The textual referent originates from a passage in Raimondi's notebook titled *I disegni "romani" di Ingres*. *Agosto 1958*, written in 1958, while the iconic referent corresponds to Ingres' painting, linked to its Wikidata entity for precise identification. The ekphrastic annotation selects a specific passage from the text where the annotator identifies a dynamization—an "animated" description of Mademoiselle Riviére's dress[8]. This annotation is anchored to an iconographical subject, namely the figure of Caroline Riviére to which it is conceptually tied. Additionally, the annotation is linked to a IIIF image annotation which visually highlights the woman figure. Finally, the annotation represents an act of literary interpretation and analysis performed by scholars based on their expertise.

4. EKPHRASTIC ANNOTATOR PROTOTYPE

To evaluate the relevance of the GEkO model we developed a prototype annotation client[9] (Fig. 3). The prototype is built as an extension of the existing Mirador[10] Annotation plugin and shares its core features and part of its user interface while incorporating the possibility to produce annotations according to classes.

In particular, this extension was designed to support ekphrastic text (Fig. 3, left side) as well as multi-level semantic annotations, while integrating domain standards into the JSON-LD model produced through the Mirador annotator. The annotator prototype incorporates NER functionality to facilitate the recognition of person names, fictional characters, and descriptive attributes.

[8] "[...] Infilato fra il busto leggero e i lunghi guanti di vitello giallo, il boa, il serpente di piuma (vuoto all'interno) di latte spumoso, allunga infinitamente le sue spire come il collo timido del cigno di Leda. Il capino del cigno, del boa, si perde in un ciuffo di alghe verdi."

[9] The annotation prototype client is available at the following link: https://github.com/friendlynihilist/geko-annotator.

[10] https://projectmirador.org/.

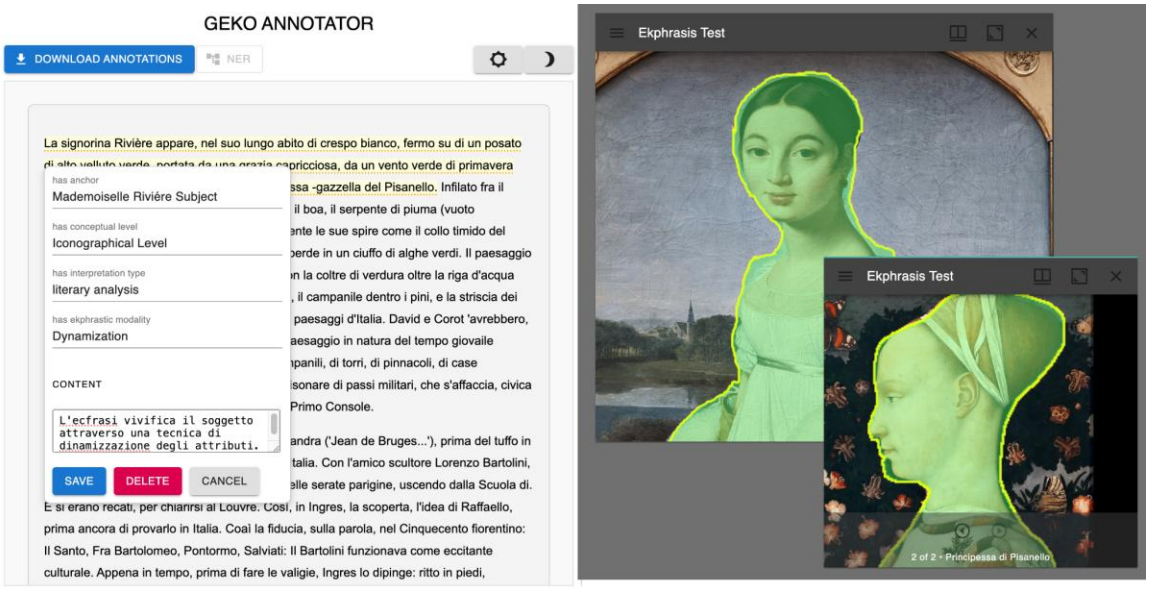


Figure 3. A screenshot of the annotator prototype. When a textual annotation is manually selected, the corresponding iconic annotation is highlighted.

From a User Interface perspective, a user can annotate both the image and the text. By highlighting a portion of the text or selecting an area of the image, a popup with autocomplete select dropdowns appears, allowing to populate GEkO properties by defining values from a taxonomy or to create new entries on the fly. The custom dropdowns dynamically inject extended metadata into a JSON-LD, which is later serialized in RDF and stored in a triplestore. This setup ensures that the annotations are stored persistently and fully interoperable. Finally, after having manually created annotations, SPARQL queries based on the Competency Questions were employed to validate the adherence of the model to the extended data schema[11]. Future work will include conducting structured evaluations with domain experts and digital humanities practitioners to assess and refine the prototype's usability and practical impact in real-world scholarly contexts, ensuring GEkO effectively meets diverse scholarly needs.

5. RESULTS AND FINDINGS

The study introduces the GEkO formal model, designed to represent ekphrasis in the digital domain. GEkO effectively bridges textual and visual references by extending and integrating established domain ontologies. The model captures crucial rhetorical modalities of ekphrasis (denotation, dynamization, and integration) and enables multi-level semantic annotations linking text and images. Specifically, the study analyzed a dataset of Raimondi's ekphrases, showcasing how textual elements dynamically interact with visual references. Using GEkO, the text-image relationship was formalized as structured annotations, highlighting the rhetorical techniques and interpretive layers embedded in Raimondi's prose.

The prototype annotation client, developed to evaluate the GEkO model, enables users to create semantically enriched annotations by linking portions of text to specific areas of an image while dynamically integrating metadata into RDF outputs.

The primary goal of the GEkO model is currently focused on scholarly documentation and enhanced information retrieval. However, future developments will concern further extensions of the model, considering, in particular, its presence in inventories and wills to trace the chain of custody and verify the authenticity of artworks over time, as well as providing automatic ways (e.g. leveraging Semantic Web reasoners) to compare and suggest candidate annotations and exploring the use of AI-driven image recognition to automatically link visual elements with corresponding textual references.

ACKNOWLEDGMENTS

This work was partially funded under the National Recovery and Resilience Plan (PNRR), Project PE 0000020 CHANGES - CUP J33C22002850006, Mission 4 Component 2 Investment 1.3, funded by the European Union - NextGenerationEU.

[11] The resulting knowledge graph was validated against competency questions to ensure it adheres to domain requirements. All relevant files can be found in the Zenodo repository: https://doi.org/10.5281/zenodo.14743534.

REFERENCES

Aub, M. (1992). Jusep Torres Campalans. Sellerio. (Original work published 1958).

Boot, P. (2009). Mesotext: Digitised Emblems, Modelled Annotations and Humanities Scholarship. Amsterdam University Press. https://doi.org/10.5117/9789085550525.

Bradley, J. (2012). Towards a Richer Sense of Digital Annotation: Moving Beyond a 'Media' Orientation of the Annotation of Digital Objects. Digital Humanities Quarterly, 006(2).

Chau, T. K., Jaquet, D., & Kenderdine, S. (2024). Towards an Annotation Data Model for a Scholarly Semantic Annotation Platform in Visual Heritage: A Case Study Using the Murten Panorama. In Bikakis, A., Ferrario, R., Jean, S., Markhoff, B., Mosca, A., Nicolosi Asmundo, M. (Eds.), SWODCH'24: International Workshop on Semantic Web and Ontology Design for Cultural Heritage. CEUR Workshop Proceedings. https://ceur-ws.org/Vol-3809/paper1.pdf.

Cometa, M. (2012). La scrittura delle immagini: Letteratura e cultura visuale. Raffaele Cortina Editore.

Eco, U. (2002). Sulla letteratura. Bompiani.

Hollander, J. (1995). The gazer's spirit: Poems speaking to silent works of art. University of Chicago Press.

Krieger, M. (1992). Ekphrasis: The Illusion of the Natural Sign. The John Hopkins University Press.

Pasqual, V., & Tomasi, F. (2022). Linked open data per la valorizzazione di collezioni culturali: Il dataset mythLOD. AIB studi, 62(1), 149–168. https://doi.org/10.2426/aibstudi-13301.

Pedretti, C. T., Bocchi, M. F., Tomasi, F., & Vitali, F. (2024). What Do We Annotate When We Annotate? Towards a Multi-Level Approach to Semantic Annotations. In Salatino, A., Alam, M., Ongenae, F., Vahdati, S., Gentile, A., Pellegrini, T. & Jiang, S. (Eds.), Knowledge Graphs in the Age of Language Models and Neuro-Symbolic AI. (pp. 370–385). IOS Press. https://doi.org/10.3233/SSW240030.

Rossetti, D. G. (2010), Scritti, poesie, lettere. Abscondita.

Sanderson, R., Ciccarese, P., & Van de Sompel, H. (2013). Designing the W3C open annotation data model. Proceedings of the 5th Annual ACM Web Science Conference, 366–375. https://doi.org/10.1145/2464464.2464474.

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, 16(3), 1–38. https://doi.org/10.1145/3594724.

Tortora, M. C. (2021). Un profilo di Giuseppe Raimondi (1898-1985). KEPOS, Anno IV(I), 130–155.

Vouilloux, B. (2005). La peinture dans le texte: XVIIIe-XXe siècles. CNRS éd.

Winckelmann, J. J. (2001), Pensieri sull'imitazione. Aesthetica Edizioni. (Original work published 1755).

Preservation of Memory and Digital Cultural Heritage

Life and Death of DH Projects: A Preliminary Investigation of Their Lifecycles in Italy

Erica Andreose[1] , Giorgia Crosilla[1],
Remo Grillo[2,3] , Gianmarco Spinaci[2,3]

[1] Digital Humanities and Digital Knowledge, Department of Classical Philology and Italian Studies, University of Bologna, Bologna, Italy, erica.andreose@studio.unibo.it, giorgia.crosilla@studio.unibo.it

[2] Patrimonio Culturale nell'Ecosistema Digitale, Università di Bologna, Bologna, Italy

[3] I Tatti - The Harvard University Center for Italian Renaissance Studies, Firenze, Italy, rgrillo@itatti.harvard.edu, gspinaci@itatti.harvard.edu

ABSTRACT (ITALIANO)

La significativa proliferazione dei progetti di Digital Humanities (DH) a cui si è assistito negli ultimi anni ha comportato altrettante sfide relative alla conservazione digitale. Infatti, una parte dei progetti risulta ad oggi non più accessibile a causa di problemi quali l'obsolescenza tecnologica e il mancato mantenimento dei server.

Questo studio rappresenta un primo tentativo di disamina del ciclo di vita e della persistenza in rete dei progetti DH in ambito italiano, considerando 270 progetti realizzati dal 1998 al 2024, provenienti da fonti diverse quali atti di conferenze, piattaforme online di centri di ricerca e aggregatori di iniziative del settore. Dall'analisi è emerso che su 250 progetti con URL identificato, 213 sono raggiungibili, 31 sono accessibili parzialmente e 37 presentano problemi di risoluzione degli indirizzi web (con sito non funzionante o dove il reindirizzamento non è corretto). Inoltre, esaminando i progetti non disponibili, lo studio ha rilevato una loro durata in vita mediamente di 3 anni e 9 mesi. I dati evidenziano una crescita costante del numero di nuovi progetti a partire dagli anni 2000, con un incremento significativo nell'ultimo decennio, seguito da un calo dal 2022 in poi. Ciò che emerge sono delle fragilità legate alle infrastrutture digitali nel settore delle DH, che evidenziano la necessità di adottare strategie strutturate a garantirne la sostenibilità e l'accessibilità nel tempo. L'obiettivo finale del lavoro è sensibilizzare sui rischi sistemici connessi alla conservazione sul lungo termine dei progetti di ricerca del settore, al fine di garantire una maggiore persistenza e tutela dei contributi legati al patrimonio culturale e accademico.

Parole chiave: progetti DH; obsolescenza tecnologica; preservazione digitale;

ABSTRACT (ENGLISH)

In recent years, there has been a significant proliferation of Digital Humanities (DH) projects, but this growth has been accompanied by increasing challenges in digital preservation. In fact, many of these projects are not currently accessible due to issues such as technological obsolescence and server maintenance difficulties.

This study represents the first attempt to examine the online lifecycle and persistence of DH projects in Italy through an analysis of 270 initiatives from 1998 to 2024. The chosen research methodology combines data from multiple sources, including conference proceedings, research center websites, and existing project aggregators, to track projects' availability and longevity.

The analysis revealed that 250 projects had associated websites, with 213 currently functioning, 31 partially inaccessible, and 37 experiencing URL-related issues (with a non functioning website or having an incorrect redirection). Examining discontinued projects, the study found an average lifespan of 3 years and 9 months. Additionally, the data highlights a steady increase in new project launches since the early 2000s, with notable growth during the last decade, followed by a decline in net growth from 2022 onwards.

Our findings unveil the fragility of digital infrastructures in DH, highlighting the need for enhanced strategies to ensure the sustainability and accessibility of these initiatives over time. By addressing systemic risks and prioritizing preservation planning, the field can better safeguard its contributions to cultural and scholarly heritage.

Keywords: DH projects; technological obsolescence; digital preservation;

1. INTRODUCTION

The field of Digital Humanities (DH) is an interdisciplinary domain in constant evolution, with projects often tied to temporary funding and rapidly changing technologies. The lifespan of a DH project reflects not only technical aspects but also the sustainability of research practices. The loss of access to such projects entails the disappearance of data, tools, and knowledge, invalidating the time and resources invested and hindering progress in specific micro-disciplinary fields.

This research provides a preliminary analysis of the state of DH projects in Italy, focusing on their longevity and the factors influencing their lifecycle. Key questions include: What is the expected lifespan of a DH project? How many projects are “born” and ”die” over time?

It is worth noticing that this study focuses exclusively on DH projects whose output is publicly accessible online via a website or digital edition. As such, it does not account for projects that exist only locally or whose results are disseminated through non-digital means. This is a structural limitation, as it captures only a portion of the broader DH landscape, though likely the most visible and fragile over time. Using project URLs, we track their online activity to assess their status and establish their initial presence. This methodology facilitates systematic evaluation and lays the groundwork for scalable computational approaches applicable to international contexts.

For this study, a DH project’s "life" begins with its first verifiable online presence and ends when it becomes inaccessible or loses essential functionalities. To avoid the binary implications of "death", we describe projects no longer functional or accessible as "unavailable." We also identify intermediate states, including partially functional projects, irregularly maintained projects, and those no longer updated but still online.

Focusing initially on the Italian context has enabled us to conduct a controlled preliminary study, allowing for manual data verification and ensuring the accuracy of our findings. This approach has been instrumental in identifying methodological challenges and refining our framework. These insights will inform future expansions to encompass DH projects on a global scale, contributing to strategies for enhancing their long-term sustainability and accessibility.

2. RELATED WORKS

While extensive research exists on the disappearance of digital resources[1], our work focuses specifically on DH projects, an area where studies remain limited and primarily concern long-term sustainability methods. DH projects differ from general digital resources due to their purpose of preserving cultural and humanistic heritage, often acting as an alternative, complement, or enhancement to traditional physical preservation practices. If DH projects aim to achieve this goal of long-term preservation, their lifespan on the web should ideally be unlimited. However, this is often not the case, as many projects face challenges such as technological obsolescence, lack of active maintenance, and eventual inaccessibility.

Previous studies on the longevity and decline of DH projects have primarily focused on the North American context. For instance, VandeCreek (2022:99) highlights the importance of institutional support in mitigating project decline, noting that academic projects are 4.15 times more likely to sustain their digital resources than non-academic ones. Furthermore, it also reveals that 68% of early funded projects (1996-2003) remain accessible online.

The longevity of DH projects has also been analyzed by *Shelf Life: Identifying the Abandonment of Online Digital Humanities Projects* (Meneses & Furuta, 2019). Their study, based on an analysis of URLs from conference abstracts within the Alliance of Digital Humanities Organizations, concludes that the average lifespan of unmaintained online DH initiatives is approximately five years before disappearance. However, the methodology used in this work has its limitations: it includes URLs that are irrelevant to the analysis, such as those linked to external tools or unrelated projects, and excludes projects not presented at conferences. Furthermore, the exclusive reliance on HTTP responses provides only a partial understanding of a project's status, failing to capture complex issues such as functional failures of seemingly active websites.

Moreover, the inability of preserving digital humanities beyond a few years has been investigated further in the *Sustainability of digital humanities projects as a publication and documentation challenge* (Edmond & Morselli, 2020). They identify evolving needs, limited resources, inadequate sustainability planning, and

[1]Chapekis Athena, Bestvater Samuel, Remy Emma, and Rivero Gonzalo, *When online content disappears*, Pew Research Center Data Labs, May 17, 2024, https://www.pewresearch.org/data-labs/2024/05/17/when-online-content-disappears/. Accessed January 12, 2025.

"orphaning" as primary causes of project discontinuation[2]. Using the Bamboo Project[3] (Dombrowski, 2014) as a case study, they illustrate how late attempts at sustainability following the end of funding could not save the initiative. Key factors in the project's failure included insufficient funding, lack of planned interoperability, and a lack of strategic planning[4] and a cohesive vision among stakeholders. Broadening the discussion, technical neglect and the discontinuation of software, such as Adobe Flash Player, a phenomenon often referred to as "bit rot", pose significant threats to the sustainability of DH projects (VandeCreek, 2022: 93).
To our knowledge, this study represents the first effort to examine the sustainability and lifespan of DH projects in Italy, addressing this critical gap in the literature.

3. METHODOLOGY

To analyze the lifecycle of Italian DH projects, we defined geographical and thematic boundaries of what is considered part of this group (see the Introduction). We deemed these projects' URLs relevant for this analysis so that we can understand the projects' health state based on status codes and Internet Archive's Wayback Machine[5] snapshot dates of HTTP requests against their URLs. On the other hand, we imply that a project is its URL, and this is not always the case. Also, Internet Archive snapshots are estimates, not a comprehensively reliable point in time describing the health of a website. Currently, the literature lacks an aggregator of projects with the characteristics above. Hence, we adopted the methodology presented in (Spinaci et al., 2022) in an attempt to create such an aggregator in the form of a dataset, integrating external sources manually assessed.
With these premises, the experiment started by detecting the typologies of *sources* where information about projects could be found: aggregators, proceedings, and research center websites. For the aggregators, we scraped the DH Awards website[6], a website where the DH community can nominate resources, and an international committee vote for the nominated based on categories. Because of its main drawbacks (i.e., it lists international projects and tools alike), we drastically removed many entries that did not conform to our target projects. Other three project aggregators are the KNOT Catalogue[7], an archive containing digital scholarly resources among projects and objects, the dataset "Classification of web-based Digital Humanities projects" (Battisti, 2024) and the "Catalogue of Digital Editions" (Franzini, 2012) from where we extracted the ones relevant to the experiment[8]. For the proceedings, we manually assessed the AIUCD Conference Proceedings[9] dated 2013 and those ranging from 2016 to 2024, looking for articles describing or referencing projects, preferably with URLs. Another set of conference proceedings is IRCDL[10], the "Information and Research Science Connecting to Digital and Library Science", of which we manually assessed the period from 2005 to 2019. The last typology container considered is the research center websites. We started by collecting AIUCD-affiliated research centers[11] and integrating the list with the missing ones of our knowledge. We scraped each of their websites to get a list of their published projects. Although their coverage is far from "complete", these sources have been selected based on their availability and relevance to provide a comprehensive overview of DH projects in the Italian landscape.
During the manual assessment, each project was checked to agree to the conformity parameters defined in the previous sections. This step's output is a list of projects per source, each with its corresponding URL.
The second step is to merge the project lists. We decided to review the list's content before applying computational techniques to disambiguate and eliminate the duplicates. We noticed some projects repeated with different URLs, a behavior explainable by their physiological evolution over time. These cases were handled manually, keeping only the row with an existing website, i.e., with a resolving URL.

---

[2] Robin Camille, *The final death(s) of digital scholarship—An ongoing case study of DH2005 projects* [Conference presentation], Digital Afterlives Symposium, Bard Graduate Center, New York, NY, March 1, 2019. https://robincamille.com/presentations/death_of_digital_scholarship/. Accessed January 12, 2025.
[3] Dombrowski, Quinn. T*owards a Taxonomy of Failure*, January 30, 2019. https://quinndombrowski.com/blog/2019/01/30/towards-taxonomy-failure/. Accessed April 8, 2025.
[4] One of the attempts of formalization of good principles and practices to avoid early obsolescence and inaccessibility is *The Endings Project*, https://endings.uvic.ca/index.html. Accessed January 12, 2025.
[5] https://web.archive.org/. Accessed January 12, 2025.
[6] http://dhawards.org/. Accessed January 12, 2025.
[7] https://projects.dharc.unibo.it/knot/ . Accessed January 12, 2025.
[8] The ongoing project ATLAS: A Digital Cartography of DH Projects. GitHub Pages, 2024. https://dh-atlas.github.io , when completed, could be implemented as a source for enhancing the dataset.
[9] https://umanisticadigitale.unibo.it/pages/boa. Accessed January 12, 2025.
[10] https://dblp.org/db/conf/ircdl/index.html. Accessed January 12, 2025.
[11] https://www.aiucd.it/centri/. Accessed January 12, 2025.

After this preliminary review, we could automatically remove rows with duplicate URLs. The final result is a list containing 270 Italian Digital Humanities projects, of which 3.1% from Catalogue of Digital Editions, 4.4% come from the DH awards website, 12.2% from IRCDL proceedings, 15.9% from AIUCD proceedings, 24.8% from the KNOT Catalogue and 39.6% from Italian DH institutions.
The next step is to enrich the data with additional information. We encountered cases where a showcase website or a landing page existed (quite common when public grants fund projects) while, to the best of our knowledge, we could not find the project URL. This has been noted for 20 projects. For the remaining 250 projects, we controlled the status code of HTTP requests against their URL and analyzed their coverage over time by accessing the Wayback Machine APIs through a custom script. This script retrieves snapshots of each website, which, when filtered based on 2xx status codes, can provide an estimate of the period during which the website was or is available. Wayback Machines's crawlers do not give the exact date when a website appeared or disappeared from the web, but they offer a temporal guide. We can infer that a website was published before the first snapshot timestamp. Regarding the last snapshot, if a website is not resolved, we can infer that the website went offline after the last snapshot timestamp. In addition to the script, we manually inspected the Wayback Machine snapshots of unavailable projects to perform a qualitative check and determine whether the projects existed by analyzing the website's content. For each project, we evaluated whether it had a website, if it was accessible, whether its content was available, and if we encountered an outdated or inactive URL. In the latter case, we manually verified whether the project had been migrated to another domain and whether there was a redirect from the old one.

4. RESULTS

As of 11 April 2025 (the date cutoff that defines this analysis), this preliminary study identified 270 Digital Humanities projects (available online[12]) from Italy, spanning the years 1998–2024. Among these, 20 do not yet have a project website, although a platform had previously been announced. Of the remaining 250, 213 have a functioning website, while 31 are partially inaccessible (due to permission issues, obsolete technologies, or lack of maintenance). Additionally, 37 projects face URL-related issues: 26 do not resolve, and 11 lack a proper redirect after migrating to a new domain.

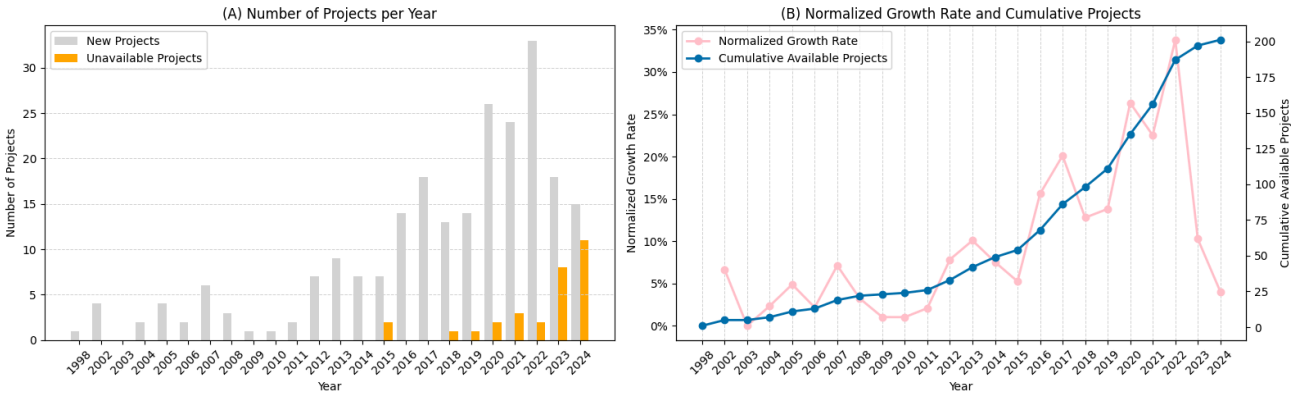


Figure 1. Growth Rate and Number of DH Projects in Italy per year

The resulting data outline an evolution of DH projects over time, as illustrated in (Fig. 1, A). The trend indicates a steady increase, with negligible fluctuations, in new online project launches per year starting from the early 2000s, with a significant surge in activity during the last decade. This increase is likely also correlated with the broader institutionalization of DH in Italy, including the establishment of new dedicated research centers and, therefore, funding opportunities (Sprugnoli et al., 2019). The surge in new projects also correlates with global trends in the DH field, as evidenced by Google Trends data reflecting increasing interest and activity in the domain nationally.[13]
The Growth Rate (Fig. 1, B), normalized by the number of cumulative projects, further contextualizes this trend. The chart reflects the annual difference between new and not-anymore-available projects,

[12] https://github.com/gspinaci/Vita-e-morte-DH-projects. Accessed April 11, 2025.
[13] https://trends.google.it/trends/explore?date=2004-01-01%202025-04-11&q=digital%20humanities&hl=it. Accessed April 11, 2025.

emphasizing the consistent growth of the DH landscape. Despite fluctuations, the positive net difference observed in most years indicates a general expansion of the field, especially during the period from 2016 to 2021, which saw the most substantial increase in the number of active projects in absolute terms. However, the declining growth from 2022 onwards might suggest an emerging gap between project launches and project discontinuations, possibly reflecting the challenges of maintaining long-term sustainability amidst growing complexity or resource constraints or that after more than 20 years, a critical mass of projects is being reached whose natural life cycle is coming to an end. It could also be due to natural project cycles or the conclusion of funding for upcoming releases tied to current national initiatives (e.g., PNRR/PRIN), potentially leading to a new increase in activity between 2025 and 2026. In either case, among discontinued projects, the average lifespan was calculated based on the period between the first and last snapshots recorded by the Wayback Machine. Preliminary findings suggest that, among the currently unavailable projects, their average lifespan was found to be relatively short, with the majority ceasing activity within 3 years and 9 months of their launch. However, this estimate should be interpreted with caution, as the current number of unavailable projects is relatively small.

While the increasing number of DH initiatives signals growth and innovation, the high rates of unavailability raise concerns about the durability of their contributions. The limited dataset and geographic focus constrain broader generalizations but provide a crucial starting point for more extensive investigations. Future expansions could integrate data from other contexts, facilitating cross-national comparisons and enabling comprehensive assessments of global DH sustainability.

5. CONCLUSIONS AND FUTURE DEVELOPMENTS

This study has provided an initial assessment of DH projects in Italy, identifying 270 initiatives and highlighting trends in their availability and sustainability. While the increasing number of projects reflects growing institutional and scholarly interest, the average lifespan of currently unavailable projects, from our list, — 3 years and 9 months — underscores challenges presumably tied to temporary funding cycles, technical obsolescence, or limited strategic planning, among others that might be worth investigating in a future study. One of the limitations of this project lies in the way data is aggregated and cleaned, as well as in the limited subset of DH projects analyzed. Due to the current lack of an aggregator that comprehensively tracks all past and ongoing DH projects, manual assessment was necessary. However, this approach is not a deterministic method for categorizing projects. Further standardized distinctions within DH project aggregators, such as the efforts undertaken by the KNOT Catalogue and ATLAS, could help eliminate the need for manual evaluation and reduce ambiguities during dataset cleaning.

To build on these findings, future research should move beyond aggregation and consider qualitative aspects of DH projects, such as their scale, funding sources, and institutional sponsorship. Understanding the internal histories of projects, including periods of downtime or partial unavailability, could provide valuable insights into the dynamics of their lifecycles. Additionally, developing a taxonomy of risk factors, such as reliance on outdated technologies or insufficient maintenance, could guide strategies to improve sustainability. While this research does not aim to offer a dynamic and continuously updated observatory, the dataset and the data collection workflow have been made openly available for replication and expansion. This "snapshot" approach provides a reproducible baseline for further studies, allowing for future analyses and tracking of trends over time. Scaling the methodology to automate parts of the project identification and verification process is also crucial for broader applications. While the manual process used in this study ensured accuracy through proper human investigation and implicitly assessing the principle of "Findability" as embodied in the FAIR principles (Tóth-Czifra & Edmond, 2020), it was time-intensive and error-prone, highlighting the need for more efficient approaches on a larger scale. Expanding the research internationally would enable cross-national comparisons and might uncover region-specific factors affecting DH project longevity. By addressing these areas, future studies can contribute to ensuring the sustainability and accessibility of DH projects, preserving their scholarly and cultural impact for the long term.

BIBLIOGRAPHY

Battisti, T. (2024). *Classification of web-based Digital Humanities projects leveraging information visualisation techniques* (1.0.0) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.14192758

Dombrowski, Q. (2014). What Ever Happened to Project Bamboo? *Literary and Linguistic Computing*, *29*(3), 326–339. https://doi.org/10.1093/llc/fqu026

Edmond, J., & Morselli, F. (2020). Sustainability of digital humanities projects as a publication and documentation challenge. *Journal of Documentation*, *76*(5), 1019–1031. https://doi.org/10.1108/JD-12-2019-0232

Franzini, G. (2012). A Catalogue of Digital Editions. Zenodo. https://doi.org/10.5281/zenodo.1161425

Meneses, L., & Furuta, R. (2019). Shelf life: Identifying the abandonment of online digital humanities projects. *Digital Scholarship in the Humanities*, *34*(Supplement_1), i129–i134. https://doi.org/10.1093/llc/fqy079

Spinaci, G., Colavizza, G., & Peroni, S. (2022). A map of Digital Humanities research across bibliographic data sources. *Digital Scholarship in the Humanities*, *37*(4), 1254-1268.

Sprugnoli, R., Pardelli, G., Boschetti, F., & Del Gratta, R. (2019). Un'Analisi Multidimensionale della Ricerca Italiana nel Campo delle Digital Humanities e della Linguistica Computazionale. *Umanistica Digitale*, *3*(5). https://doi.org/10.6092/issn.2532-8816/8581

Tóth-Czifra, E., & Edmond, J. (2020). The risk of losing the thick description: data management challenges faced by the arts and humanities in the evolving fair data ecosystem. Digital Technology and the Practices of Humanities Research, 126.

VandeCreek, D. (2022). "Where Are They Now? The 2020 Status of Early (1996–2003) Online Digital Humanities Projects and an Analysis of Institutional Factors Correlated to Their Survival". *Preservation, Digital Technology & Culture*, *51*(3), 91–109. https://doi.org/10.1515/pdtc-2022-0011

Research on Street Art in the Digital Space

Aleksandra Tselikova
University of Bologna, Italy - aleksandra.tselikova@studio.unibo.it

ABSTRACT (ENGLISH)

The purpose of this proposal is to research methods for preserving memories of street art objects in virtual space, as well as potential alternatives to preserve them. A single database of works worldwide can be created using photographic documentation in a common digital archive, which is considered the most suitable option.This proposal examines the potential problem of creating digital archives and preserving memory in the online space, as well as the potential use of databases. Possible uses of digital works include maps that display existing and disappeared street art objects, and educational initiatives that utilize preserved street art in digital spaces. Creating an online repository accessible to all can help increase the inclusiveness of the urban environment by making street art available to more people, and also underline the importance of intercultural dialogue in a single urban environment.
Keywords: street art; digital archives; cultural memory; virtual maps

ABSTRACT (ITALIANO)

Ricerca sulla Street Art nello Spazio Digitale. L'obiettivo di questa proposta è quello di ricercare metodi per preservare la memoria degli oggetti di street art nello spazio virtuale, oltre a potenziali alternative per preservarli. Una banca dati unica delle opere mondiali può essere creata utilizzando la documentazione fotografica in un archivio digitale comune, che è considerata l'opzione più adatta. Questa proposta esamina il problema potenziale della creazione di archivi digitali e della conservazione della memoria nello spazio Internet, nonché l'uso potenziale di database digitali. Gli usi possibili delle opere digitali includono mappe che mostrano oggetti di street art esistenti e scomparsi, e iniziative educative che utilizzano l'arte di strada conservata negli spazi digitali. La creazione di un archivio online accessibile a tutti può contribuire ad aumentare l'inclusività dell'ambiente urbano rendendo l'arte di strada accessibile a più persone e sottolinea inoltre l'importanza del dialogo interculturale in un ambiente urbano condiviso.
Parole chiave: street art; archivi digitali; memoria culturale; mappe virtuali

1. INTRODUCTION

The preservation of street art is one of the most pressing issues in the field of contemporary art of the 20th and 21st centuries. As a young and developing art form, it is increasingly gaining approval from both professional communities of artists and researchers of contemporary art, as well as from audiences who encounter street art daily in open urban spaces. By addressing sharp social, political, historical, and cultural issues, street art, including murals, graffiti and stickers, is often “protected” through relocation to enclosed institutions or *in situ* conservation, such as protective glass (Pagnin et al., 2023). However, these preservation methods frequently contradict the very essence of street art — its integration within open and natural urban environments (Cadetti, 2020). Street art becomes fragile, gradually collapsing under the influence of the climate, the work of urban utilities and authorities, painting over by other artists or vandals. The ephemeral nature of street art makes it particularly relevant to consider ways to digitize artists' work, similar to what can be seen in the field of performances and short-term installations (Owens, 2018: 25). The context of Italian cities emphasizes that they receive a large number of tourists every year, underlining the importance of studying this issue.

It is crucial to emphasize from the beginning the distinction between street art and new muralism in an urban setting. Street art often features a wide variety of works, typically smaller in size, fast and dynamic in making, oftenly illegal, and they are made using stencil techniques, bombing with spray and markers, and stickers. The new muralism is a vast current that is only beginning to be actively considered by the academic community. It contains conceptual and space-appropriate high quality murals, often large scale and not always, but more than street art, coordinated with the community and administration. It is interesting that the New Muralism arose from Latin American street art, showing a social context such as the Mexican Revolution. New murals are also included inside of them a message for the public. Such murals often harmoniously fit into space, gaining much wider acceptance than classical street art (Mezzadri, 2021: 2515-2516). The preservation of high-quality murals will be significantly higher than, for

instance, the wall work.

Street art, especially high-quality ones like "Madonna with a Gun" in Naples by British artist Banksy or the mural "Dante" in Ravenna by Brazilian artist Kobra, reflect the country's openness to intercultural dialogue and the expansion of contemporary art objects in its space, which means they do not deny the consideration of issues of digitalization of controversial but vibrant street art (Amato, 2015: 12-13). Some good digital archives already exist and combine not all, but many works of art that are presented in GLAM, like Google Arts and culture database (Wani, Ganaie, & Ali, 2019).[1] At the same time, there is still no single archive of art works on the streets of cities. Moreover, not even an active project on researching this type of street art archives for improving them. Monumental works of art with the high significance to humanity are most often protected and restored, while even high-quality works of modern street art, even if the authors wish, are doomed to "death". The memory of even those works that were saved through photo-capture is lost even in infinite Internet space due to the lack of a single and accessible digital database. It is difficult to deny that the quality of the work, having polisemia and not destroying other objects, can complement the complex urban space, and their disappearance sometimes makes the space more opaque (Paris, 2021). The presence of professionally executed street art facilities allows to decorate the city and organize local residents and tourist routes more interactive and memorable. If the transfer to real institutions seems to some artists wrong and violates the real nature of street art — being available and located on the open street, then the creation of digital archives often does not contradict the idea of artists. This can be seen in the example of many projects, for instance, the active idea Street museum of art in 2010, which brings together artists from different cities, with different backgrounds on one city site and fixing the resulting and changed urban reality. As a result, the website is available for examination as one small archive.[2] An equally interesting project is Tales of Golden Apples, in the Republic of Tatarstan, Russia.[3] It shows a whole archive and map of interconnected street art objects. However, it does not exhaust all the possibilities and all the artists who exist and work outside of this project. Very similar to this idea is Street Art Storage in Saint Petersburg, Russia.[4] They try to save the pieces of street art into the museum's closed space, but with showing it in virtual places, such as social networks.

Particular attention is paid to how modern tools and methodologies can be applied to protect this type of cultural heritage and further present it to a more global audience. The proposal attempts to explore and study the possibilities and means of preserving street art in digital form. The use of such ways to display and preserve valuable street art objects is applicable to Italy and other countries in the world, allowing street art. The research related to the topic of preservation of memory and digital heritage, which is in the available conference topics, aims at finding and eliminating the shortcomings of possible ways of preserving objects of street art in digital design. There are many theoretical sources for studying this issue today. Some sources consider the topic of street art conservation through different methods, but do not provide a full analysis of the prospects of a digital street art archive. For instance, Pagnin et al. (2023) shows different methods of preservation from environmental threats, but their research does not address the potential for creating a comprehensive digital archive of street art. Some of the works analyze how to find the significance and meanings of street art in urban spaces, like Paris (2021) and Amato (2015), and how to preserve it outside of the streets, like in the works Ciancabilla & Omodeo (2016) and Ciancabilla (2015). A lot of dialogues were conducted with artists, because their opinions, in the end, are the main, that is seen in the part with interview Cadetti (2020). Exploring existing experiences of preserving art or culture in digital form is presented, for example, Owens (2018) and Wani et al. (2019). The article Mezzadri (2021) examines the issue of conservation of works in urban spaces, including murals, using striking examples and with the participation of institutions in real life, but it might be important for the creation of digital spaces for such a process. However, they do not consider all existing options. An interesting alternative to real street art is the broadcasting of images using a projector. This option is explored in the paper by Jung (2015). One of the most famous and significant works describing the origins of street art and its influence on the modification of the city is the book Chalfant & Cooper (1984). It was created by photographers and street art researchers, and has been reprinted several times. Finally, a large base will be papers describing the action of the FAIR principles and copyrights in digital spaces, such as Wilkinson et al. (2016) and Yarc (2020). The attention is paid to the geographic information system (GIS) and usage of it in the cultural sphere in the research Carlucci et al. (2012). Their research is more

[1] https://artsandculture.google.com/ (cons. 05/01/2025)
[2] https://www.streetmuseumofart.org/ (cons. 08/01/2025)
[3] https://almetpublic.art/en (cons. 08/01/2025)
[4] https://street-art-storage.com/ (cons. 02/04/2025)

technical, which provides information about the real-world uses of databases and the law cases associated with them.

2. RESEARCH OBJECTIVES

The main objectives of the study may be considered as:

- Analysis of existing methods for the digital documentation of street art objects and possible online archives and repositories collecting and displaying street art objects.
- Study of FAIR principles (findability, accessibility, interoperability, and reusability) regarding their application to the issue of preservation of street art, as well as the main contradictions that may arise when analyzing the compatibility of these principles and street art (Wilkinson et al., 2016).
- Evaluation of the integration of digital platforms and metadata to ensure the sustainability and accessibility of digital street art collections.

3. METHODOLOGY

The main methods used in this work are:
Ethnographic approach.

- Interviews with artists and a population survey. The main task will be to conduct new interviews with artists, conduct a survey of viewers and, possibly, projects, and also conduct an analysis of existing interviews and conversations with artists.
- Field research, in the form of observation and description of processes of creation of objects of street art. Using this method, photographic field notes will be taken.
- Monitoring social networks with reference and placement of street art. The main social networks to research will be Instagram, Tiktok and Telegram.

Digitalization and visualization of street art.

- Photo-capture and scanning of street art objects for further consolidation into archives. This can be done with the use of a digital camera, mobile photography and AdobeScan scanning.
- Mapping of existing street art objects and possible integration of existing and disappeared objects into routes. Such a project can be created using simple and accessible open platforms, such as Google Maps, as well as more complex ones, such as StoryMaps Blog or Datawrapper.[5] When integrating street art into digital form, the implementation of geographic information systems (GIS) helps improve the infrastructure of the archival system. GIS records the spatial location of each artwork in the form of geographic coordinates, i.e. latitude and longitude, and is often used in both the cultural and user fields. This allows researchers, the general public, and other institutions to spatially access the artwork in its original location, which is essential for understanding the contemporary physical environment in which the artwork was located. Additionally, recording the date of creation, documentation, deterioration, or removal allows the archive to add a dynamic temporal layer of information, which is important for perishable street art. This feature makes it easier to track changes over time and provides insight into the relatively transient quality of street art, its life cycle, urban shifts, or the impact of policy. GIS also provides multi-layered visualizations that allow users to sift through data based on specific attributes, including neighborhood, style, artist, or subject matter. This transforms the archive from a simple static repository into an active resource for dynamic research, education, cultural policy, or even comparative and thematic analysis (Carlucci et al., 2012). Mapping art at a global or city level gives users an intuitive understanding of the geography of street art.
- Study of creation and use of digital graffiti as an alternative to digital archives of street art. A study of available technologies and programs used to create 3D graffiti in the form of projectors and mapping software, as well as a comparison of examples of projection mapping from different cities, including their goals, technical implementation, and audience opinion (Jung, 2015: 361-365).

Analysis of data from active online archives of street art objects.

[5] https://www.google.com/maps (cons. 05/01/2025)

- Identifying the advantages and disadvantages of existing archives. Comparison and identification of pros and cons can be done using a scale of assessment, including user-friendly interface, website accessibility, translation into other languages, and data quality.
- Creating a proposal to optimize and improve the work of digital preservation of street art. To improve existing street art digital archives, it is proposed to develop a more intuitive interface that can provide convenient access to data for both specialists and a wider audience, as well as to implement uniform metadata standards (for example, the studied FAIR principles), including more accurate geolocation, authorship and description of objects.[6]
- Research into the possibilities of creating virtual tours and educational initiatives based on street art facilities. The study will cover existing digital tours, as well as the possibility of creating similar excursions in text or live format.

4. RESULTS AND DISCUSSION

The expected results will be a full analysis of the problems of exhibition and preservation of street art in the digital environment, as well as the possible creation and existence of the best archive of works using FAIR principles:

- Develop optimal methodology to use the recommendations of the study in the process of creating and disseminating digital archives with street art. This methodology will help to better manage the photographic survey of works, make their meta-description, carry out the determination of geolocation and combine them into stored, convenient and accessible archives. This will help improve the systematization related to the preservation not only of works of street art, but also of any objects possessing ephemeral nature.
- Identification of the problems related to the preservation of street art and accessibility of urban environments in the form of the disappearance of works due to climate, time and urbanization. The definition of complexity can help not only to preserve street art, but also to pay attention to the development of urban space.
- Application of FAIR principles to the creation of conceptual models that allow the development of common storage bases for works related to street art.
- Creation of a project of an interactive archive map, especially with usage of GIS, with various works on the example of one city. This virtual map in the future will be able to be kept for a long time in digital format, and also to be distributed, used and created on the basis of maps of other cities. Through such mapping, virtual or real tours can also be built that introduce interested people to the cultural diversity of cities.
- Assistance in creating an accessible and inclusive environment. These archives are an option for exploring and virtual travel for people who cannot visit a place, for example, for health reasons. The digital archive of street art can be considered an alternative to reviewing works, while not excluding the possibility of live visits and the existence of real original works.

As a topic of discussion, not studied but addressed in the research, questions related to methodology may arise. For example, how should archives treat copyright? The ethnographic approach used in the research can show us the opinion of the artists themselves. Studying archives such as The Street Museum of Art or Google Arts and Culture can show how artists, associations or institutions join the project. In addition, respect for personal rights or the use of Creative Commons can also be brought up in the discussion (Yarc, 2020: 270-280).

5. CONCLUSIONS

This study focuses on the possibility of preserving graffiti and street murals in the digital space, which could be a significant step toward preserving the memory of contemporary street art and passing information about it to future generations. The future practical steps are based on the conceptual framework laid down in this work. This study reflects the main problems arising from various studies, for example ethnographic, but it is important to consider a future digital project that can take its origin from this conceptual stage of work. The long-term goal is to eventually develop an accessible and free platform for modern street art, both classic disappearing graffiti and wider murals. This database/platform can directly contain a structure informed by the results of this study.

[6] https://www.go-fair.org/fair-principles/ (cons. 05/01/2025)

Before the widespread accessibility of the Internet, street art that had not been captured by a camera was often lost, highlighting the unique ephemeral nature of this art form. Modern digitization already enables the preservation of photographs, modeling of artworks, integration into routes, and mapping their locations, effectively storing and preserving the memory of possible routes in digital form (Pagnin et al., 2023: 1-5). The creation of such a digital space emphasizes possible intercultural dialogue between cities and countries, as well as promotes the development of urban space. Street art has not yet won full love among the audience, but every year, with the appearance of new bright names and with the beginning of commercialization of street art objects, which we can see in the example of Sotheby's auctions with pieces of street art works, the stigma, warning and mythological space surrounding street art began to fade, as well as the desire of institutions to preserve this type of art (Ciancabilla & Omodeo, 2016).[7]
It is important to certainly divide the quality street art, which makes the city space more vivid, from vandalism or primitive works that are perceived as hostile by people (Chalfant & Cooper, 1984). There are many important names in the field of street art in different countries that create special art, capable of their beauty or sometimes repulsive ugliness to make a place in the city unusual and congenial (Cadetti, 2020: 166-356). Hundreds of bright names surround the street art iconography, and deserve to be remembered for these works. Sometimes the key work is to destroy it, but still, many artists are keen to preserve their work as much as possible (Ciancabilla, 2015: 49-73).
Beyond this study there are many perspectives and opportunities. This work may to some extent allow changing the psychological perception of street art, but does not make it a priority, although it contributes to it. In the future, this work will allow:

- Digitize and create a complete archive or several archives of metadata related to street art.
- Develop the study of fixing non-temporary works, for example, performances and their placement in virtual space.
- Find accessible routes, maps and educational programs that will make not only virtual but real space even more inclusive.

This proposal can help develop the urban space, access to its people, and also facilitate the task of artists in the field of preservation of finished works in digital spaces. It may help to do it without contradiction with the main idea of street art. Moreover, such online archives can be filled with interested people, just as it happens in social networks where users can publish, share, comment on the visual materials of other users.
In conclusion, it should be noted that such an initiative will expand the idea of the possibility of memory retention in the virtual space of non-durable objects and actions, and also emphasizes the importance of such steps.

ACKNOWLEDGEMENTS

This research was conducted without external funding but is an important part of my academic studies at my University. I would like to express my appreciation to the street art communities, whose participation has substantially enhanced this study.

REFERENCES

Amato, F. (2015). Il lungo dialogo tra arte e geografia. Il paesaggio urbano in trasformazione: i murales nei quartieri disagiati di Napoli. Estetica. Studi e Ricerche, 2(2015). https://doi.org/10.4434/ESR.20396635.102015.01

Carlucci, R., Grava, S., & Coz, P. (2012). Il database dei beni culturali e il sistema GIS per la 'mappa del rischio' del patrimonio artistico siriano. Archeomatica, 3.

Cadetti, A. (2020). Conservare la street art. Le problematiche del muralismo contemporaneo in Italia. Edifir. ISBN: 978-8892800205.

Ciancabilla, L. (2015). The sight gallery. Salvaguardia e conservazione della pittura murale urbana contemporanea a Bologna. Bononia University Press. ISBN: 978-8869230349.

Ciancabilla, L., & Omodeo, C. (2016). Street art. Banksy and co. L'arte allo stato urbano. Ediz. illustrata (Illustrated edizione). Bononia University Press. ISBN: 978-8869231315.

[7] https://www.sothebys.com/en/digital-catalogues/public-intervention-art-of-the-street (cons. 10/01/2025)

Chalfant, H., & Cooper, M. (1984). Subway art. Thames & Hudson. Reprinted 1999. ISBN: 0-500-27320-0.

Jung, S. (2015). Effect of 3D projection mapping art: Digital surrealism. International Conference on Virtual, Augmented and Mixed Reality. Lecture Notes in Computer Science. Springer, 361-365. https://doi.org/10.1007/978-3-319-21067-4_37

Mezzadri, P. (2021). Contemporary murals in the street and urban art field: Critical reflections between preventive conservation and restoration of public art. Heritage, 4(3), 2515–2525. https://doi.org/10.3390/heritage4030142

Owens, T. (2018). The theory and craft of digital preservation. Johns Hopkins University Press. ISBN: 978-1421426976.

Pagnin, L., Guarnieri, N., Izzo, F. C., Goidanich, S., & Toniolo, L. (2023). Protecting street art from outdoor environmental threats: What are the challenges? Coatings, 13(12), 2044. https://doi.org/10.3390/coatings13122044

Paris, O. (2021). Strategie di ridefinizione semantica della città: La street art e i segni di riscrittura urbana. Ocula, 22(May), Article 2. https://doi.org/10.12977/ocula2021-2

Yarc, M. (2020). Mural mural on the wall: Revisiting fair use of street art. UIC Review of Intellectual Property Law, 19(267), 267–292.

Wani, S. A., Ganaie, S. A., & Ali, A. (2019). The digitally preserved old-aged art, culture and artists: An exploration of Google Arts and Culture. PSU Research Review, (ahead-of-print). https://doi.org/10.1108/PRR-08-2018-0026

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., & Appleton, G. (2016). The FAIR guiding principles for scientific data management and stewardship. Scientific Data, 3(160018). https://doi.org/10.1038/sdata.2016.1

Motion Visualisation of Dancers' Performances

Giacomo Alliata[1], Loïc Serafin[1], Samy Mannane[1], Sarah Kenderdine[1]
[1] Laboratory for Experimental Museology, EPFL – Switzerland – name.surname@epfl.ch

ABSTRACT (ENGLISH)

The paper presents research on a large collection of audiovisual recordings of dance performances from the Prix de Lausanne archive, exploring how dancers' motion on stage could be visualised in alternative ways. Four Motion Visualisations (MVs) are automatically generated based on features extracted from monocular video, with each MV focusing on specific aspects of the performance. Namely, human positions and velocities, dancers' cutouts and depthmaps are computed and recombined in four modes: Analytics, Skeleton, Afterburn and Particles, through a scalable computational pipeline. Finally, the paper presents the results of a user's evaluation, highlighting a preference for Afterburn and confirming that both casual users and dance experts enjoy these alternative modes of motion visualisation.

Keywords: dance archive; motion visualisation; computational pipeline

ABSTRACT (ITALIANO)

Visualizzazioni di Movimento di Performance di Danza

L'articolo presenta una ricerca su un'ampia collezione di registrazioni audiovisive di spettacoli di danza appartenenti agli archivi del Prix de Lausanne, esplorando come il movimento dei ballerini sul palco possa essere visualizzato in modi alternativi. Quattro Visualizzazioni di Movimento (MV) sono state generate automaticamente sulla base di caratteristiche estratte da video monoculari, con ciascuna MV incentrata su aspetti specifici della performance. In particolare, le posizioni e le velocità dei ballerini, i loro ritagli e le mappe di profondità vengono calcolate e ricombinate in quattro modalità: Analytics, Skeleton, Afterburn e Particles, attraverso una pipeline semi-automatizzata. Infine, l'articolo presenta i risultati della valutazione degli utenti, evidenziando la preferenza per Afterburn e confermando che sia gli utenti occasionali sia gli esperti di danza apprezzano queste modalità alternative di visualizzazione del movimento.

Parole chiave: archivio di danza; visualizzazione del movimento; pipeline computazionale

1. INTRODUCTION

In the *Convention for the Safeguarding of the Intangible Cultural Heritage*, UNESCO recognizes the performing arts, such as music, dance and theatre, as part of Intangible Cultural Heritage (ICH), and the importance of preserving them, including measures on "promotion, enhancement, transmission [and] revitalization".[1] Based on these guidelines, this paper explores novel ways to visualise and represent the motion of dancers on stage. Certainly not a novel idea, the earliest attempts at visualising motion actually go back to the 19th century with the famous chronophotography experiments conducted by Étienne-Jules Marey, who sought to abstract the human body to "pure movement" (McCarren, 2003, p.29). In this research, we worked with the Prix de Lausanne archive, an audiovisual collection of 1500 performances on stage, capturing the finals of this world-famous dance competition for young dancers. Due to the scale of the dataset, it was necessary to develop a semi-automatic pipeline, where visual features are extracted with Artificial Intelligence (AI) models and recombined to generate four Motion Visualisations (MV), each focusing on different aspects of human bodies moving in and through space. A users' perception study, conducted during the week of the Prix de Lausanne 2024 in Lausanne, Switzerland, validates the MVs created and investigate the effect of dance knowledge on users' appreciation.

2. RESEARCH CONTEXT

Over the last decades, professionals of the sector and scholars alike have developed a variety of ways to preserve ICH, including the digital capture of performances (Hou et al., 2022). In the specific case of

[1] https://ich.unesco.org/en/convention#art2

dance practices, the *i-Treasures* initiative (2013-2017) embarked on a series of digitization endeavours aimed at immortalizing the living traditions encompassing folk dances (Dimitropoulos et al., 2014; Grammalidis & Dimitropoulos, 2015). In a parallel vein, famous choreographers have preserved their works through digital means. Wayne McGregor contributed with the Google Arts & Culture Lab to create *Living Archive*, a web-based platform to explore a digital landscape of all the human poses in its repertoire of performances.[2] Similarly, William Forsythe documented its practice with audiovisual recordings of rehearsals, performances, and installations since the 1970s, resulting in the *Performative Archive* (ZKM, 2023). He also co-created *Synchronous Objects*, a research investigation into how to transform dance into choreographic objects through data (Zuniga Shaw, 2014). Such practices are also present in other fields, like martial arts, with works such as *Kung Fu Motion Visualisation*, by media artist Tobias Gremmler.[3] Similarly, visual artists Quayola and Memo Akten won the Golden Nica at Ars Electronica Prix in 2013 for their series of animated movies *Forms*,[4] generated using 3D match-moving techniques, a common method in computer-generated (CG) pipelines, on footage of the Commonwealth Games.

3. DESIGNING MOTION VISUALISATIONS

The main goal of this research was to automatically create alternative ways to visualise dancers' motions from monocular video recordings of their performances on stage. The first step therefore involved extracting various features from the audiovisual footage. Namely, we employed MediaPipe BlazePose GHUM 3D model to compute human skeleton data from the monocular videos (Bazarevsky et al., 2020). This model outputs 33 keypoints in three-dimensional space, normalised to the frame size. From the skeleton positional data, velocity information for each keypoint was then computed. Moreover, the model also created a mask, used to obtain "cutouts" of the dancers by removing the background from the frames. Finally, we used the MiDaS model for monocular depth estimation to generate depth maps (Ranftl et al., 2022). For each frame, we thus ended up with keypoints positions and velocity, dancers' cutouts and depth maps. The latter two more "visual" features are demonstrated in Figure 1.

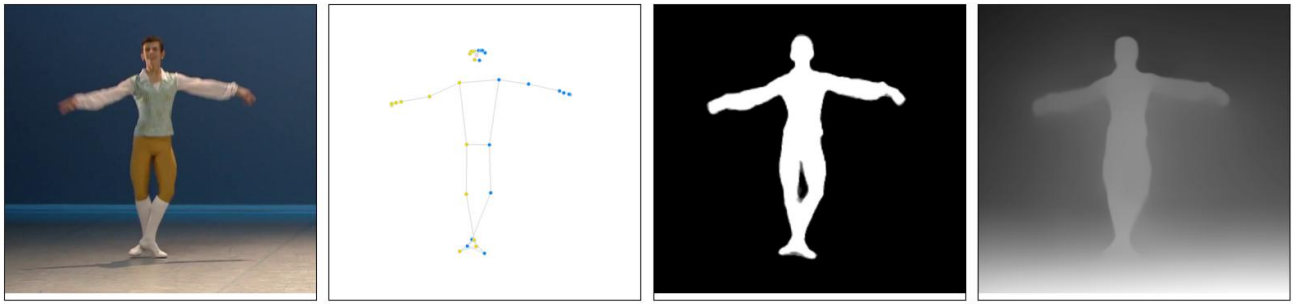


Figure 1. Visual features extracted. From left to right: original frame, keypoints, cutout and depthmap.

These elements were then combined to create four different Motion Visualisations, as illustrated in Figure 2. First, Analytics focuses on the dancers' hands and feet. It provides a simplified view of the full skeleton where only the head, the centre of the hips, the hands and the feet are visualised, tracing golden paths on a black background based on the interpolated positions. Three-dimensional boxes scale around each keypoint based on the corresponding velocity, visualising dancers' slower or faster movements in space. Second, Skeleton displays the full body of the dancers with a stick-like figure based on the keypoints extracted, inspired by Marey's early chronophotography experiments. Five positions at intervals of one second each are shown with decreasing alpha values, creating an effect of trace through time. Third, Afterburn also plays with the idea of tracing through time using this time the "cutout" of the dancers. Here as well the previous "cutouts" have temporarily increased transparency, creating an ethereal ghost-like effect that homages dance ephemeral value as a live performing art. For this MV in particular, we were inspired by Norman McLaren's 1968 experimental dance film *Pas de Deux*. Finally, Particles mixes the dancers' masks and the depthmaps with a particle simulation created in Unreal Engine where the particles

[2] *Living Archive*: https://experiments.withgoogle.com/living-archive-wayne-mcgregor

[3] *Kung Fu Visualisation*: https://www.youtube.com/watch?v=RwJG62tRjGU

[4] *Forms*: https://www.memo.tv/works/forms/

"falling" in the dancer cutout are brightly coloured while the others are in a deep blue. Furthermore, the camera is set to slowly rotate around the dancer, to recreate a more immersive feeling. This particular MV focuses therefore on the physical place that the dancer takes on stage. Due to the size of the collection, a Python script automatically process all the videos, computing the aforementioned visual features. The four MVs are then rendered in Python (Afterburn), Unity (Analytics and Skeleton) or Unreal Engine (Particles), due to technical requirements.



Figure 2. Examples of the four Motion Visualisations of dancers' performances. From left to right: Analytics, Skeleton, Afterburn, Particles.

4. EVALUATION

The four Motion Visualisations we created were showcased during the week of the Prix de Lausanne in February 2024, as part of an interactive installation whose goal was to let visitors explore the 1500 recordings of dance performances on a 4k touch screen (description of the full installation is outside the scope of this paper but interested readers can find more information in Alliata & Kenderdine (2024)). The interactive interface allowed visitors to freely switch between the original video and the four MVs. During the week, visitors were randomly approached and asked if they would answer a short survey about the MVs on an iPad, using the survey tool *muse* (Kocsis & Kenderdine, 2013).[5] Our objective was to answer the following research questions: which MV visitors have preferred and how are these MV rated in general in terms of attractiveness, confusion and interest overall (RQ1); and whether dance knowledge impact visitors' appreciation of the different MVs (RQ2).

Over the course of the week, 26 answers were collected, with a mean age of 40.2±3.6 years (ranging from 15 to 76 years old) and 73% of the respondents having identified as female. Dance knowledge was reported at 5.85±0.59 out of 10, although 8 persons indicated a 9/10, explained by the presence of many professionals in the dance field for this event. Table 1 reports the average visitors' scores for the seven Likert scale items. We observe that visitors rated positively all four of the MVs, with a preference for Afterburn and Particles. The watch times (in second) recorded by the application are consistent with the results from the survey insofar as people spent more time watching Afterburn (13.05±1.55) and Particles (15.54±3.17) than Analytics (10.74±1.25) and Skeleton (11.11±1.69). Table 1 also show that, overall, our MVs were rated highly interesting and attractive and only slightly confusing. We consider these results to be a clear indication that the MVs we designed were well received by the general public, answering RQ1, although it would seem that the Analytics one should be reworked.

To better understand these results, we conducted a correlation analysis between questions L1 to L4 and G1 to G3.[6] Figure 3 reports the results of the Spearman's computation. Note that we employed Benjamini-Hochberh correction to control the false discovery rate, due to conducting multiple comparisons. The

[5] Since the MVs were shown during a public exhibit, the requirements for the user survey were quite important. In particular, it had to be very short and could therefore not delve deeper into each of the four MV. The number of questions was kept to a minimum, with only ten questions in total: three on demographics (age, gender and dance knowledge) and the seven Likert scales items reported in Table 1.

6 This solution was preferred instead of directly asking the three questions G1 to G3 for each MV to keep the questionnaire as short as possible.

analysis reveals that the four MV could actually be split into two groups, based on their appreciation by visitors. Respectively, Skeleton highly correlates positively with Analytics ($r=0.64$, p-value<0.05) and Afterburn with Particles ($r=0.54$, p-value<0.05). Although the four MV were presented as a single group of four alternative modes of visualising dancers' motions, these results indicate that visitors were able to perceive the similarity between Analytics and Skeleton. The two MVs are indeed both focused on dancers' movements, with the first highlighting the velocity of key body parts while the second expresses the trace of the whole-body movement.

Furthermore, we note that Afterburn was significantly positively correlated with the "interesting" score ($r=0.47$, p-value<0.05), potentially explaining why it received the highest score. Conversely, Analytics was deemed the most confusing MV (although the result is not significant, $r=-0.32$, p-value$=1.00$), potentially explaining why it received the lowest score. This MV is indeed the most complex, with multiple layers of information being displayed between the positions of the selected keypoints and their corresponding velocities. Since these visualisations were shown without additional explanations, we can infer from these results that the Analytics MV was not intuitive enough and would therefore require a certain re-design. This claim is further supported by informal comments visitors expressed during the exhibit, asking more questions specifically on the Analytics MV. In particular, one visitor mentioned they would have found it easier to understand if shown alongside the original video. This is also in concordance with Afterburn being the most appreciated MV, since it is the only one showing the original dancer.

Moving on RQ2, we wanted to find out if people more knowledgeable about dance would appreciate more or less these MVs. Thus, we investigate the effect of self-reported dance knowledge on the seven scores by fitting a simple linear regression model. Table 1 however reveals that dance knowledge had no significant effect on any of the seven scores. Indeed, it explained at maximum 5.8% of the total variance (in the case of the "interesting" score G1), while it is below 1% for the majority of the individual scores L1 to L4. Although such results might seem inconclusive, we posit it actually indicates that the MV we designed can equally be well appreciated independently of dance knowledge, thus being effective at conveying dancers' motion for casual and expert users alike.

Question (Likert scale)	**Score**	**Fitted model**	**R2**	**F(1,24)**	**p-value**
How much do you like the Analytics visualization? (0-10)	6.73±0.39	6.2995+0.0738x	0.013	0.3092	0.585
How much do you like the Skeleton visualization? (0-10)	7.12±0.42	7.1036+0.0020x	0.000	0.0002	0.989
How much do you like the Afterburn visualization? (0-10)	7.96±0.33	8.2713-0.0530x	0.009	0.2262	0.639
How much do you like the Particles visualization? (0-10)	7.39±0.44	7.0651+0.0547x	0.005	0.1311	0.720
These alternative views of movements visualizations are interesting. (1-5)	4.35±0.15	4.6932-0.0594x	0.058	1.483	0.235
These alternative views of movements visualizations are attractive. (1-5)	4.23±0.15	4.5131-0.0483x	0.037	0.9117	0.349
These alternative views of movements visualizations are confusing. (1-5)	1.92±0.21	1.4467+0.0815x	0.055	1.389	0.250

Table 1. Quantitative questions and effect of dance knowledge on the corresponding scores, based on the results of a linear regression model. No significant effect was found.

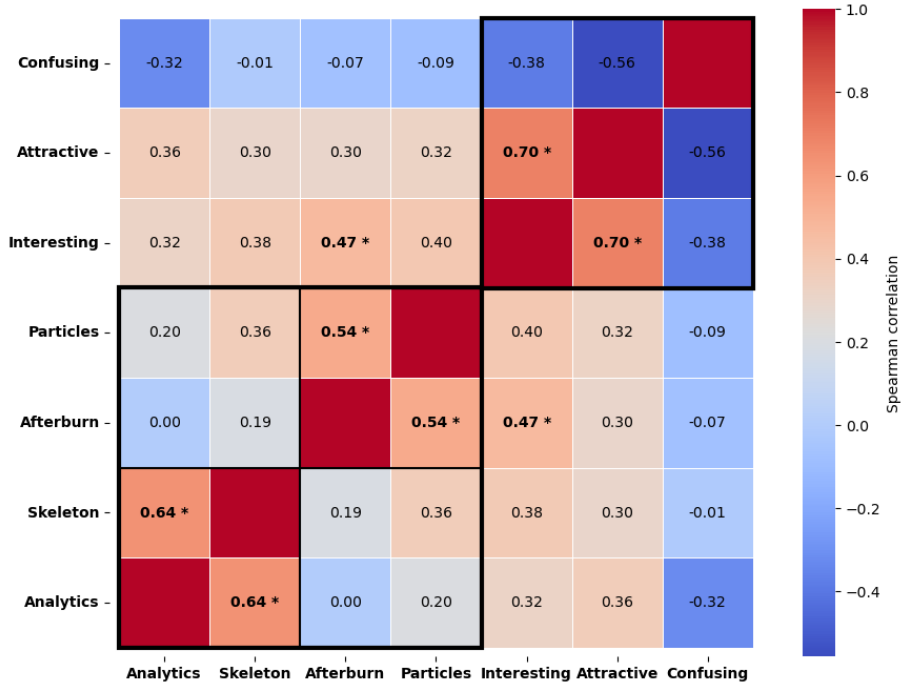


Figure 3. Spearman's correlation on the seven scales results with Benjamini-Hochberh correction. Significant values are marked in bold with an asterisk (at α=0.05).

5. KEY TAKEAWAYS ON VISUALISING MOTION

This research shows that it is possible to visualise motion in a variety of ways. The four MVs we have designed are indeed only a glimpse into an endless realm of possibilities but we believe they are sufficient to demonstrate the benefits of our approach. In particular, they show how computational approaches can be leveraged to generate these visualisations at scale, an important aspect given the politics of mass digitization of recent decades (Kenderdine et al., 2021). Motion can be operationalised by computing a variety of features even from simple monocular video recordings of dance performances (in contrast to the complex datasets generated by motion-capture technologies), and this wealth of new data can then be used to generate these "motion-as-data visualisations", hence Motion Visualisation. In comparison with the arguably greater artistic value of Gremmler's *Kung Fu visualisations* or *Forms* by Quayola and Memo Akten, our research therefore contributes an experiment at larger scales, better suited to face the challenges resulting from mass digitization.

Applying Artificial Intelligence (AI) to extract such features, however, is not a perfect process. There are inconsistencies in the results and missing data. We noticed for instance that the BlazePose model sometimes failed to detect the dancer on stage when the camera was too far away. Similarly, the dancers' cutouts and the depthmaps obtained with MiDaS are occasionally inexact because of more flowy costumes. These "errors" are then replicated in the MVs, with blanks where data is missing or unnatural shapes where dancers' bodies and costumes blend. The MVs we designed are thus a visualisation as much of dancers' motion as of the inaccuracies AI models produce. In our opinion, these "errors" should therefore not be treated as such but be considered as the kind of specificity that arises from the introduction of AI methods in these more artistic practices. It is however also important to note that new (and arguably) better models are released regularly, these errors should therefore be less and less present.

6. CONCLUSION

The four Motion Visualisations we have created are the result of a will to combine the wealth of digital archiving that has been produced with performing arts and the advances in computer vision enabling the extraction of a variety of features from audiovisual recordings. As shown in this paper, these alternative ways of visualising dancers' motions on stage can be efficiently generated at scale with computational methods to first extract relevant features and then combine them in different renderings.

We have furthermore validated that these Motion Visualisations are appreciated by casual and expert audiences alike, with a preference for Afterburn and Particles. We were particularly pleased to observe users did not find the MVs as confusing, although we note that Analytics was potentially too distracting and might benefit from a certain rework. The average watch times on the four MVs during the week were also consistent with the results of the survey, further strengthening our conclusions.

ACKNOWLEDGEMENTS

This research is supported by the Swiss National Science Foundation through a Sinergia grant for the interdisciplinary project *Narratives from the Long Tail: Transforming Access to Audiovisual Archives*, led by co-author Prof. Sarah Kenderdine (grant number CRSII5_198632, a project description is available at https://www.futurecinema.live/project/). The authors are grateful to the Prix de Lausanne for the opportunity to work on such an important and interesting archive, as well as to exhibit this work during the festival. Furthermore, the Motion Visualisation are the result of a collaboration with Swiss designer Patrick Donaldson, to whom the authors would like to express their gratitude.

REFERENCES

Alliata, G., & Kenderdine, S. (2024). *Dancing through Time: Walking through a Dance Audiovisual Archive*. The Eurographics Association. https://doi.org/10.2312/gch.20241251

Bazarevsky, V., Grishchenko, I., Raveendran, K., Zhu, T., Zhang, F., & Grundmann, M. (2020). *BlazePose: On-device Real-time Body Pose tracking* (No. arXiv:2006.10204). arXiv. https://doi.org/10.48550/arXiv.2006.10204

Dimitropoulos, K., Manitsaris, S., Tsalakanidou, F., Nikolopoulos, S., Denby, B., Kork, S. A., Crevier-Buchman, L., Pillot-Loiseau, C., Adda-Decker, M., Dupont, S., Tilmanne, J., Ott, M., Alivizatou, M., Yilmaz, E., Hadjileontiadis, L., Charisis, V., Deroo, O., Manitsaris, A., Kompatsiaris, I., & Grammalidis, N. (2014). Capturing the intangible an introduction to the i-Treasures project. *2014 International Conference on Computer Vision Theory and Applications (VISAPP)*, *2*, 773–781. https://ieeexplore.ieee.org/abstract/document/7295018

Grammalidis, N., & Dimitropoulos, K. (2015). Intangible Treasures—Capturing the Intangible Cultural Heritage and Learning the Rare Know-How of Living Human Treasures. *International Congress on Digital Heritage - EXPO*. International Congress on Digital Heritage - EXPO. https://doi.org/TBA

Hou, Y., Kenderdine, S., Picca, D., Egloff, M., & Adamou, A. (2022). Digitizing Intangible Cultural Heritage Embodied: State of the Art. *Journal on Computing and Cultural Heritage*, *15*(3), 55:1-55:20. https://doi.org/10.1145/3494837

Kenderdine, S., Mason, I., & Hibberd, L. (2021). Computational Archives for Experimental Museology. In M. Shehade & T. Stylianou-Lambert (Eds.), *Emerging Technologies and the Digital Transformation of Museums and Heritage Sites* (pp. 3–18). Springer International Publishing. https://doi.org/10.1007/978-3-030-83647-4_1

Kocsis, A., & Kenderdine, S. (2013). I sho u: An innovative method for museum visitor evaluation. In *Digital Heritage and Culture* (pp. 245–259). WORLD SCIENTIFIC. https://doi.org/10.1142/9789814522984_0016

McCarren, F. M. (2003). *Dancing Machines: Choreographies of the Age of Mechanical Reproduction*. Stanford University Press.

Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., & Koltun, V. (2022). Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-Shot Cross-Dataset Transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(3), 1623–1637. IEEE Transactions on Pattern Analysis and Machine Intelligence. https://doi.org/10.1109/TPAMI.2020.3019967

ZKM. (2023). *The Performative Archive: William Forsythe and Peter Weibel in Conversation | ZKM*. https://zkm.de/en/event/2023/02/the-performative-archive-william-forsythe-and-peter-weibel-in-conversation

Zuniga Shaw, N. (2014). *Synchronous Objects: Degrees of Unison*. Blurb Books. http://www.blurb.com/b/5174567-synchronous-objects

Entità in relazione: policies, soluzioni tecnologiche e modelli lessicali per un (eco)sistema informativo integrato

Herbert Natta[1], Michela Tardella[2], Eleonora Lattanzi[3], Gianluca Rossi[4], Roberta Maggi[5]
[1] Istituto di Matematica applicata e tecnologie informatiche IMATI-CNR, Italy – herbert.natta@ge.imati.cnr.it
[2] Istituto per il Lessico intellettuale italiano ed europeo ILIESI-CNR, Italy – michela.tardella@cnr.it
[3] Istituto per il Lessico intellettuale italiano ed europeo ILIESI-CNR, Italy – eleonora.lattanzi@iliesi.cnr.it
[4] Istituto di Matematica applicata e tecnologie informatiche IMATI-CNR, Italy – gianluca.rossi@ge.imati.cnr.it
[5] Istituto di Matematica applicata e tecnologie informatiche IMATI-CNR, Italy – maggi@area.ge.cnr.it

ABSTRACT (ITALIANO)
L'attività di ricerca che qui presentiamo è stata svolta nell'ambito del progetto *Portale delle fonti per la storia della Repubblica italiana*, un'iniziativa connotata da importanti implicazioni teoriche e metodologiche, ma anche da notevoli risvolti civili. Attraverso la collaborazione tra tre istituti del Consiglio Nazionale delle Ricerche - CNR, diversi enti pubblici e dodici istituti di cultura privati afferenti all'Associazione delle istituzioni di cultura italiane - AICI, si è cercato di elaborare un sistema concettuale e tecnologico capace di integrare risorse fra loro eterogenee, sia nei metodi di descrizione che nei formati e nelle tecnologie di trasmissione e condivisione dei dati. Il presente contributo si concentrerà sulle soluzioni individuate per gestire tale varietà, in particolare nel processo di acquisizione dei dataset provenienti dalle istituzioni di cultura private, come la progettazione e lo sviluppo di una *pipeline* ETL flessibile, capace di ricondurre la varietà dei dati in *input* al modello logico del database del *Portale*, nonché l'elaborazione di un vocabolario controllato dei livelli di descrizione archivistica.
Parole chiave: patrimonio culturale; descrizione archivistica; sistemi informativi; vocabolari controllati

ABSTRACT (ENGLISH)
Entities in relationship: policies, technological solutions and lexical models for an integrated information (eco)system - The research activity presented here was carried out within the project *Portale delle fonti per la storia della Repubblica italiana*, an initiative characterized by important theoretical and methodological implications, but also by notable civil ones. Through the collaboration between three institutes of the italian National Research Council, several public bodies and twelve private cultural institutes belonging to the Associazione delle istituzioni di cultura italiane - AICI, an attempt was made to develop a conceptual and technological system capable of integrating heterogeneous resources, both in the description methods and in the formats and technologies for data transmission and sharing. This contribution will focus on the solutions identified to manage this variety, in particular in the process of acquiring datasets from private cultural institutions, such as the design and development of a flexible ETL pipeline, capable of bringing the variety of input data back to the logical model of the database, as well as the development of a controlled vocabulary of the archival description levels.
Keywords: cultural heritage; archival description; information systems; controlled vocabularies

1. INTRODUZIONE
Il progetto *Portale delle fonti per la storia della Repubblica italiana*, ha rappresentato una sfida tanto culturale quanto tecnologica, con importanti implicazioni civili, teoriche e metodologiche, e che vede la collaborazione tra tre istituti del CNR, enti pubblici, e dodici istituti di cultura privati afferenti all'Associazione delle istituzioni di cultura italiane - AICI (cfr. Tardella *et al*, 2024).
Il progetto nasceva dalla ferma convinzione di creare uno strumento rivolto non solo ad un'utenza esperta, ma anche e soprattutto al pubblico cosiddetto "generalista", composto da non addetti ai lavori, dagli studenti delle scuole, da chi si approccia per la prima volta al mondo degli archivi.
Conseguenza naturale di questa vocazione, che potremmo dire civile, e di questa apertura al dialogo con tipologie di utenti dai profili eterogenei, è stata la creazione di uno strumento semplice e di immediata comprensione, che garantisse l'apertura e la condivisione dei dati, delle descrizioni e delle tecnologie, ma che tenesse in considerazione le esperienze pregresse e coeve in campo archivistico.

Uno strumento semplice, ma tecnologicamente evoluto che permettesse di fruire del patrimonio culturale, con particolare riguardo a quello archivistico, conservato dagli enti pubblici[1] e dagli istituti privati e messo a disposizione di questo specifico progetto, attraverso il quale si mira a rendere più agevole la consultazione dei documenti degli organi costituzionali e degli apparati amministrativi dello Stato, insieme a quelli prodotti e conservati dalle associazioni private. L'impegno principale è stato rivolto all'elaborazione di un sistema concettuale e tecnologico capace di integrare risorse caratterizzate da una notevole eterogeneità, sia nei metodi di descrizione che nei formati e nelle tecnologie di trasmissione e condivisione dei dati. In questo scenario, il *Portale* ambisce a diventare un ecosistema nel quale giungere ad una integrazione dei dati e dei sistemi, superando quel "particolarismo informativo" che caratterizza l'eterogeneità e la frammentarietà delle informazioni e che si riscontra anche nella proliferazione dei siti per la ricerca archivistica (Cfr. Cacioli, 1996). Per perseguire questo obiettivo, i gruppi di lavoro, a carattere fortemente interdisciplinare (si compongono infatti di archivisti, bibliotecari, informatici, filosofi ed esperti di linguaggi), hanno sviluppato strategie volte all'armonizzazione di patrimoni archivistici, dei modelli concettuali e dei formati di rappresentazione dei dati. Il presente contributo si concentra, in particolare, sulle soluzioni teoriche, metodologiche e tecnologiche adottate per l'integrazione dei dati in un unico modello logico e per la elaborazione di un vocabolario controllato dei livelli di descrizione archivistica.

2. LE ISTITUZIONI DI CULTURA, I FONDI, I SISTEMI INFORMATIVI

La partecipazione delle istituzioni di cultura afferenti all'AICI si è svolta secondo i criteri definiti dall'accordo quadro stipulato tra l'Associazione stessa e il CNR nel luglio del 2022. L'accordo, finalizzato a favorire la cooperazione tra le parti per la realizzazione di iniziative progettuali - in aree tematiche di comune interesse e, soprattutto, nella collaborazione al progetto *Portale delle fonti per la storia della Repubblica italiana* - prevedeva l'indizione di un bando per la presentazione di manifestazioni di interesse. La selezione delle proposte pervenute si è basata su una serie di criteri, secondo i quali i fondi acquisibili dovevano rappresentare i diversi orientamenti politici, sociali e culturali dell'Italia repubblicana, nonché la varietà tipologica delle fonti (archivistiche e bibliografiche). Per rispondere agli obiettivi di progetto, sono stati ritenuti di notevole interesse gli archivi contenenti materiali relativi all'attività istituzionale del soggetto produttore e, in particolare, quelli non ancora digitalizzati. Sono stati inoltre presi in considerazione sia la consistenza dei fondi in rapporto ai tempi del progetto e alle risorse finanziarie disponibili sia, ai fini della sperimentazione tecnologica, i software utilizzati e i formati di interscambio.
Le istituzioni di cultura selezionate e i fondi proposti per la partecipazione al progetto sono indicati nella tabella 1.

ISTITUTO STURZO	Fondo Francesco Bartolotta
	Archivio Giulio Andreotti (2 serie)
FONDAZIONE GRAMSCI	Archivio Mosca
FONDAZIONE UGO SPIRITO E RENZO DE FELICE	Fondo Nino Tripodi
FONDAZIONE LUIGI MICHELETTI	Fondo Repubblica Sociale italiana
FONDAZIONE PASTORE	Fondo Giulio Pastore
	Fondo Lamberto Giannitelli
FONDAZIONE LELIO E LISLI BASSO	Fondo Ada Alessandrini
	Fondo Lelio Basso (2 serie)
ISTITUTO NAZIONALE FERRUCCIO PARRI	Fondo Ferruccio Parri
	Carte Ferruccio Parri
FONDAZIONE GRAMSCI TORINO	Fondo Partito comunista italiano - Federazione di Torino
	Fondo Unione donne italiane

[1] L'Archivio Storico della Presidenza della Repubblica e l'Archivio Storico della Camera hanno partecipato attraverso dataset già disponibili ed esposti come linked open data. L'Archivio Storico del Senato ha invece condiviso dati in formato Csv, relativi in particolare all'anagrafica dei Senatori. Per informazioni di dettaglio si veda Tardella *et al.*, 2024.

FONDAZIONE GRAMSCI EMILIA ROMAGNA	Fondo Triumvirato insurrezionale
	Fondo Giuseppe Dozza
ARCHIVIO AUDIOVISIVO DEL MOVIMENTO OPERAIO E DEMOCRATICO	Collezioni film e audiovisivi
FONDAZIONE GIUSEPPE E SALVATORE TATARELLA	Fondo Maselli Campagna
	Fondo Fotografico
FONDAZIONE UGO LA MALFA	Fondo Fotografico
	Fondo La voce della Donna
	Fondo Ugo La Malfa
	Fondo La Malfa - Appendice

Si tratta di nuclei archivistici profondamente difformi gli uni dagli altri, sia per quanto riguarda la storia della sedimentazione delle carte, sia in relazione al loro trattamento archivistico. Realtà variegate cui corrispondeva un eterogeneo livello di descrizione e di maturazione tecnologica. Per sintetizzare, dati diversi con diversi gradi di analiticità descrittiva, prodotti con software diversi.
Le relazioni tra il CNR e le singole Istituzioni sono state normate attraverso convenzioni bilaterali, nelle quali sono stati formalizzati gli aspetti progettuali e fiscali, con particolare riguardo alle attività e agli obiettivi da raggiungere, ai tempi e alle condizioni di svolgimento e ai termini del supporto finanziario relativo alle attività condotte per il raggiungimento dei comuni obiettivi scientifico-culturali. Erano parte integrante delle stesse convenzioni anche alcuni allegati tecnici, contenenti: le informazioni di dettaglio relative al piano di progetto e ai relativi fondi proposti dalle Istituzioni; le linee guida per la gestione e la rendicontazione delle attività, e, soprattutto, le linee guida per la descrizione archivistica, la digitalizzazione e le policies di condivisione dei dati e degli oggetti digitali, fondamentali queste ultime per la definizione di criteri di descrizione e di modalità di digitalizzazione dei documenti condivisi.

3. PROCEDURE DI ACQUISIZIONE DEI DATI

L'eterogeneità delle fonti, degli strumenti di descrizione e dei formati di interscambio ha richiesto la progettazione e sviluppo di un sistema di componenti software (*pipeline*), per l'estrazione, trasformazione e caricamento (ETL) dati, flessibile, capace di ricondurre la varietà dei dati in *input* al modello logico del database del *Portale*, minimizzando la perdita di contenuto informativo.
L'analisi dei dati ricevuti ha messo in evidenza, nell'ampio margine di variabilità, pattern ricorrenti. In particolare, i formati e le strutture sono risultati dipendenti dagli strumenti di descrizione utilizzati e riconducibili a tre tipologie: documenti Word (.doc/.docx), cartelle di lavoro Excel (.xls/.xlsx), XML strutturati in base a diversi schemi, principalmente EAD e diverse configurazioni di xDams, la piattaforma per il trattamento, la gestione e la fruizione di archivi storici utilizzata da alcuni degli Istituti di cultura.
In aggiunta alle naturali differenze negli schemi dei dataset, che hanno richiesto la predisposizione di strumenti di *mapping*, sono state riscontrate generali difformità nelle modalità di rappresentazione dell'informazione temporale, sia date puntuali sia intervalli, e delle voci di indice, che hanno richiesto specifiche procedure di trasformazione, finalizzate, nel primo caso, a normalizzare la rappresentazione degli estremi cronologici e, nel secondo, ad agevolare l'identificazione degli agenti relazionati.
Sulla base di queste analisi preliminari, sono stati sviluppati algoritmi *ad hoc* per i diversi *dataset* acquisiti, basati su un modello architetturale comune. Si tratta di micro-applicazioni, sviluppate in linguaggio python e composte di: script principale (*importer*) eseguibile, contenente l'intera procedura ETL (dalla lettura del/dei dataset sorgenti alla trasformazione alla generazione di output, prima tabellari, per permettere la verifica dei risultati, procedendo infine all'inserimento dei dati nel database del *Portale*), file con le funzioni richiamate dall'*importer*, modificate in relazione al formato di input e alle procedure di trasformazione necessarie, file di setup, con la configurazione delle costanti (parametri di connessione, percorso dei dataset in input, eventuali parametri di mapping), file di mapping, con dizionari contenenti la corrispondenza tra la struttura di input e il modello della base dati del *Portale*, costruiti sulla base delle indicazioni definite dagli esperti di dominio.

In particolare, è stata utilizzata la libreria *pandas* per l'elaborazione dei formati tabellari, la libreria *xml* per il processamento degli XML e la libreria *docx* per i documenti Word.
La normalizzazione delle date è stata sviluppata mediante progressiva generalizzazione e adattamento di specifiche funzioni, definite nel file dedicato e basate sulla libreria *datetime*, in base alle caratteristiche dei dataset pervenuti. Le funzioni sono strutturate in una principale pubblica che, ricevendo in input la stringa contenente la data e l'indicazione relativa al fatto che si tratti di data iniziale o finale, restituisce in output la data correttamente formattata, richiamando due funzioni private: una che procede alla sostituzione delle stringhe relative ai mesi con la corrispondente notazione numerica, l'altra che integra giorno e/o mese dove assenti e riformattando la stringa in base al pattern richiesto, riconoscendo l'eventuale indicazione di intervalli temporali (e distinguendo quindi la data iniziale da quella finale).
Il trattamento delle voci di indice ha invece richiesto un'elaborazione più complessa, dovuta, da un lato, all'eterogeneità delle fonti e, dall'altro, alla loro rilevanza per l'attivazione di relazioni semantiche trasversali all'interno del sistema Portale. L'approccio metodologico adottato ha previsto l'entificazione dei nomi di enti/persone/famiglie/congressi (agenti) e luoghi mediante la creazione di una scheda dedicata per ogni voce. In questo modo è stato possibile sia riferire le risorse acquisite a un'entità univoca sia dare modo agli operatori di arricchire la semplice denominazione con altri dati relativi all'entità rappresentata.
Per trattare la variabilità delle voci di indice si è però resa necessaria un'elaborazione in più fasi, inclusiva sia di procedure automatiche sia di una validazione *expert based* dei risultati parziali.
In particolare, i casi più ricorrenti sono stati: i) voci di indice distinguibili nella struttura del dataset tramite specifici elementi (es. <controlaccess>), tipologia distinguibile (es. <persname>, <corpname>, <geogname>) e denominazione formattata in modo uniforme (es. 'Cognome, Nome'); ii) voci di indice distinguibili, ma non è distinguibile la tipologia e/o la denominazione non è formattata in modo uniforme.
Per la prima condizione si è generata la scheda corrispondente, riportando gli elementi della denominazione nei relativi campi e mantenendo l'identificativo (dove presente) del sistema di origine; per la seconda, si sono elaborate le voci di indice tramite librerie specifiche (quali *names_dataset*) per ricostruire, tramite ranking[2], la probabilità che un record corrisponda a nome o cognome di persona o a denominazione di ente. Prima di importare i dati, il risultato è verificato manualmente per minimizzare l'errore. Come si evince, se generalmente le voci di indice sono identificabili nei dataset importati[3], il riconoscimento della tipologia o della formattazione della denominazione, funzionale a strutturare il dato coerentemente con il sistema di destinazione (distinguendo per esempio nome, cognome, qualifica per le persone, denominazione e ulteriori denominazioni per gli enti, denominazione, luogo e data di svolgimento di congressi).
Il numero complessivo delle voci di indice importate è di 32.014 agenti, dei quali 23.614 persone, 8.355 enti, 38 congressi, 7 famiglie, e 9.938 luoghi, dei quali 6.524 riconducibili a comuni italiani, 45 stati, 19 regioni italiane e 3.354 non classificati (es. microtoponimi, città o regioni estere, ecc.).
Per gli agenti, nonostante la strategia di deduplicazione adottata in fase di importazione, che ha previsto la verifica della corrispondenza con agenti già precedentemente importati (in questi casi non è stata creata una nuova scheda, ma attivato un nuovo collegamento a schede esistenti), sono risultate inserite nel sistema 1.828 denominazioni univoche associate a più di un identificativo.
In questi casi, si è deciso di mantenere la duplicità dei record, uniformandoli però attraverso la relazione allo stesso identificativo in sistemi di rappresentazione esterni (VIAF, Geonames, Wikidata). Questa procedura di arricchimento, che ha riguardato l'intero corpus di voci importate, è stata sviluppata attraverso un flusso di lavoro semiautomatico, che ha previsto, in una prima fase, la definizione ed esecuzione di uno script di *matching*, basato sull'interrogazione delle API utilizzando come parametri gli attributi noti delle entità (toponimo, denominazione, nome e cognome, qualifica, ecc.). Le associazioni ottenute, sono state poi oggetto, in seconda fase, di una revisione *expert based*, volta a risolvere ambiguità, duplicazioni e falsi positivi.
La necessità, da un lato, di garantire la coerenza dei dati importati rispetto alle fonti e, dall'altro, di metterli a sistema in uno strumento di descrizione integrato, nel quale i dati potessero convivere

[2] Nei casi di formattazione non uniforme, la denominazione è stata tokenizzata, ogni token è stato classificato come first name (FN), last name (LN), mixed (M) in base al ranking generato da names_dataset, componendo così, per ogni stringa, un pattern, ulteriormente elaborato considerando il posizionamento del token nella stringa, la lunghezza del token e la presenza di elementi tipici dell'onomastica (es. 'Di', 'De' tipici di alcuni cognomi italiani).

[3] Per i documenti Word (.doc/.docx) le voci di indice sono state isolate o attraverso l'analisi preliminare del documento e l'individuazione di specifiche caratteristiche di formattazione o attraverso l'elaborazione dei marcatori di indice (index markers).

generando nuovo contenuto informativo, ha richiesto inoltre la definizione di vocabolari condivisi rispetto ai quali procedere alla normalizzazione dei dati importati.

4. I VOCABOLARI CONTROLLATI

La definizione delle entità relative al dominio archivistico utilizzate nel *Portale* ha posto la necessità di utilizzare un vocabolario controllato. Come è noto, un vocabolario controllato è costituito da una serie di lemmi organizzati e "definisce concetti (temi, stili artistici, autori) che vengono utilizzati come valori nei metadati. [...] Un vocabolario rappresenta quindi una lista chiusa di valori controllati consentiti per un elemento" (Guerrini & Possemato, 2015 : 99).
Per la realizzazione del *Portale delle fonti* è stato richiesto ai partner di conferire descrizioni molto analitiche e quindi in molti casi vengono presentati i singoli documenti che compongono i complessi documentari, come ad esempio fotografie, manifesti, cartoline etc.
Si è però ravvisato come la terminologia applicata dai diversi istituti culturali non fosse sempre semanticamente univoca, nonostante l'ambito disciplinare specifico[4], ma presentasse delle sfumature che dovevano essere uniformate per poter procedere all'attività di importazione dei dati e alla loro successiva esposizione come linked open data. Termini legati alla struttura multilivellare degli archivi si presentavano infatti caratterizzati da usi dipendenti da diverse interpretazioni e, pertanto, applicati secondo sensi e in modalità differenti, come, ad esempio, "Complesso archivistico", che rimanda ad insieme di documenti di varia natura, oppure di concetti quale "Collezione" o "Subfondo".
Questa distorsione semantica si ripercuote anche nelle diverse definizioni date a questi stessi concetti nell'ambito dei diversi standard, che risentono dell'ambiente culturale e, inevitabilmente, del contesto storico di produzione.
Sono stati quindi individuati e definiti[5] i seguenti 14 lemmi: Classe/Categoria, Collezione/Raccolta, Fondo, Unità archivistica, Sottofascicolo/Inserto, Unità documentaria, Altro livello, Complesso archivistico, Subfondo, Serie, Sottogruppo, Sottoserie, Articolazione interna al fascicolo, Parte di un documento.
Tale lavoro si colloca pienamente nella prospettiva di arricchire il modulo archivistico dell'ontologia ArCo[6] e permettere così l'esposizione nel *Portale* dei dati relativi alle risorse che sono state tipizzate tramite il vocabolario[7].
Lo studio ha preliminarmente preso in considerazione le tipologie in uso a livello nazionale validate dall'Istituto centrale per gli archivi (ICAR) e le ha confrontate rispetto a quanto prescritto da Encoded Archival Description[8], standard che garantisce l'interoperabilità fra diversi sistemi archivistici.
Successivamente è stata effettuata una ricerca nella letteratura scientifica per poter attribuire delle definizioni univoche alle tipologie da accettare nel vocabolario controllato. Dopo aver consolidato il vocabolario è stato necessario renderlo operativo anche nel sistema GeCa (Maggi, 2023; Natta, 2024) verificando quali tipologie fossero già presenti e aggiornandolo in modo da poter trattare in modo corretto i dati descrittivi importati dai diversi dataset.
Il *Portale delle fonti*, come abbiamo visto, viene popolato da dati derivati da sorgenti differenti: solo in alcuni casi sono già disponibili come linked open data mentre quelli provenienti dagli Istituti culturali che partecipano al progetto devono subire un processo di lodificazione. In questo frangente si sono riscontrate alcune criticità per quello che riguarda i dati a valle della importazione nell'infrastruttura GeCa. Il sistema utilizza diversi standard internazionali per la lodificazione a seconda del dominio specifico. La prima problematica affrontata riguarda il fatto che i termini del vocabolario implementato devono avere una corrispondenza con l'ontologia relativa alle risorse archivistiche, Records in contexts (RiC-O), versione 1.0[9].
Questa ontologia è basata su un modello concettuale che prevede una entità "Record Resource"[10] che si

[4] Si noti già la polisemia del termine "archivio", che può essere riferito al complesso di documenti, all'istituto di conservazione o a quella parte di un edificio destinata al deposito della documentazione e che richiede dunque una disambiguazione. Per una estesa riflessione sul termine archivio si vedano Ciandrini *et al*, 2023; Carucci, 2010; ISAD(G), 2000:10-11.
[5] Le fonti di riferimento per le definizioni in italiano sono il Glossario della Direzione Generale Archivi, il volume di Paola Carucci e Mariella Guercio, *Manuale di archivistica* (Roma, 2021) e le Norme Internazionali per la Descrizione Archivistica. Per la versione in inglese, sono stati invece utilizzati il Dictionary of Archives terminology, l'Encoded Archival Description Tag Library e la versione inglese delle le Norme Internazionali per la Descrizione Archivistica.
[6] Vedi: http://wit.istc.cnr.it/arco (cons: 07/01/2025).
[7] In tal senso il passo successivo sarà quello di rispettare le raccomandazioni previste per la pubblicazione dei dati in *linked open data*. Si veda: W3C, 2014.
[8] Vedi: https://www.loc.gov/ead/EAD3taglib/EAD3-TL-eng.html#attr-level (cons: 07/01/2025).
[9] Vedi: https://www.ica.org/resource/records-in-contexts-ontology/ (cons: 07/01/2025).
[10] Rispetto allo standard di descrizione precedente, ISAD(G), questa entità "is conceptually comparable to unit of description" (*Records In Contexts – Conceptual Model Version 1.0:* 20).

distingue in Record Set, Record, Record Part. In particolare, per i Record set esiste una ulteriore categorizzazione che nell'ontologia di Records in contexts definisce quattro tipi[11] (ICA, Expert group on archival description, 2023):

- Fonds, ovvero "the whole of the records, regardless of form or medium, organically created and/or accumulated and used by a particular person, family, or corporate body in the course of that creator's activities and functions";
- Series, ovvero "documents arranged in accordance with a filing system or maintained as a unit because they result from the same accumulation or filing process, or the same activity; have a particular form; or because of some other relationship arising out of their creation, receipt, or use. A series is also known as a records series";
- Collection, ovvero "an artificial assemblage of documents accumulated on the basis of some common characteristic without regard to the provenance of those documents. Not to be confused with an archival fonds";
- File, ovvero "an organized unit of documents grouped together either for current use by the creator or in the process of archival arrangement, because they relate to the same subject, activity, or transaction. A file is usually the basic unit within a record series".

Il 14 lemmi del vocabolario sopra elencati sono stati dunque allineati e implementati nell'infrastruttura e ricondotti alle differenti entità.
L'altra problematica in fase di risoluzione riguarda invece il rapporto tra Records in contexts e ArCo, l'ontologia del *Portale delle fonti*, e il modo in cui quest'ultima deriva queste sei tipizzazioni senza perdere l'informazione presente nel dato originario, che invece ne distingue dodici.

CONCLUSIONI

L'acquisizione e integrazione di dati da fonti eterogenee in un unico sistema informativo ha richiesto l'adozione di strategie operative, soluzioni tecnologiche e scelte teorico-metodologiche volte a massimizzare l'attivazione di relazioni latenti e, in generale, l'interoperabilità del sistema, minimizzando la perdita di contenuto informativo e l'ambiguità semantica.
La strada aperta da questo lavoro ha il chiaro e naturale obiettivo, orientato dal paradigma dei *linked open data*, di formalizzare le scelte semantiche effettuate nella definizione del vocabolario, integrandole stabilmente nell'ontologia del *Portale delle fonti*, basata su ArcO, la rete di ontologie per la descrizione del patrimonio culturale a livello nazionale.
L'auspicabile prosecuzione del progetto, oltre a prevedere un ampliamento rispetto all'arco cronologico preso in esame (1943-1953) e quindi ai contenuti, dischiude anche ulteriori prospettive.
Il *Portale* si è rivolto principalmente al patrimonio archivistico proponendo fonti custodite da Istituti culturali che si presentavano, seppur in modo vario e a un diverso livello di conformità agli standard, come dataset riconducibili a modelli di descrizione archivistica. Risulterebbe interessante un'integrazione di queste fonti, costituite per la maggior parte da documenti originali, con le risorse bibliografiche, tra cui gli studi e le ricerche condotte sin dalla nascita della Repubblica. In particolare si possono prendere in considerazione le pubblicazioni del Quirinale, della Camera e del Senato, come ad esempio la "Bibliografia del Parlamento"[12], che comprende anche i singoli contributi analitici presenti nelle pubblicazioni, suddivisi per argomento. Alla modellazione dei dati secondo ontologie specifiche del dominio bibliografico, così come fatto rispetto al dominio archivistico, si aggiungerebbe la possibilità di relazionare le risorse con il catalogo del Servizio bibliotecario nazionale per consentire all'utente il loro reperimento, nel caso non sia accessibile una loro versione in formato digitale.
Un'ulteriore evoluzione del Portale potrebbe riguardare i prodotti di natura divulgativa, audio e video, realizzati specificamente per questo progetto. Si tratta di una serie di videointerviste e podcast che contestualizzano le risorse documentali e propongono alcune tematiche fondamentali per tracciare la storia della Repubblica italiana. Questi prodotti potrebbero quindi essere oggetto di una metadatazione dettagliata e dello sviluppo di funzionalità che permettano la realizzazione di sottotitoli in più lingue, per gli audiovisivi, e la trascrizione degli interventi.Queste prospettive dunque rientrano nell'orizzonte di un progetto che non si limita a mettere a disposizione una raccolta di fonti per la storia politica e istituzionale nazionale relativa alla seconda metà del Novecento, ma mira a fornire strumenti di approfondimento orientati ad ampliare la fruizione da parte di un pubblico diversificato, favorendo e migliorando l'accessibilità alle risorse e al loro contenuto.

[11] Questi tipi riprendono quanto definito da ISAD(G), cui si rimanda.
[12] Vedi: https://storia.camera.it/bpr/ (cons: 09/04/2025).

BIBLIOGRAFIA

Cacioli, M. (Eds.). 1996. Gli archivi dei partiti politici: atti dei seminari di Roma, 30 giugno 1994, e di Perugia, 25-26 ottobre 1994. Roma: Ministero per i beni culturali e ambientali, Ufficio centrale per i beni archivistici.

Carucci, P. (2010). Le fonti archivistiche: ordinamento e conservazione, Roma: Carocci.

Ciandrini, P., Lattanzi, E., Maggi, R., Tardella, M. (2023). Archivi e contaminazioni disciplinari: dai linguaggi ai modelli, dai metodi alle tecniche. Migrazioni e contaminazioni tra le scienze. Metodi e linguaggi interdisciplinari. Laureti, S., Marras, C., & Peddis, D. (Eds.). Collana PLURIMI – IV, 2023. CNR Edizioni. ISBN 978 88 8080. DOI 10.36173/PLURIMI-2023-4

Guerrini, M., & Possemato, T. (2015). Linked data per biblioteche, archivi e musei. Editrice Bibliografica. ISBN 978-88-7075-830-6

ICA, Expert group on archival description (November 2023). Records in Contexts conceptual model. Version 1.0. International Council on Archives
https://www.ica.org/app/uploads/2023/12/RiC-CM-1.0.pdf

ISAD(G), 2000. General international standard archival description adopted by the Committee on Descriptive Standards, Stockholm, 19-22 September, Ottawa.
https://icar.cultura.gov.it/fileadmin/risorse/docu_standard/RAS_2003_1.pdf

Maggi, R., Pasciuto, T., Mazzoleni, M., Artese, M.T., Gagliardi, I., & Albertoni R. (2023). GECA 3.0 - A new tool for cataloguing and enjoying cultural heritage. La memoria digitale. Proceedings del XII Convegno Annuale AIUCD, Siena, 5-7/06/2023, 978-88-942535-7-3.

Natta, H., Rossi, G., Maggi, R. (2024). Luoghi comuni: metodi e strategie di sviluppo software in ambito GLAM, dalle voci di autorità all'esplorazione cartografica. Me.Te. Digitali. Mediterraneo in rete tra testi e contesti. Proceedings del XIII Convegno Annuale AIUCD, Catania 28-30 maggio 2024, Università di Catania. Di Silvestro, & A., Spampinato D. (Eds.). ISBN 978-88-942535-8-0. DOI 10.6092/unibo/amsacta/7927

Pavone, C., & D'Angiolini, P. (1981-1994). Guida generale degli archivi di Stato italiani. Roma: Ministero per i beni culturali e ambientali, Ufficio centrale per i beni archivistici, voll. 1-4.

Tardella, M., Maggi, R., Lodi, G., Albertoni, R., Natta, H., Rossi, G., Pasciuto, T., Ciandrini, P., Sinopoli, L., Artese, M.T., Gagliardi, I., Lattanzi, E., Ventroni, S., Tizzoni, E., Russo, A., & Porena, M. (2024). Un futuro per la memoria. Strumenti, modelli e sinergie per l'integrazione dei dati nel Portale delle fonti per la storia della Repubblica italiana. Me.Te. Digitali. Mediterraneo in rete tra testi e contesti. Proceedings del XIII Convegno Annuale AIUCD, Catania 28-30 maggio 2024, Università di Catania. Di Silvestro, & A., Spampinato D. (Eds.). ISBN 978-88-942535-8-0. DOI 10.6092/unibo/amsacta/7927

W3C (2014). Best Practices for Publishing Linked Data. W3C Working Group Note. Hyland, B., Atemezing, G., Villazón-Terrazas, B. (Eds.). World Wide Web Consortium
https://www.w3.org/TR/ld-bp/

Preserving and enhancing cultural heritage: the Digest project

Alessandra Cinini[1], Paola Marongiu[2], Eva Sassolini[3] & Monica Monachini[4]
[1]Istituto di Linguistica Computazionale "A. Zampolli", (CNR - Pisa), Italia, alessandra.cinini@cnr.it
[2]Istituto di Linguistica Computazionale "A. Zampolli", (CNR - Pisa), Italia, paola.marongiu@cnr.it
[3]Istituto di Linguistica Computazionale "A. Zampolli", (CNR - Pisa), Italia, eva.sassolini@cnr.it
[4]Istituto di Linguistica Computazionale "A. Zampolli", (CNR - Pisa), Italia, monica.monachini@cnr.it

ABSTRACT[1] (ENGLISH)
This paper aims to describe and analyze the activities related to the preservation and valorization of textual corpora of cultural-historical value, produced over a long period of time. We intend to share with the scientific community the problems related to the advancement of technologies used for data creation/processing, as well as the issues related to the transition from proprietary to standard formats to enable data sharing and interoperability. Finally, we will outline the complex transition to open data paradigms and the necessary steps for migrating data into international research infrastructures. Specifically, we will describe the translation process of Justinian's Digest, emphasizing how this has evolved alongside technological progress. Our contribution will offer the DH community a point of view on the evolution of the digitization/computerization of large textual resources, based on a real use-case. To illustrate this, we will examine how specialized translation from Latin to Italian can be enhanced using textual analysis tools. Furthermore, we will describe the transformation of both the original and translated texts into a parallel bilingual corpus; its conversion into XML TEI format; the steps taken for depositing the data in the CLARIN research infrastructure.
Keywords: texts corpora; standard TEI format; digital preservation

ABSTRACT (ITALIANO)
Valorizzare e custodire il patrimonio culturale: alla scoperta del progetto Digesto. Il contributo vuole discutere delle attività connesse alla salvaguardia e alla valorizzazione di corpora testuali di valore storico-culturale prodotti in un lungo arco temporale. Analizzeremo le problematiche legate al progresso delle tecnologie utilizzate per la creazione/elaborazione dei dati che sono mutate nel tempo, nonché l'arduo passaggio dai formati proprietari a quelli standard per una migliore condivisione e interoperabilità dei dati. Infine, descriveremo la non facile transizione verso paradigmi di dati aperti e i passaggi necessari per migrare i dati verso infrastrutture di ricerca internazionali.
In particolare, descriveremo le fasi del progetto di traduzione del Digesto di Giustiniano che hanno richiesto il supporto tecnologico del nostro gruppo di ricerca, e come questo contributo è cambiato di pari passo con il progresso tecnologico. Vogliamo offrire alla comunità DH un punto di vista reale sull'evoluzione della digitalizzazione/informatizzazione di grandi risorse testuali. Nel caso specifico, discuteremo del supporto alla traduzione specializzata dal latino all'italiano con strumenti di analisi testuale, e della trasformazione dei testi originali e tradotti in un corpus bilingue parallelo; di come questo è stato convertito in formato XML TEI; infine, delle operazioni necessarie al deposito dei dati all'interno dell'infrastruttura di ricerca CLARIN.
Parole chiave: corpora testuali; standard TEI; preservazione digitale

1. INTRODUCTION
The activities described in this paper are part of the TESTO project, which currently focuses on the development of a parallel Latin-Italian corpus of Justinian's Digest. The translation and digitization of the Digest began in the late 1990s and advanced gradually over an extended period (Schipani 1994[1996], 2005; Schipani & Lantella 2005a, 2005b, 2007, 2011). Initially, technological contributions were limited to digitization efforts. However, by the time the third book was completed, the need for more structured translation support had become evident. The fragmentary nature of the technological support provided led to challenges comparable to those encountered in many recent projects involving the digitization of archives encoded in outdated formats (Sassolini et al., 2014). The development of the project can be divided into three main phases, each closely tied to technological advancements. In the first phase, IT efforts focused on developing prototypes and stand-alone programs for acquiring the bilingual corpus and

[1] As part of a joint elaboration, Eva Sassolini specifically oversaw the drafting of §§ 1 and 5, Alessandra Cinini of §§ 3, while Paola Marongiu focused on §§ 4 and 4.1. Paragraph 2 is a joint work. Monica Monachini participated as the project manager.

enabling textual searches. These early tools were created using then-available development environments and software systems, and were mainly developed *ad hoc*, later revealing compatibility issues with the new environments. In the second phase, the growing diffusion of the Internet prompted the transition of translation support tools to web-based platforms. This shift mitigated the issues related to the evolution of software environments and fostered collaborative and shared working practices. The third and ongoing phase addresses a new set of emerging challenges, primarily concerning data protection, standardization and interoperability of representation formats, and efficient data sharing strategies.

The creation of the bilingual parallel corpus of the Digest represents a peculiar initiative in the landscape of resources for legal studies and historical linguistics. Various corpora have been developed for legal studies, both mono- and multilingual. The CLARIN infrastructure hosts a "Legal Corpora" Resource Family,[2] which includes resources such as Europarl (Koehn 2005), a collection of European Parliament proceedings in 21 languages, and the JRC-Acquis corpus (Steinberger et al. 2006), a sentence-aligned parallel corpus in 22 languages. However, to our knowledge, there remains a notable absence of multilingual corpora specifically focused on historical legal texts. On the other hand, significant efforts have been made in the digitization of historical texts. Among other initiatives, we can mention the Perseus Digital Library (Crane 1987; Crane et al. 2006), which provides a large digital collection of Ancient Greek and Latin texts; the five treebanks for Latin, now aligned with the Universal Dependencies project annotation guidelines (Haug and Jøhndal 2008; Bamman and Crane 2011; Passarotti 2019; Cecchini et al. 2020; Cecchini et al. 2020); Corpus Corporum, another freely accessible digital collection of Latin texts (Roelli 2014); Musisque Deoque (Boschetti, Del Grosso & Spinazzè 2021), a digital collection of Latin poetry. Concerning parallel/comparable corpora of historical texts, it is worth mentioning the work by Yousef and Berti (2015), who automatically aligned Ancient Greek and Latin texts with their English translations at the sentence and word level. Our work on the Digest thus brings together the challenges and methodological considerations inherent in two research domains: the development of specialized parallel corpora and the encoding and alignment of historical texts. In this paper, we will describe the activities carried out in constructing the parallel corpus of the Digest, including the text alignment procedures (section 2), the rationale behind the selection and use of the XML-TEI encoding format (section 3), and the steps that will be taken to publish the corpus in compliance with Open Access principles (section 4). Section 5 outlines our plans for future developments in our work.

2. TEXT CORPUS OF CONCORDANCES

The Justinian's Digest, together with the *Codex Iustiniani*, the *Iustiniani Institutiones* and the *Novellae Constitutiones*, is part of the *Corpus Iuris Civilis*, a collection of Roman legal texts compiled under the order of Emperor Justinian the Great (527–566 CE) between 529 CE and 534 CE (Banchich et al. 2015; Dingledy, 2016). Justinian inherited from his predecessor Justin I a weak empire: the West Empire was now under the influence of Germanic tribes and Justinian saw in the project of the *Corpus Iuris Civilis* an opportunity to unite the empire under Christianity. The Digest represents the most important part of the *Corpus Iuris Civilis*, as it offers the wealthiest collection of Roman legal texts. The Digest encompasses selected excerpts from 1528 writings produced by 39 Roman jurists, starting from the Twelve Tables (449 BCE) until the jurist Hermogenianus (350 CE), for a total of 800 years of Roman legal history (Dingledy, 2016: 5; Ribary & McGillivray, 2020). The jurists' writings are organized in 50 books, divided into five groups depending on the topic: Public Law (Book I); Private Law (Books II–XLVII); Criminal Law (Book XLVIII); Appellate Procedure and Treasury (Book XLIX); Municipal Law, Specialized Law, and Definitions (Book L). The books are divided into 432 titles. Each title is divided into laws, and laws into paragraphs. Each law is preceded by an inscription stating the name of the jurist who wrote it, and the title and volume of the book from which the excerpt was taken (Banchich et al. 2015: 5). Studying the Digest nowadays is for jurists functional to the understanding and development of a system of principles on which each state's law is based, regardless of their specific juridical systems (Schipani 2005: xxvii).
For all the above, the Digest translation project concerns a specialized field (the legal domain), with all the challenges associated with the translation of specialist concepts/terms. The translators' first need was to consult past translations to facilitate their work and ensure coherent use of legal terminology throughout the text. Therefore, when our institute was involved, our main goal was to support the translators by providing them with a bilingual concordances query system, to be used as translation memories. To do this, the Latin texts of the Digest were automatically aligned with their Italian translations by using a

[2] https://www.clarin.eu/resource-families/legal-corpora

statistical method for their synchronization, refined and guided by a network of linguistic anchors/items (Marinai et al. 1992). The semi-automatic alignment process relies on the morphological engines of the Italian and Latin languages, as well as on a bilingual dictionary, to establish a network of links between source and target text units within a predefined textual window. However, this process often breaks down when the structure of the translation diverges significantly from that of the original Latin text (Simard & Plamondon, 1998). To ensure accuracy, minimum thresholds are set for the number of detected equivalent translation units; when these thresholds are not met, manual intervention is required. This requirement has a significant impact on alignment times, which may increase considerably. Furthermore, an interface (or tool) is available to review the aligned texts, allowing for targeted adjustments in specific sections, typically used for aligning Ancient Greek and Italian.
Then, we transformed the results into bilingual parallel concordances accessible with the DBT (Data Base Testuale) textual analysis engine.[3] The tool, initially conceived as a PC software, was later transformed into a client-server application. The primary objective of the concordances corpus was to offer translators a resource where they could search for specific words or family of words in Latin and check their Italian correspondents in the previous translations.
Web development extended the range of users for these resources to scholars in other research domains. This required the implementation of new functionalities and a new interface for lower specific users, e.g. historical linguists interested in the study of non-literary Latin, semantic shift and the development of specialized lexicon.
As part of the project, we also plan to develop a bilingual digital glossary of Latin legal terms and their specialized Italian translations, drawn from the bilingual corpus. The creation of the glossary is still ongoing. We have lemmatized the books translated so far by using Latinpipe (Straka, Straková, & Gamba 2024) for the Latin texts, and Stanza (Qi et al. 2020) for the Italian translations. We are currently working on the evaluation of the lemmatization results and on devising the best strategies for the extraction of specialized/domain-specific Latin terms to be included in the glossary.

3. XML TEI FORMAT: WHY AND HOW

The process of computerizing data is a technological ecosystem and as such has specific characteristics. Nowadays, the objectives of a researcher working on this type of process cannot be exclusively oriented towards improving performance. In fact, the development of user support tools requires a different approach that takes the context into account. The aim is to make research more responsive to changes that are not only technological but also social and cultural. Achieving these goals requires continuous updates aligned with recent scientific developments, dealing with issues concerning the value of data and their dissemination. It is important for us to find new, multidisciplinary and adaptive approaches that can compete with the new challenges posed by the changing scenario. A computational linguist will work not only to improve the performance of the tools, but also their flexibility, so that they can be designed and implemented in an integrated way and respond to the varied users' information needs.
In addition to adapting/improving the tools already developed, we mapped in XML TEI the parallel corpus created by the bilingual concordances and all derived resources, with the aim of disseminating the project results, preserving them from obsolescence and making the data interoperable. We have chosen to deposit our resource in the CLARIN resource infrastructure due to its robust capabilities and commitment to expanding access to linguistic data.
As a first step, we studied a representation model in TEI. Based on the chosen model, we implemented the automatic mapping process. This resulted in the output of separate files for each book, available in both the Italian and Latin language versions. Each book has a tree structure divided into sections that has been encoded with the <DIV> tag. The ordering and organization follow the Digest's textual units in their original printed form. As shown in Figure 1, each section's tag is assigned attributes that specify its type: title; fragment; principium/paragraph; an identifier for text alignment in the two languages. More specifically, Figure 1 shows how an inscription is encoded in the TEI model.

[3] Text analysis system developed by Eugenio Picchi at CNR-ILC with various software configuration possibilities, including the contrastive consultation of digital corpora.

```
<div type="frammento" resp="POMPONIUS" corresp="#libro nono ad Sabinum">
<p>
<seg xml:id="D.18.I.13_L" ana="D.18.I.13">IDEM libro nono ad Sabinum. Sed si servo meo vel ei cui mandavero
vendas sciens fugitivum illo ignorante, me sciente, non teneri te ex empto verum est.</seg>
</p>
</div>
<div type="frammento" resp="ULPIANUS" corresp="#libro vicesimo octavo ad Sabinum">
<p>
<seg xml:id="D.18.I.14_L" ana="D.18.I.14">ULPIANUS libro vicesimo octavo ad Sabinum. Quid tamen dicemus, si in
materia et qualitate ambo errarent? ut puta si et ego me vendere aurum putarem et tu emere, cum aes esset? ut puta
coheredes viriolam, quae aurea dicebatur, pretio exquisito uni heredi vendidissent eaque inventa esset magna ex
parte aenea? venditionem esse constat ideo, quia auri aliquid habuit. nam si inauratum aliquid sit, licet ego aureum
putem, valet venditio: si autem aes pro auro veneat, non valet.</seg>
</p>
</div>
```

Figure 1. Encoding of *inscriptiones*

Regarding alignment procedures, some considerations can be made concerning the linguistic anchors that connect specific words with their potential translations. These elements, initially extracted to improve the contrastive query of bilingual concordances, are generated by automatic software procedures with a significant rate of 'noise'. Since in the context of specialized translation a literal approach to the source text is considered insufficient, the translator's intent must be interpretative. For this reason, the translated text may in some cases contain integrations/additions of various kinds (explanatory, stylistic, morpho-syntactic, etc.) that make the alignment less accurate. In fact, the translator leaves out the formal correspondence between the two languages but interprets and reformulates what is necessary to achieve an equivalence of meaning (Angelucci, 2024). In our case, this scenario requires synchronization of the texts at the fragment level rather than the word level, while simultaneously opening the possibility of extending the alignment to other languages into which the Digest has been or is currently being translated. For example, the translation of the Digest in Chinese has been in progress for years (Schipani, 2010; Luo & Colangelo, 2013; Li & Colangelo, 2016). Furthermore, we are studying methods to analyze the linguistic anchors and codify them in resources to be shared.
The textual parts added by the translators, due to their peculiarity, have been encoded with a specific TEI tag (<ADD>, as shown in Figure 2) and linked to an external resource that describes and documents with examples the various types of textual additions.

```
<seg xml:id="D.01.II.2.6_I" ana="D.01.II.2.6">
        In seguito, pressappoco nel medesimo periodo, sulla base di queste leggi furono composte le azioni
<add>processuali</add> con le quali gli uomini contendessero tra loro; si volle che tali azioni fossero certe e solenni,
affinché il popolo non ne istituisse come volesse. Questa parte del diritto viene chiamata: "azioni di legge", cioè
"azioni legittime". E così, pressoché nel medesimo periodo, nacquero questi tre diritti: le Leggi delle Dodici Tavole; da
queste cominciò a fluire il diritto civile; sulla base delle medesime Tavole vennero composte le azioni di legge. Di tutto
ciò, tuttavia, sia la scienza dell'interpretare sia le azioni erano nell'ambito di competenza del collegio dei pontefici, tra i
quali si statuiva chi in ciascun anno fosse preposto a <add>rispondere alle domande de</add>i privati; e il popolo si
avvalse di questa consuetudine pressappoco per cento anni.
</seg>
```

Figure 2. Example of textual additions introduced by the translators

Finally, we worked on reconstructing the history of the project, the actors, methods and technologies used. The information reconstructed is included in the header of the TEI corpus, but also in the Data Management Plan, required for depositing the resource in CLARIN, which we will discuss in detail in the next paragraph. Once the translation is completed, it will be possible to specialize the TEI model and annotate the texts at a lower level.

4. DATA PRESERVATION: MIGRATING THE DATA INTO THE CLARIN INFRASTRUCTURE

The original purpose of the project was to provide a valuable tool to be used as a translation memory, to ensure translation normalization. Since then, the goal of the project has gradually broadened to ensure data sharing with the scientific community. For this reason and to guarantee the preservation of data and resources created in the long history of the project, we decided to convert the aligned bilingual corpus into

the XML TEI interoperable format (§3) and to share it with the humanities and social sciences communities (DHs) via the CLARIN infrastructure, all by adhering to the best practices underlying the FAIR[4] principles. Within this distributed digital infrastructure, the individual centers are the network nodes. For our project we addressed ILC4CLARIN, a type B CLARIN center, which is part of CLARIN-IT, the Italian CLARIN ERIC node.[5]
By doing so, we aim not only to preserve our developed resources but also to align them to the CLARIN framework and facilitate interoperability with linguistic services offered by the infrastructure. This strategic decision aligns with CLARIN's mission to enhance data discoverability and facilitate interoperability across linguistic and cultural studies. Additionally, the CLARIN infrastructure offers us the opportunity to publish our textual data within the evolving landscape of Linked Open Data (LOD), which is an important aspect of contemporary digital scholarship. In particular, CLARIN-IT is developing a robust Linked Open Data (LOD) platform, designed to enhance data sharing and collaborative research in the context of the H2IOSC project (Horizon 2020 International Open Science Cloud).[6]
The LOD platform aims to integrate diverse datasets from various fields into a cohesive framework, thereby promoting an interconnected approach to research practices. By utilizing Linked Open Data principles, the platform provides a means to interlink datasets, allowing researchers to navigate through related information seamlessly, regardless of the original source or format. This not only enriches the context of individual datasets but also facilitates interdisciplinary research by bridging gaps between disparate fields. By positioning our data in the 'cloud' of linguistic Linked Open Data in the future, we can ensure that it is not only accessible but also interconnected with other resources. This interlinking will enable richer, more complex queries and analyses, thus enhancing the overall utility of our data for the community.

4.1. The data management plan

As part of the process for ensuring adherence to the best practices in terms of Open Science and data FAIRification, we have included in the project activities the creation and continuous update of a Data Management Plan (DMP).
A DMP is a document that follows the life cycle of the data, from data collection, organization, documentation and quality control to data sharing, archiving and preservation (Michener 2015). The DMP is supposed to be a 'living document' that evolves together with the advancement of the project. Drafting the first version of the DMP at the very beginning of the project has been a fundamental step to ensure a structured view of our data cycle. It allowed us to think ahead of the nature of our data, including e.g. copyright and privacy-related issues, expected outcomes, overall costs throughout the project's lifespan, best practices and decisions in terms of licenses that will be applied to the data.
In our DMP we included, among other types of information, legal implications of the data creation/annotation process; long-term data preservation (type of infrastructure or repository where our data will be archived); actors involved in the data life cycle; expected costs of carrying out the project activities (Giglia 2023).
To fill out the DMP for the TESTO project we relied on ARGOS,[7] a platform powered by OpenAIRE[8] that allows for creating, modifying and archiving machine-actionable DMPs. The completion of the first version of our DMP posed various challenges, due to the very peculiar nature of the data and to the history of the TESTO project.
As mentioned earlier, the project started in the late 90s and is still underway to this day. There have been various changes throughout the years, including the project funding, directors, partners, assistants (especially translators) and technologies used to process the data. Moreover, the project as it is conceived today involves different university partners, together with our institute. Delivering the DMP proved more challenging than initially anticipated, as it required reconstructing the project's history to ensure proper

[4] https://www.go-fair.org/fair-principles/

[5] https://ilc4clarin.ilc.cnr.it/, hosted by our Institute.

[6] https://www.h2iosc.cnr.it/

[7] ARGOS offers a simplified, guided step-by-step process to complete the DMP and makes sure the user has entered all information relevant to their specific project data (https://argos.openaire.eu/splash/about/how-it-works.html).

[8] "OpenAIRE is a Non-Profit Partnership of 50 organisations, established in 2018 as a legal entity, OpenAIRE A.M.K.E, to ensure a permanent open scholarly communication infrastructure to support European research" (from https://www.openaire.eu/about).

credit was given to all former directors and collaborators. The information related to the project history, including funding, editors of the published volumes, translators for each paragraph in the published volumes, and current partners of the project, constitutes the metadata for the bilingual corpus in XML TEI, and will be accessible as an external resource linked from the header of the TEI files. The DMP details the type of metadata and information included in the TEI header.
The long history of the project has also affected the type of technologies used for data processing and analysis. As mentioned in section 2, the alignment of the Latin-Italian texts was carried out with DBT, a proprietary text analysis system that also enables the online contrastive consultation of the bilingual corpus. In the perspective of sharing the data with the research community, and adhering to the FAIR principles, we decided to convert the data and resources generated through DBT into the XML TEI format. The same workflow will be applied to the books that will be published until completion of the translation process. To ensure full transparency on the workflow and data processing, we decided to include this information in the DMP, providing details about the type of software used and the process developed to obtain the TEI files.
Additional issues stem from the fact that the project is still ongoing, such as determining the appropriate licensing for the texts and finalizing the bilingual glossary. One of the milestones of the project is a first deposit of the data, i.e., the bilingual corpus in TEI format, on the platform ILC4CLARIN. However, the translation of the Digest is still incomplete, with books 36–50 yet to be translated. Additionally, the TEI model developed from previous translations has not yet been tested against the new material; it's possible that the Latin or Italian texts in the remaining books contain types of information not present in earlier ones, which may require different treatment. Therefore, we decided to apply a restricted license to this first deposit, which only allows the people working on the project to access the data. We explained the reasons for our decision in the first version of our DMP.
Finally, as mentioned earlier, one of the outputs envisaged for the completion of the project is a bilingual glossary of Roman legal terminology. However, as the translation is still ongoing, this output can only exist in a provisional form. Therefore, we decided to announce the publication of the bilingual glossary as a future outcome of the project, leaving the details for the updated version of the DMP.

5. FUTURE PERSPECTIVES

The project is still underway, and finishing the translation remains the primary goal. We have defined the TEI model of the corpus but, as mentioned in § 3, various options exist for the representation of textual data in this format. It is our intention to study them further. The nearest future perspective is to archive all 50 books in Latin-Italian into the CLARIN infrastructure under an open license.
We also plan to experiment with integrating services and resources into centralized platforms, such as the CLARIN Switchboard. Further testing will be devoted to the publication of the project data as Linguistic Linked Open Data, incorporating them into the resource cloud for historical languages. This step will broaden the potential for sharing, hopefully fostering synergies with similar resources, projects and initiatives, e.g. by bridging translations of the Digest into other languages. This scenario is a concrete possibility, especially considering the ongoing translation of the Digest into Chinese.

ACKNOWLEDGEMENTS

Research Project PRIN 2022 PNRR P20224NJLK – SH2 – TESTO "Translating, Encoding, Sharing The Origins". From the Littera Florentina to an open-access Italian translation of Justinian's Digest". Piano Nazionale Ripresa e Resilienza (PNRR) M4C2 – Investimento 1.1 "Progetti di Ricerca di Rilevante Interesse Nazionale (PRIN)- funded by the European Union – NextGeneration EU" – CUP B53D23032480001.

REFERENCES

Angelucci, M. (2024). Computabilità e traduzione interlinguistica: considerazioni sull'impiego dell'Intelligenza Artificiale in ambito traduttivo. *DigItalia*, 19(1), 183–197.

Bamman, D. & Crane, G. (2011). The Ancient Greek and Latin Dependency Treebanks. *Language technology for cultural heritage: Selected papers from the LaTeCH Workshop Series*. Sporleder, C., Bosch, A., Zervanou, K. (Eds.), Berlin, Heidelberg: Springer. 79–98.

Banchich, T. M., Marenbon, J., & Reid, C. J. (2015). The Revival of Roman Law and Canon Law. *A Treatise of Legal Philosophy and General Jurisprudence. Volume 6: A History of the Philosophy of Law from the Ancient Greeks to the Scholastics*. Miller Jr, F. D., & Biondi, C. A. (Eds.), Dordrecht: Springer. 251–265.

Boschetti, F., Del Grosso, A. M., & Spinazzè, L. (2021). La galassia Musisque Deoque: storia e prospettive. *ANTICHISTICA. ARCHEOLOGIA*, 32, 405–419.

Branco, A., Eskevich, M., Frontini, F., Hajič, J., Hinrichs, E., de Jong, F., Kamocki, P., König, A., Lindén, K., Navarretta, C., Piasecki, M., Piperidis, S., Pitkänen, O., Simov, K., Skadina, I., Trippel, T., Witt, A., & Zinn, C. (2023). The CLARIN infrastructure as an interoperable language technology platform for SSH and beyond. *Language Resources & Evaluation*, 1–32. DOI: 10.1007/s10579-023-09658-z

Cecchini, F. M., Korkiakangas, T. & Passarotti, M. (2020). A New Latin Treebank for Universal Dependencies: Charters between Ancient Latin and Romance Languages. *Proceedings of the 12th Language Resources and Evaluation Conference*, Marseille, France 11–16 May 2020. Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., & Piperidis, S. (Eds.). 933–942.

Cecchini, F. M., Sprugnoli, R., Moretti, G., & Passarotti, M. (2020). UDante: First Steps Towards the Universal Dependencies Treebank of Dante's Latin Works. *Proceedings of the Seventh Italian Conference on Computational Linguistics CLiC-it 2020,* Bologna, Italy, March 1–3, 2021. Dell'Orletta, F., Monti, J. & F. Tamburini (Eds.). 1–7.

Dingledy, F. W. (2016). The Corpus Juris Civilis: A Guide to its History and Use. *Legal Reference Services Quarterly,* 35(4), 231–255. DOI: 10.1080/0270319X.2016.1239484

Giglia, E. (2023). Open Science cafè - Come scrivere un Data Management Plan (DMP). *Zenodo*. DOI: 10.5281/zenodo.7526256

Haug, D. & Jøhndal, M. (2008). Creating a Parallel Treebank of the Old Indo-European Bible Translations. *Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data (LaTeCH)*, Prague, Czech Republic, June 28 2007. Sporleder, C. & Ribarov, K. (Eds.). 27–34.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. *Machine Translation Summit*. Phuket: Thailand. http://people.csail.mit.edu/koehn/publications/europarl/

Li, C., & Colangelo, L. (2016). *Luoma fa Minfa Daquan fanyi xilie: Xueshouo Huizuan di shiliu juan. Dixiao yu jituo = Iustiniani Augusti Digesta seu Pandectae, libro XVI. De Compensationibus et de deposito*. Beijing: Zhongguo Zhengfa Daxue Chubanshe.

Luo, G., & Colangelo, L. (2013). *Luoma fa Minfa Daquan fanyi xilie: Xueshouo Huizuan di' erishiasan juan. Hunyin yu jiazi = Iustiniani Augusti Digesta seu Pandectae, libro XXIII. De Matrimonio et dote*. Beijing: Zhongguo Zhengfa Daxue Chubanshe.

Marinai E., Peters C., & Picchi E. (1992). A Project for Bilingual Reference Corpora. *Acta Linguistica Hungarica,* 41, 1/4, 191–204. http://www.jstor.org/stable/44308294

Michener, W. K. (2015). Ten simple rules for creating a good data management plan. *PLoS computational biology*, 11(10), e1004525.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations.* Celikyilmaz, A. & Wen, T-H. (Eds.). 101–108.

Passarotti, M. (2019). The Project of the Index Thomisticus Treebank. *Digital Classical Philology*, 10, 299–320.

Ribary, M., & McGillivray, B. (2020). A Corpus Approach to Roman Law Based on Justinian's Digest. *Informatics*, 7(4), 44. https://doi.org/10.3390/informatics7040044

Roelli, P. (2014). The Corpus Corporum, a New Open Latin Text Repository and Tool. *Archivum Latinitatis Medii Aevi (ALMA): Bulletin Du Cange*, 72, 289–304.

Sassolini, E., Cucurullo, S., & Sassi, M. (2014). Methods of textual archive preservation. *Proceedings of CLiC-it 2014 – First Italian Conference on Computational Linguistics*, vol. I. Pisa, Italy, 9–10 December 2014. 334–338.

Schipani, S. (1994[1996]). Primo rapporto sulla attività della ricerca: "Il latino del diritto e la sua traduzione. Traduzione in italiano dei Digesta di Giustiniano". *Studia et documenta historiae et iuris, 60. Studi in Memoria di G. Lombardi.* Falchi, G. A. (Dir.). Roma: Pontificia Università Lateranense. 553–568.

Schipani, S. (2005). Premessa. *Iustiniani Augusti Digesta seu Pandectae. Digesti o Pandette dell'Imperatore Giustiniano. Testo e traduzione.* Volume I (libri 1–4). Schipani, S. (Ed.) & Lantella, L. Milano: Giuffrè Editore.

Schipani S. (Ed.), & Lantella, L. (2005a). *Iustiniani Augusti Digesta seu Pandectae. Digesti o Pandette dell'Imperatore Giustiniano. Testo e traduzione*. Volume I (libri 1–4). Milano: Giuffrè Editore.

Schipani S. (Ed.), & Lantella, L. (2005b). *Iustiniani Augusti Digesta seu Pandectae. Digesti o Pandette dell'Imperatore Giustiniano. Testo e traduzione*. Volume II (libri 5–11). Milano: Giuffrè Editore.

Schipani S. (Ed.), & Lantella, L. (2007). *Iustiniani Augusti Digesta seu Pandectae. Digesti o Pandette dell'Imperatore Giustiniano. Testo e traduzione.* Volume III (libri 12–19). Milano: Giuffrè Editore.

Schipani, S. (2010). 桑德羅. 斯奇巴尼教授文集 = Sandro Schipani, *scritti di diritto romano pubblicati in cinese.* 中國政法大學出版社. *Corporis Iuris Civilis. Digesta*. Schipani, S. (Ed.) (2001–2016). Beijing: CUPL. (books published in the collection: 1, 3–4, 6, 8–9, 12–13, 16–18, 22–24, 41, 48).

Schipani S. (Ed.), & Lantella, L. (2011). Iustiniani Augusti Digesta seu Pandectae. Digesti o Pandette dell'Imperatore Giustiniano. Testo e traduzione. Volume IV (libri 20–27). Milano: Giuffrè Editore.

Simard, M., & Plamondon P. (1998). Bilingual Sentence Alignment: Balancing Robustness and Accuracy. *Machine Translation,* 13, 59-80. DOI: https://doi.org/10.1023/A:1008010319408

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufiş, D. & Varga, D. (2006). The JRC-Acquis: A Multilingual Aligned Parallel Corpus with 20+ Languages. *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. Calzolari, N., Choukri, K., Gangemi, A., Maegaard, B., Mariani, J., Odijk, J. & Tapias, D. (Eds.).

Straka, M., Straková, J. & Gamba, F. (2024). ÚFAL LatinPipe at EvaLatin 2024: Morphosyntactic Analysis of Latin. *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024*. Torino: ELRA and ICCL. 207–214.

Yousef, T., & Berti, M. (2015). The Digital Fragmenta Historicorum Graecorum and the Ancient Greek-Latin Dynamic Lexicon. *Proceedings of the Workshop on Corpus-Based Research in the Humanities (CRH)*, Warsaw, Poland, 10 December 2015. Mambrini, F., Passarotti, M., Sporleder, C. Warszawa (Eds.). 117–123.

The Staccioli Digital Archive: Using Knowledge Graphs to power digital art history catalogues and art exhibitions

Klaus Werner[1], Pietro Liuzzo[1], Alessandro Adamou[1]
[1] Bibliotheca Hertziana – Max Planck Institute for Art History, Italy – *firstname*.*lastname*@biblhertz.it

ABSTRACT (ENGLISH)

This paper presents the Staccioli Digital Archive, a comprehensive catalogue of digitized material and semantic metadata revolving around the life and career of Italian sculptor Mauro Staccioli. Through the shared usage of open standards, conventions and systems, it attests to a fruitful interaction between art history research institutions, individual scholars, and the archives and foundations providing the source material. The Staccioli Archive also powers a persistent virtual exhibition built upon the same data platform as the archive, and supported a physical exhibition that ran in 2023-24. This collaborative development follows an extensible data management model based on FAIR standards such as Linked Data, CIDOC-CRM and IIIF, from inventory to publishing. The workflow relies upon shared conventions designed in such a way that data capture, digitization, IIIF image publishing, and the identity management of archival units could proceed each at its own pace and independently of one another. This model is intended to serve as a blueprint for the creation of future explorable art history datasets with support for discoverability, to potentially pair with virtual exhibitions to come.
Keywords: digital archives, Knowledge Graph; Linked Data; Art History; 20th-century sculpture

ABSTRACT (ITALIANO)

L'Archivio Digitale Staccioli: utilizzare i Knowledge Graph per alimentare cataloghi di storia dell'arte e mostre digitali. Il contributo presenta l'Archivio Digitale Staccioli, un catalogo completo di materiale digitalizzato e metadati semantici che ruotano intorno alla vita e alla carriera dello scultore italiano Mauro Staccioli. Attraverso l'uso condiviso di standard, convenzioni e sistemi aperti, esso è testimone di una proficua interazione tra istituti di ricerca di storia dell'arte, singoli studiosi, ed archivi e fondazioni da cui provengono le fonti materiali. L'Archivio Staccioli alimenta anche una mostra virtuale persistente costruita sulla stessa piattaforma di dati dell'archivio e ha supportato una mostra fisica svoltasi nel 2023-24. Questo sviluppo collaborativo segue un modello di gestione dei dati estensibile basato su standard FAIR come Linked Data, CIDOC-CRM e IIIF, dalla fase di inventario a quella di pubblicazione. Il flusso di lavoro si basa su convenzioni condivise, progettate in modo tale che l'acquisizione dei dati, la digitalizzazione, la pubblicazione di immagini IIIF e la gestione dell'identità delle unità archivistiche possano procedere ciascuna al proprio ritmo e indipendentemente l'una dall'altra. Questo modello progettuale è destinato a fungere da base per la creazione di futuri dataset di storia dell'arte con supporto per l'esplorabilità, da abbinare potenzialmente alle mostre virtuali a venire.
Parole chiave: archivi digitali, Knowledge Graph; Linked Data; storia dell'arte; scultura contemporanea

1. INTRODUCTION

In 2023, curator Marica Antonucci of the exhibition "Mauro Staccioli: Cementing an Artistic Legacy" hand-picked a selection of visual material documenting the defining moments in the career of Italian contemporary sculptor Mauro Staccioli (1937-2018) (Dorfles, Barbero, & Bazzini, 2008). Prior to the exhibition, which was open from October 2023 to January 2024, the archive was only available as a plaintext inventory with image galleries of partially digitized material, mostly limited to photographs of Staccioli's art installations and designs. The physical inventory, however, comprised a wealth of additional documentation, including preliminary drawings, press, correspondence, and exhibition flyers, catalogues and tickets, all of which nurture a greater understanding of Staccioli's artistic production context and were hitherto available in physical form only.

Aside from providing the curator with additional material to include in the digital counterpart to the exhibition, digitizing these resources also had the purpose of adding a dimension of serendipitous discoverability, thus allowing the viewer to branch out of the exhibition and explore other digital, if not curated ad-hoc, resources. A structured and standardized knowledge organization of the data for these documents is key to achieving this goal, for which Linked Data provide an ideal implementation. Additionally, setting up such a complex environment requires the contribution of departments and teams—the Library, Photo archive, digitization team, curators and scholars carrying out onsite fieldwork—that

operate on greatly differing schedules: while their respective outputs must be integrated seamlessly, the paces at which they produce it must be taken into account without disrupting their workflows.
For the above reasons, the outcome of this effort, the Staccioli Digital Archive,[1] rather than addressing specific research questions in art history, focuses on how to integrate the existing infrastructure, including assets which are usually isolated, such as an archive, a database and an exhibition. The front-facing component of the project is implemented on a fork of the ResearchSpace platform, which sits atop the knowledge graph of the Bibliotheca Hertziana, and on the facilities of its Photo archive for publishing its visual material using the IIIF framework, while the ePIC identifier system provides IDs to archival units. These components operate independently, yet are seamlessly integrated through a shared knowledge of how the material is organized, which is encoded in the knowledge graph using the CIDOC-CRM core model. The proposed set-up has the added benefit of leaving space for further possible layers of reading, narrative and interpretation to be added explicitly or simply formed by users in their experience with the application.

2. BACKGROUND

The physical archive of Mauro Staccioli's oeuvre is located at his landhouse in Mazzolla, Tuscany. The material is stored in 190 boxes sorted per year and/or project; 40 moleskin notebooks containing mostly theoretical, partly already published writings; 10 horizontal folders with plans and sketches; 30 rolled-up plans; a cabinet full with slides used for seminars at the University of Milan; and 15 maquettes of his projects. A project may be intended as the preparatory work for, or documentation of, an art exhibition or installation, or the design of an individual sculpture. The criterion around which they are organized is primarily event-driven and chronological.
A digital inventory of the physical archive contents does exist, however, its entries are limited to listing the amount of units per type contained in each folder[2] in one single HTML page in Italian, despite the website being bilingual. Since a full digitization of the complete archive would have taken two to three years, our digitization campaign hand-picked a selection of the most relevant projects, detailed in Section 4.

3. FIELDWORK AND DATA CAPTURE

To accommodate the schedule of the art historian tasked with scrutinising Staccioli's physical archive in Mazzolla and recording its data, the timescales of a native Linked Data implementation of the data capture workflow were not feasible: a field database was then developed on top of a commercial relational DBMS, MementoDB, which is also available for tabletop systems and could be installed on the scholar's tablet device for the on-site recording of data from the archive's folders. Although this part of the implementation was relational, it was design reusing terms from the CIDOC-CRM ontology to facilitate subsequent transformations. Whereas the database was supposed to only give us a crude overview of the material—box, folder, and item with its relevant project title and timeframe—it quickly evolved to accommodate a more fine-grained representation and, most importantly, linking to external datasets and authority files. The final schema for data capture was thus as follows:

- **Document**(ID, box, folder, item, title, date, conservation, note, Keyword, P2_has_type[Type], P62_depicts[Work], P62_depicts[Person], P67_refers_to[Person], P45_consists_of[Material], P108i_was_produced_by[Person], P70_documents[Event], P67_refers_to_Institution[Institution])
- **Work**(ID, title, date, fate, Wikidata, description, P53_has_former_or_current_location[Place], P94i_was_created_by[Person], P103_was_intended_for[Event], P45_consists_of[Material])
- **Material**(ID, name, AAT)
- **Institution**(ID, name, GND, P69_has_association_with[Event])
- **Person**(ID, name, ULAN, Wikidata)
- **Place**(ID, name, geometry, Wikidata)
- **Event**(ID, title, date, P7_took_place_at[Place], P147i_was_curated_by[Person])
- **Keyword**(ID, name, GND)

Simplifications with regards to CRM include the fact that P94i_was_created_by or P108i_was_produced_by were linked directly to persons, rather than to activities as mandated by CRM, as the original archive does not distinguish individual production activities, therefore the reification of the activities required by the CRM could be deferred to a later stage, where the RDF data for the knowledge graph were produced. Links to the German authority file GND, to the Getty thesauri AAT and ULAN, and to Wikidata, as per the eponymous fields in the schema, were added manually after each fieldwork session.

1 Staccioli Digital Archive: https://staccioli.biblhertz.it/ (cons. 24/01/2025).

2 Original Staccioli inventory: https://maurostaccioli.org/it/inventario-dellarchivio-mauro-staccioli/ (cons. 22/01/2025).

4. DIGITIZATION AND DIGITAL IMAGE STORAGE

As the material was progressively described, the curator selected a subset of it to prioritize for digitization targetting the planned exhibition. Of more than 150 events (exhibitions, showcases etc.) to which Staccioli took part between 1965 and 1988, which had been documented in the database, three were selected, owing to the fact that they were well-documented and covered an important phase in the development of the artist. These are: Staccioli's solo exhibition in Volterra, 1972; the 1978 Biennale in Venice, and "Arte e Critica" at the National Modern Art Gallery in Rome, 1981 (GNAM'81).

The boxes were transferred to the Bibliotheca Hertziana, where the digitization of the papers, sketches, photos and slides was performed. The storage schema closely resembled the physical order, as indicated in the "Document" entity of the database: boxes, folders and items as nested (sub)directories containing scans with 4-figure numbers. The compiled material amounted to over 22,000 scans of ca. 7,500 items.

The digital images produced were stored on a Nexus-POSIX filesystem, which already hosted all the images of the Photo archive of the Bibliotheca Hertziana, and were similarly processed on a high-performance computing (HPC) cluster for transforming them to the JPG2000 format and longterm archival. The JPG2000 derivates are served using the Internet Image Interoperability Framework by a Cantaloupe image server backed by an RDF triple store. For each image a minimal set of triples was produced, which included the absolute path to the Nexus- POSIX storage and the accessibility information (property v:access) needed by the server to check whether it has the rights required for publishing it in open access.

5. SEMANTIC DATA MODELLING

Whereas setting up a relational data model served as a stratagem to maximize the throughput of data entry, achieving compliance to the FAIR principles (findability, accessibility, interoperability, reusability) requires that a Linked Open Data derivative of the data be produced and published.

The data collected revolved around a restricted set of entity classes and relations between them:

- Sculptures and other artworks authored by Staccioli himself or his collaborators.
- Events in which said artworks were featured or discussed.
- The documentation that constitutes the actual digitized material of our collection.
- Places may be event locations, or places where the artworks rest or are conserved.
- People related to artworks or documents as their authors, or as being depicted or mentioned.

It was ascertained that most of the above could be modelled using version 7 of the CIDOC CRM core ontology (Bekiari, Bruseker, Doerr, Ore, Stead, & Velios, 2021), with the Keyword entity in the original relational database being modelled as a SKOS thesaurus, and with minimal usage of other vocabularies like schema.org. External links are modelled as owl:sameAs.

```
event:QFJDK2YycilqZCNEanhhI01Sb3U/collection
  a crm:E78_Curated_Holding, crm:E72_Legal_Object,
    crm:E24_Physical_Human-Made_Thing ;
  rdfs:label "\"Alternativa - II Festival Internacional de Arte Viva\""@it;
  crm:P19i_was_made_for event:QFJDK2YycilqZCNEanhhI01Sb3U ;
  crm:P46i_forms_part_of <http://data.biblhertz.it/curated_holding/archivio_staccioli> .

event:QFJDK2YycilqZCNEanhhI01Sb3U a crm:E7_Activity ;
  crm:P7_took_place_at place:Rio5Tj47dCtYSm87UW1AQVlTajg ;
  crm:P19_was_intended_use_of event:QFJDK2YycilqZCNEanhhI01Sb3U/collection .

place:Rio5Tj47dCtYSm87UW1AQVlTajg a crm:E53_Place ;
  rdfs:label "Almada" ;
  owl:sameAs wd:Q328015 .
```

Listing 1. Example of RDF/Turtle rendering of a collection associated to an event

Listing 1 shows an example of how an event, the "Festival Internacional de Arte Viva" that took place in Almada, Portugal, was modelled in the knowledge graph as an E7_Activity in CRM, which is associated to a collection of materials that is part of the larger Staccioli archive.

The mapping from the initial database to RDF was performed using SPARQL-Anything, a software library for querying data in various formats using the SPARQL language with an extended syntax for query

federation and additional utility functions (Asprino, Daga, Gangemi, & Mulholland, 2022). Because MementoDB exposes its data using a cloud service with a Web API in JSON format, it was possible to query it directly in SPARQL-Anything, use queries of type CONSTRUCT to generate new RDF data in the desired schema based on CIDOC-CRM, and store the result in the knowledge graph infrastructure of the project. A dataset of nearly 250,000 RDF triples was produced for this project.

6. END-USER APPLICATION

Due to its flexible templating mechanism and versatile support for different semantic storage solution, ResearchSpace (Oldman & Tanase, 2018) was adopted as the semantic platform upon which to build the catalogue and digital exhibition. Unlike default ResearchSpace installations, the instance of the Staccioli archive does not have its own RDF store, but is instead fed from the knowledge graph of the Bibliotheca Hertziana, which hosts multiple datasets along with the Staccioli one: therefore, no further data integration was necessary after the re-engineering from the original MementoDB instance. This was also made possible by the fact that the editorial workflow was implemented outside the platform.

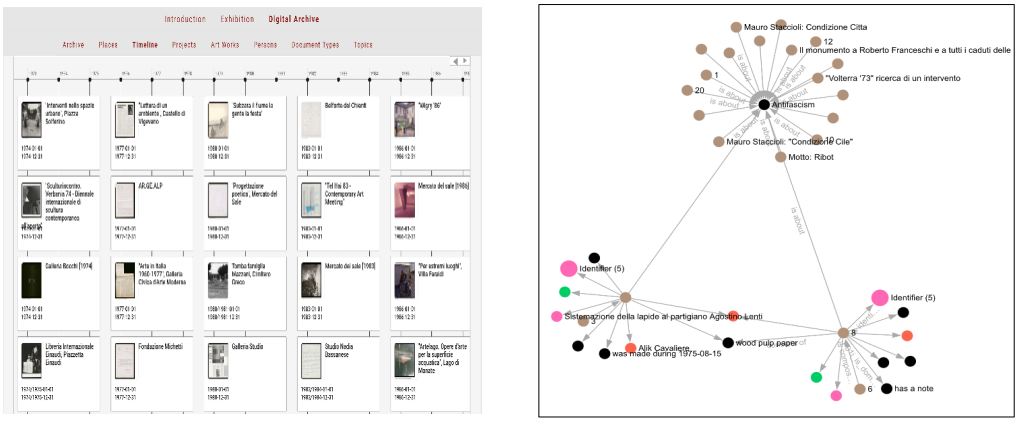


Figure 1. Exploring the Staccioli Digital Archive through a timeline view (left) and the SigmaGraph (right).

Visualizations were implemented using off-the-shelf components of ResearchSpace, such as the timeline view of events and projects, map-based geospatial exploration, and entity-based search. Additionally, the standard graph-based visualization, which displays a fixed set of notes retrieved from a simple SPARQL, was replaced with a third-party component, the Semantic SigmaGraph[3] (Figure 1), which allows dynamic query expansion upon clicking on individual nodes, thereby encouraging the serendipitous exploration of knowledge around the sculptor's career, beyond the curated content of the exhibition (Po, Bikakis, Desimoni, & Papastefanatos, 2020). The Staccioli Digital Archive thus blends in one place both the exhibition contents selected from the archive and the digitized part, in its entirety, of the archive itself. While this is an unusual choice, because a scholarly dataset is typically not something one would want to expose to the visitors of the exhibition, it allows users to choose the level of complexity at which they wish to explore the content (Foo, Theng, Go, & Na, 2009). There were also noticeable benefits in the development stage, as it did not prove necessary to implement one system for the researcher interested in the data, and one for the visitor interested in further documentation to the items on display at the exhibition. In this way the restricted pre-organized view into the material which the exhibition favoured was preserved, but also the possibility was given to enhance one's knowledge by allowing a wider, and at the same time more refined, view into the available documentation.

7. IMAGE COLLECTIONS, RESOURCE IDENTITY, AND SEAMLESS INTEGRATION

A dedicated parallel structure was also built on the basis of the existing material exposed by the image servers, so that collections and manifest documents conform to the IIIF Presentation API could be constructed from the knowledge graph, independently from the ResearchSpace application. This makes it possible for these collections to be usable in any IIIF viewer, including the one set up as a standard for the Photographic collection of the Bibliotheca Hertziana, which hosts the digital archive.
Collections and Manifests are generated from an XQuery module stored in eXist-db. The module is integrated into a declarative structure and queries the knowledge graph directly over HTTP. The SPARQL

3 F. Kräutli. Visualising Networks. The Sigma Graph Component: https://www.sari.uzh.ch/en/Insights/Visualising-Networks.-The-Sigma-Graph-Component.html (cons. 24/01/2025).

queries are tailored each Collection and Manifest. Starting from a main Collection, several subcollections are available, which query the knowledge graph on the fly to list available items. This bore at least two advantages to the various workflows at play in the realization of this project:

- The Staccioli catalogue and exhibition can use the IIIF image viewer of the Photographic collection, without any need to host its own IIIF substructure, yet with interlinking between the two.
- Both the collections on the IIIF platforms, and the image links in the Staccioli catalogue, could be generated altogether, without having to wait for the digitization department to finish its job. Any empty collections or broken image links would automatically be fixed as the corresponding scan was performed and the resulting images processed and saved to longterm storage.

For each archival unit, a persistent identifier (PID) was also generated and stored into the knowledge graph, which replicates the box-folder-unit organization of the archive, upon querying the knowledge graph itself. As the identity provider, the ePIC mechanism[4] was selected due to its flexibility and open source API. For instance, PID http://hdl.handle.net/21.11153/biblhertz_staccioli_box_1_folder_6_unit_19 resolves to the 19th unit catalogued in folder 6 of box 1.

8. CONCLUSION AND OUTLOOK

Not only does the Staccioli Digital Archive offer an instance of co-habiting digital catalogue and exhibition, but also its implementation tested the ability of different departments with different workflows to interact with one another, while preserving their priorities and setups. By relying upon open standards, shared and well-understood conventions, and open source software, this endeavour repurposed existing technologies that are being adopted in other projects, with minimal adaptation. It also proved to be an effective way forward for several other similar projects, and set a precedent for a more integrated implementation of said existing resources in future efforts.

Funding is actively being sought to bring the digitization of the entire archive in Mazzolla to completion. This will be accompanied by further enhancements of the front-facing application, potentially expanding it to address the research questions of scholars in contemporary art history, naming providing alternative visualizations and prototypical queries to the knowledge graph. Due to the evolution of other ResearchSpace-based projects in the meantime, this second iteration aims at implementing the data curation workflow natively as Linked Data, thereby abating the need to perform an extraction process from a relational database.

ACKNOWLEDGEMENTS

The project was developed in collaboration with Associazione Archivio Mauro Staccioli, Fondazione Cassa di Risparmio Volterra, Galleria Il Ponte, Firenze, the Galleria d'Arte Niccoli, Parma, and the Swiss Art Research Infrastructure, Zurich. For detailed acknowledgements visit https://staccioli.biblhertz.it.

REFERENCES

Asprino, L., Daga, E., Gangemi, A. & Mulholland, P. (2022). Knowledge Graph Construction with a façade: a unified method to access heterogeneous data sources on the Web. *ACM Transactions on Internet Technology*. Association for Computing Machinery. https://doi.org/10.1145/3555312.

Bekiari, C., Bruseker, G., Doerr, M., Ore, C.-E., Stead, S., & Velios, A. (Eds.). (2021, April). Volume A: Definition of the CIDOC Conceptual Reference Model, Version 7.1.1. CIDOC CRM Special Interest Group. https://doi.org/10.26225/FDZH-X261.

Dorfles, G., Barbero, L.M., Bazzini, M. (Eds.). 2008. *Mauro Staccioli. All'origine del fare*. Mantova: Corraini Edizioni. ISBN: 978-8875701901.

Foo, S., Theng, Y., Goh, D. H., & Na, J. (2009). From Digital Archives to Virtual Exhibitions. *Handbook of Research on Digital Libraries: Design, Development, and Impact*. Y. Theng, S. Foo, D. Goh, & J. Na (Eds.), IGI Global Scientific Publishing. 88-100. https://doi.org/10.4018/978-1-59904-879-6.ch009.

Oldman, D. & Tanase, D. (2018). Reshaping the Knowledge Graph by Connecting Researchers, Data and Practices in ResearchSpace. *ISWC 2018 - 17th International Semantic Web Conference*, Monterey, CA, USA 8-12 November 2018. Vrandecic, D., Bontcheva, K., Suàrez-Figueroa, M.C., Presutti, V., Celino, I., Sabou, M., Kaffee, L.A. & Simperl, E. (Eds.). 325-340. https://doi.org/10.1007/978-3-030-00668-6_20.

Po, L., Bikakis, N., Desimoni, F., & Papastefanos, G. 2020. *Linked Data Visualization: Techniques, Tools, and Big Data*. Springer Cham. https://doi.org/10.1007/978-3-031-79490-2.

4 European Persistent Identifier Consortium: https://www.pidconsortium.net/ (cons. 24/01/2025).

Fantàsime: *Interactive Drama* per la valorizzazione del Patrimonio Culturale

Maria Chiara Provenzano[1], Eleonora Miccoli[2], Mario A. Bochicchio[2]
[1] Università Telematica Pegaso, Italia - mariachiara.provenzano@unipegaso.it
[2] Università degli Studi di Bari Aldo Moro, Italia - {eleonora.miccoli, mario.bochicchio}@uniba.it

ABSTRACT
Collocandosi al crocevia tra ingegneria informatica, storia e drammaturgia, l'obiettivo realizzativo *Itinerari turistici in Puglia* del PON "RASTA - Realtà Aumentata e Storytelling Automatizzato per la Promozione del Patrimonio Culturale e degli Itinerari Turistici" ha come obiettivo quello di rendere accessibile il patrimonio culturale delle residenze fortificate del Salento ad un pubblico eterogeneo attraverso una proposta teatrale interattiva e immersiva, grazie all'integrazione di strumentazione XR ed AI.
In questo contesto viene presentato l'approccio metodologico e le tecnologie adottate per combinare ricerca storica, scrittura drammaturgica, storytelling automatizzato basato su AI e creazione di ambienti virtuali, al fine di promuovere la valorizzazione del patrimonio culturale individuato come caso di studio, nella fattispecie il Castello Carlo V di Lecce, la cittadella fortificata di Acaya e il castello De' Monti di Corigliano d'Otranto, in Salento.
Parole chiave: digital storytelling; interactive storytelling; interactive drama; digital performance; phigital theatre.

ABSTRACT
Interactive Drama for the Valorisation of Cultural Heritage. At the crossroads of computer engineering, history and dramaturgy, the objective "Tourist Itineraries in Apulia" of the Project 'RASTA - Augmented Reality and Automated Storytelling for the Promotion of Cultural Heritage and Tourist Itineraries' aims to make the cultural heritage of the fortified residences of the Salento more accessible to a heterogeneous public through an interactive and immersive theatrical proposal, thanks to the integration of XR (eXtended Reality) and AI (Artificial Intelligence) technologies.
This paper presents the methodological approach and the technologies adopted to combine historical research, dramaturgical writing, AI-based automated storytelling and the creation of virtual environments, in order to promote the valorisation of the cultural heritage identified as a case study, in this case the Carlo V Castle in Lecce, the fortified citadel of Acaya and the De' Monti Castle in Corigliano d'Otranto, Salento, Apulia, Italy.
Keywords: digital storytelling; interactive storytelling; interactive drama; digital performance; phigital theatre.

1. INTRODUZIONE

Le performance teatrali costituiscono una forma d'arte che fin dalle sue origini ha fatto uso di tecniche ed effetti scenici in grado di emozionare e coinvolgere il pubblico. L'intelligenza artificiale (AI) e le tecniche di extended reality (XR) da questo punto di vista, con le loro peculiarità, contribuiscono ad arricchire la collezione di strumenti che registi, scenografi e artisti performativi possono utilizzare per raggiungere il loro obiettivo.
Il lavoro descritto in questo articolo riguarda un progetto di ricerca finanziato dal Ministero dell'Università e della Ricerca nell'ambito del "PON 12 aree", ovvero il Programma Nazionale di Ricerca Industriale e Sviluppo Sperimentale nelle 12 Aree di Specializzazione Smart. Il progetto è denominato "RASTA - Realtà Aumentata e Storytelling Automatizzato per la Promozione del Patrimonio Culturale e degli Itinerari Turistici" e l'area di specializzazione smart è "Patrimonio Culturale".
L'obiettivo di RASTA è definire un framework tecnologico per la promozione del patrimonio culturale e del turismo in Italia, anche attraverso la realtà estesa, le tecnologie di storytelling intelligente per la personalizzazione della narrazione e sofisticati sistemi di raccomandazione e profilazione degli utenti.
In questo contesto, l'articolo descrive il lavoro che l'unità di ricerca dell'Università del Salento sta portando avanti riguardo al tema "Itinerari turistici in Puglia". Il focus della nostra unità di ricerca è la produzione, in realtà estesa, di tre performance teatrali allestite in tre castelli medievali pugliesi riguardo alla storia medievale di questi luoghi. Qui descriviamo la struttura del sistema che abbiamo progettato e che stiamo realizzando per produrre gli eventi teatrali.

L'obiettivo è quello di creare un sistema in grado di supportare il regista e gli attori teatrali nella creazione e messa in scena di performance dal vivo che sfruttino tecnologie come extended reality, intelligenza artificiale e storytelling intelligente per la personalizzazione della narrazione e della *user experience*. Nel seguito è descritto l'approccio drammaturgico adottato per l'ideazione e la messa in scena degli spettacoli assieme all'architettura del sistema informatico che si sta realizzando, considerando come premessa di fondo quanto le arti della scena siano, per natura e statuto, da sempre vocate all'innovazione e, inevitabilmente, multimediali e il fatto che dagli anni '70 del Novecento, con l'avvento dei new-media, i linguaggi artistici si siano innovati accogliendo il digitale.
Il teatro, come la filosofia, è una *pratica* che interroga l'esistenza ponendo al centro l'uomo, producendo immagini – *fantàsime, allucinazioni,* per dirla con Pirandello - che interrogano il reale. Lo statuto *finzionale* del teatro non dipende, però, dalla "virtualità" dell'accadimento ma dal fingere di crederci, grazie alla lente che lo spettatore indossa per "estendere" la propria realtà. Infatti lo spettatore (*theoròs*) riporta una testimonianza, una teoria derivata dalla visione/contemplazione della quale i suoi occhi e le sue orecchie si sono intrisi, riproducendo e fissando nella propria mente immagini sonore.

2. SISTEMA DRAMMATURGICO E STORYTELLING AUTOMATIZZATO

Una folta schiera di definizioni e neologismi nell'ambito della *performance digitale* (*Cybernarratore, Synthetic Actor, Digital Storyteller, Hyperactor, Augmented Reality Theatre, Cyber Performance, Digital Performance, Computer Theatre, ecc.*) fa da eco alla riformulazione del sistema drammaturgico di matrice aristotelica, che diventa **iperdramma** – ovvero scrittura ipertestuale che utilizza tecnologie audiovisive, digitali e interattive (Balzola, 2005) – o **sistematurgia**, cioè: "una drammaturgia che ha bisogno dell'informatica, basata sul principio della gestione della complessità del computer. La sistematurgia è un processo interattivo che indaga attraverso nuovi prototipi, un arco di mediazione che include l'interfaccia, il calcolo e i nuovi mezzi di rappresentazione; sta al servizio di una narrazione, di un racconto, di un organismo teatrale ma lo fa in maniera interattiva usando uno strumento ipermediale" (Antunez Roca, in Monteverdi 2011). Quali sono le nuove forme espressive di scrittura scenica, ovvero – per prendere spunto dall'interrogativo di Lev Manovich: "**Come possono le nostre nuove capacità** – archiviare masse di dati, classificarli, indicizzarli, collegarli, ricercarli automaticamente e recuperarli istantaneamente – **realizzare nuove tipologie di narrazione?"** (Manovich, 2001, 294).
Come sintetizza Antonio Pizzo "La sperimentazione tecnologica degli anni '80 ha fatto emergere una pratica di scrittura scenica multimediale e tecnologica e una diversa attenzione alla realizzazione artistica in una riconfigurazione dei codici scenici e del linguaggio drammatico", (Pizzo, 2013, 13).
Andrea Balzola sostiene che il concetto di drammaturgia ha assunto molteplici significati nel Novecento, che ne hanno fatto "una definizione aperta e piuttosto ambigua" un campo semantico che contiene varie concezioni di teatro e ne riflette l'evoluzione, annullando il binomio testo/rappresentazione e, per quel che concerne la scena multimediale, evidenzia il passaggio da testo lineare a ipertesto polimorfo, e da teatro spettacolo a spettacolo laboratorio, in cui il drammaturgo si trova a **generare eventi performativi polisensoriali in tempo reale** (Balzola, 2009).
L'integrazione delle tecnologie digitali nella creazione scenica è ormai una pratica consolidata. Ambienti immersivi, personaggi virtuali, nuove modalità di interazione tra attori, oggetti di scena e pubblico, e l'uso esteso degli effetti acustici stanno trasformando profondamente il linguaggio teatrale. Numerose sperimentazioni ne sono testimonianza.
In (Vasilakos, Athanasios V., et al., 2008), ad esempio, è descritta una forma di teatro interattivo basata su tecnologie di realtà mista e intelligenza ambientale, progettato per fornire ai *performer* strumenti creativi che estendano la grammatica del teatro tradizionale. Attori e danzatori, anche distanti fisicamente, sono ripresi da più telecamere e renderizzati in 3D, consentendo loro di esibirsi e danzare in tempo reale nello stesso luogo virtuale.
In (Bernini, et al., 2011) è presentato il concetto di coreografia aumentata, in cui gli attori virtuali si esibiscono al fianco di performer reali, influenzando la percezione scenica e lo spazio stesso.
In (Marner et al., 2012) è illustrato uno spettacolo teatrale in realtà aumentata in cui il pubblico interagisce attivamente, con gli attori reali e virtuali, votando su come far avanzare l'azione scenica. Il set è rappresentato come un ambiente virtuale 3D. Gli attori sono seguiti da sensori e telecamere mentre si muovono sul palco e i contenuti proiettati sono controllati dai loro movimenti. I risultati delle votazioni del pubblico vengono proiettati direttamente nell'ambiente virtuale.

In (Lisowski et al., 2023) è presentata un'ampia e aggiornata discussione sul rapporto tra tecnologie digitali e teatro, in cui la performance teatrale dal vivo è intesa come una forma d'arte in evoluzione e la realtà aumentata (AR) offre nuovi modi per connettere e coinvolgere il pubblico ed espandere il potenziale del palcoscenico. Inoltre, l'aggiunta del tracking 3D e delle proiezioni interattive consente nuove opportunità per le performance teatrali e la narrazione.
Infine, in (Papathanasiou et al., 2022) è discusso l'uso combinato di intelligenza artificiale, narrazioni basate sulla localizzazione e partecipazione degli utenti come mezzi per creare esperienze di realtà estesa efficaci e di impatto, aumentando l'empatia e il coinvolgimento.
La **performance multimediale assume caratteristiche proprie**, su piani differenti: **narratologico** (lineare vs. ipertestuale), **estetico** (spettacolo conchiuso vs laboratorio aperto) e **operativo** (da spazio scenico della prosa a spazio sensibile e interattivo della performance digitale). Su questi piani si è inteso operare nella scrittura di una drammaturgia originale atta a narrare le fortificazioni elette come soggetti e siti per la realizzazione del PON RASTA, nella fattispecie ci soffermeremo qui sullo *study case* del castello Carlo V di Lecce.

3. "CIBERNETICA E FANTASMI". INTERACTIVE HISTORY-TELLING NEL CASTELLO

Nella prolusione per la conferenza dell'Associazione Culturale Italiana del 1967, Italo Calvino esplora l'orizzonte della narrazione come processo combinatorio, riflettendo sulle potenzialità delle macchine intelligenti: «L'uomo sta cominciando a capire come si smonta e come si rimonta la più complicata e la più imprevedibile delle macchine: il linguaggio. […] Stabiliti questi procedimenti, affidato a un computer il compito di compiere queste operazioni, avremo la macchina capace di sostituire il poeta e lo scrittore? Così come abbiamo già macchine che leggono, macchine che eseguono un'analisi linguistica dei testi letterari, macchine che traducono, macchine che riassumono, così avremo macchine capaci di ideare e comporre poesie e romanzi?» (Calvino, 1980, 169-70). Quello che per lo scrittore era ancora uno scenario futuribile è diventato oggi reale grazie alla diffusione delle Intelligenze Artificiali generative, con tutti gli interrogativi di natura deontologica che ne conseguono. Anche il settore dello spettacolo contemporaneo è abitato da inquietudini e preoccupazioni dovute all'introduzione dell'AI nella creazione artistica e il teatro, nello specifico, si interroga da tempo sulla potenzialità / pericolosità della modalità creativa automatizzata.
Il progetto scientifico RASTA, nel suo obiettivo realizzativo dedicato alla valorizzazione dei castelli presenti nella provincia di Lecce, si pone in questo orizzonte di indagine, prevedendo come *output* la realizzazione delle tre performance teatrali in *digital liveness*, nei tre siti individuati da progetto: il castello Carlo V di Lecce, il Castello de' Monti di Corigliano d'Otranto e il Castello d'Acaya.
All'uopo, sono riunite competenze storico-artistiche, drammaturgiche e informatiche per la creazione di uno storytelling digitale, con l'utilizzo di un *Large Language Model* addestrato su un ampio *corpus* di fonti storiografiche e iconografiche inerenti le architetture fortificate del Salento, e declinate secondo lo stile della scrittura drammatica del XVI sec., con un lessico prossimo a quello dei letterati italiani che hanno fondato la Commedia Nuova in volgare (es. Machiavelli, Ariosto).
Nello specifico il lavoro di ricerca si è avviato a partire dal reperimento dei materiali bibliografici per la creazione di un *repository* di progetto e sulla creazione di una banca dati contenente studi archeologici, storici, linguisti, artistici, immagini e testi teatrali che fungessero da modelli per la creazione dello storytelling automatizzato, da trasformare ulteriormente in modello drammaturgico interattivo grazie alla collaborazione con la compagnia teatrale Factory Transadriatica (Lecce).
Nel caso in oggetto, difatti, potremmo parlare piuttosto di *Interactive History-telling* (Colombo, 2020), essendo la drammaturgia spiccatamente interattiva – per il maggior coinvolgimento del pubblico – ma basata su una ricerca scientifica preventivamente svolta, selezionando le fonti secondo criteri di **autenticità e accuratezza**; **interdisciplinarità**; **rilevanza culturale e tematica**.
Il processo di scrittura di scena – ovvero la dinamica performativa che riunisce testualità, visione e movimento – è qui caratterizzato da una redazione dei testi in modalità *phigital* (umana e automatica) – attualmente *in fieri* – integrata dalla realizzazione degli ambienti sonori e visivi immersivi che riproducono, con tecnologie VR e AR, archelogicamente gli ambienti storici dei siti di interesse, al fine di offrire al pubblico esperienze interattive, immersive e accessibili a diverse tipologie di utenza (es. bambini, turisti, non vedenti).

4. Deus ex Machina

Tra gli obiettivi della parte tecnica del Progetto RASTA, l'Università del Salento si occupa dell'ideazione e realizzazione di un sistema in grado di supportare il regista e gli attori teatrali nella messa in scena di performance dal vivo in realtà estesa, capaci di coinvolgere il pubblico.
Ciò ha condotto alla scelta di tecnologie "non invasive" (proiettori laser portatili e di ridotte dimensioni), compatibile con la valorizzazione dei luoghi scelti per l'allestimento degli spettacoli (castelli medioevali) senza comprometterne l'integrità, e capaci di evocare scene virtuali immersive tali da porre lo spettatore in uno stato di immersione attiva nel luogo di realizzazione della performance teatrale. L'uso combinato di tecniche di narrazione innovative, azioni performative e gaming ha consentito di coinvolgere ulteriormente lo spettatore interessato a conoscere la storia dei luoghi.
Il lavoro di ideazione del sistema digitale è partito da un'attenta analisi delle modalità adottate dal regista e dallo scenografo per creare, portare in scena e mettere a punto lo spettacolo teatrale. Anche la fase di esecuzione della performance é stata accuratamente analizzata per valutare possibili contributi performativi di tipo digitale. Il risultato delle analisi ha condotto il gruppo di ricerca alla progettazione e realizzazione di due sistemi digitali denominati "Machina" e "Deus".
Il componente denominato *Machina*, viene utilizzato dal regista e dallo scenografo per ideare e comporre il copione di performance teatrali con elementi AI e XR. Il componente denominato *Deus*, invece, coadiuva il regista e gli attori durante la fase di esecuzione della performance a partire dal copione redatto con *Machina*. Il copione può essere elaborato a partire da un testo teatrale nuovo o preesistente, mediante opportune annotazioni testuali che descrivono i comportamenti, le scene e gli effetti desiderati. Questa scelta mutua lo stile adottato per dialogare con i Large Multimodal Models (LMM) che usano descrizioni testuali opportunamente formulate ed arricchite (prompt engineering, Retrieval Augmented Generation, etc.) per generare testi, immagini, animazioni, musiche, scene, modelli 3D, voci etc..
Nello specifico, le annotazioni aggiunte al copione dal regista sono trasformate in istruzioni per la configurazione delle scene, per la descrizione del comportamento di un attore virtuale e per definire un insieme di comandi utilizzati dal performer in una determinata scena per attivare delle azioni (es. cambio di scena, luci, musiche, effetti, etc.). I comandi possono essere vocali, gestuali o provenire da telecomandi nascosti, o possono provenire da azioni del pubblico, e la corrispondenza tra comando e azione può essere definita a livello locale, in una determinata scena, o globale, ovvero valide per ogni scena dell'azione performativa. Altri esempi di azioni possono essere: rumori di sottofondo, passaggio da una scena alla successiva, gesti, spostamenti, azioni e battute dell'attore virtuale in scena.
Ogni scena del copione viene associata, tramite annotazioni, ad una scena virtuale, assimilabile ad un livello di gioco di un videogame, e viene riprodotta tramite proiettori in forma di scenografia immersiva che "avvolge" sia il pubblico che gli attori (se proiettata in stanze del castello di dimensioni opportune e con pareti adatte al tipo di proiezione) nella quale può essere presente un attore virtuale che interagisce con l'attore reale sulla base del copione, ma è guidato in tempo reale da una AI generativa e perciò è in parte "libero" di adattare le proprie parole agli stimoli dell'attore reale.
Una volta ottenuta la versione annotata del copione, questo viene dato in input alla *Machina* che trasforma le annotazioni in strutture dati come quelle rappresentate in figura 1.
La struttura dati così ottenuta è direttamente interpretabile da *Deus* (il secondo componente software) che si occupa del popolamento delle scene, dell'animazione e della voce dell'attore virtuale, della raccolta dei comandi provenienti dall'attore reale e della loro trasformazione in elementi multimediali o azioni di scena come precedentemente descritte. In altre parole, *Deus* organizza e costruisce gli elementi digitali dello spettacolo in base al copione e secondo le direttive impartite da performer umano con la narrazione. In figura 2 è rappresentata una performer (parte destra dell'immagine) che interagisce con la scena e l'attore virtuale (parte sinistra) mediante i gesti della mano che vengono rilevati e interpretati in tempo reale dal sistema Deus in base a quanto scritto nel copione per quella determinata scena (nello specifico: il gesto della mano attiva un'azione dell'attore virtuale). In una scena possono essere presenti uno o più attori virtuali con cui il performer può interagire. Le informazioni presenti nella struttura dati estratta dal copione consentono di indicare agli attori virtuali di cosa parlare e come far evolvere la conversazione con l'attore reale. La conversazione tra attore reale e virtuale viene generata durante la performance, in tempo reale, utilizzando un componente software appositamente realizzato in linguaggio python che, ad ogni battuta dell'attore umano:

- ne decodifica il parlato (conversione *speech to text*) mediante il modello di *machine learning* (ML nel seguito) *Whisper Medium*, di OpenAI;

- inserisce il testo della battuta in un *prompt* costruito per generare risposte in linguaggio cinquecentesco, guidate dal copione della scena teatrale in corso di svolgimento;
- invia il *prompt* e la battuta al *Large Language Model* GPT V3.5-turbo-fine_tuned, che risponde con la battuta dell'attore virtuale. La battuta, in testo libero, è comunque vincolata dal *prompt* a restare coerente con il copione e il ruolo del personaggio virtuale;
- trasforma il testo della risposta in un file audio mediante il modello ML *text to speech* di *Eleven Labs* opportunamente configurato per garantire efficacia espressiva e toni adeguati alla scena;
- importa il file audio nel sistema di rendering *Unreal Engine* V5.3
- usa le capacità di sincronizzazione labiale dell'estensione "MetaHumanSDK" di *Unreal Engine* per conferire realismo ed efficacia al volto dell'attore virtuale.

Questo consente di ottenere un'efficace interazione in tempo reale tra attore umano ed attore virtuale, garantendo ampia flessibilità e possibilità di personalizzazione dei testi, del linguaggio, del timbro e del tono vocale, dei gesti, della figura, dell'abbigliamento dell'attore virtuale e del contesto ambientale in cui l'attore è collocato. Ciò consente la messa in scena di spettacoli teatrali filologicamente adattabili alle finalità di *history telling* dichiarate dal progetto RASTA, includendo al tempo stesso elementi di improvvisazione e/o libera interpretazione sia per l'attore umano che per quelli virtuali. Queste caratteristiche distinguono in modo sostanziale il metodo e il software di narrazione sviluppato per il progetto RASTA da analoghi sistemi di animazione di avatar attualmente esistenti.
A conclusione delle rappresentazioni previste per il completamento del progetto RASTA è prevista l'intervista ai direttori dei castelli in cui saranno ambientate le rappresentazioni e la somministrazione e di un questionario agli spettatori ed al personale della compagnia teatrale per rilevare l'efficacia narrativa, il livello complessivo di gradimento, l'accuratezza storica e filologica delle rappresentazioni teatrali e il livello di integrazione delle tecnologie nel contesto drammaturgico.

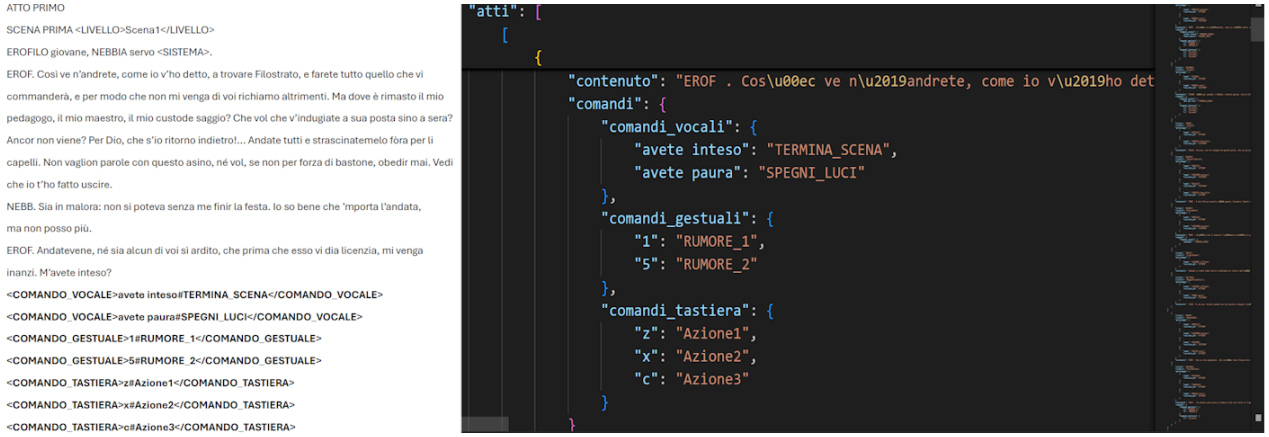


Figura 1. Copione con annotazioni e struttura dati corrispondente



Figura 2. Interazione tra attore umano e attore virtuale attivata tramite gesti

RINGRAZIAMENTI

Il progetto RASTA è finanziato dal Ministero dell'Università e della Ricerca con risorse del PON "Ricerca e Innovazione" 2014–2020 e con Fondi FSC ai sensi e per effetto dell'Art. 13 dell'Avviso per la presentazione di progetti di Ricerca Industriale e Sviluppo Sperimentale nelle 12 aree di specializzazione intelligente individuate dal PNR 2015–2020 e dell'Art. 1 del DD n. 551 del 27 aprile 2020.

BIBLIOGRAFIA

Arcagni S., Argano L. (2024). *onLive. Libro bianco sullo spettacolo digitale dal vivo in Italia*, Luiss University Press.

Balzola A. (2009). *Una drammaturgia multimediale. Testi teatrali e immagini per una nuova scena*, Editoria&Spettacolo.

Bernini, D., De Michelis, G., Plumari, M., Tisato, F., & Cremaschi, M. (2012). *Towards augmented choreography. In Arts and Technology: Second International Conference*, ArtsIT 2011, Esbjerg, Denmark, December 10-11, 2011, Revised Selected Papers 2 (pp. 10-17). Springer Berlin Heidelberg.

Boccia Artieri G. (2023), *OTONI - Oggetti Teatrali Online Non-Identificati: prospettive sulla liveness del teatro digitale,* Connessioni Remote, 6, 2023, 9-23.

Calvino I. (1980). *Cibernetica e fantasmi (Appunti sulla narrativa come processo combinatorio)*, in *Una pietra sopra. Discorsi di letteratura e società*, Einaudi.

Colombo P. (2020). *History Telling. Esperimenti di storia narrata*, Vita e Pensiero.

Gemini L., Brilli S. (2023). *Gradienti di liveness*, Franco Angeli.

Lisowski, D., Ponto, K., Fan, S., Probst, C., & Sprecher, B. (2023), *Augmented Reality into Live Theatrical Performance.* In Springer Handbook of Augmented Reality (pp. 433-450). Cham: Springer International Publishing.

Manovich L. (2002). *Il linguaggio dei nuovi media*, Olivares.

Marner, M. R., Haren, S., Gardiner, M., & Thomas, B. H. (2012, November). *Exploring interactivity and augmented reality in theater: A case study of Half Real.* In 2012 IEEE International Symposium on Mixed and Augmented Reality-Arts, Media, and Humanities (ISMAR-AMH) (pp. 81-86). IEEE.

Monteverdi A. M. (2011). *Nuovi media, nuovo teatro. Teorie e pratiche tra teatro e digitalità*, Franco Angeli.

Papathanasiou, A., Chavez, N. T., Bokovou, P., Chrysanthopoulou, C., Gerothanasi, A. D., & Dimitriadi, N. M. (2023), *Artificial Intelligence, Augmented Reality, Location-based narratives, and user participation as means for creating engaging Extended Reality experiences,* Proceedings of 5th International Conference Digital Culture & AudioVisual Challenges (DCAC-2023), 14p. 2022.

Pizzo A. (2013). *Neodrammatico digitale: scena multimediale e racconto interattivo*, Accademia University Press.

Pizzo A., Lombardo V., Damiano R. (2021). *Interactive storytelling. Teorie e pratiche del racconto dagli ipertesti all'Intelligenza Artificiale*, Dino Audino.

Vasilakos, Athanasios V., et al. (2008), *Interactive theatre via mixed reality and Ambient Intelligence.* Information Sciences 178.3, 2008: 679-693.

Soluzioni phygital e mediazione culturale: riflessioni digiteconomiche nell'era dell'IA*

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

ABSTRACT (ITALIANO)

L'interazione con la cultura e le sue manifestazioni e relative espressioni sta subendo una profonda trasformazione nell'era digitale, guidata dalla digitalizzazione e da approcci innovativi. In questo scenario, le problematiche legate alla lettura e all'interazione con libri e biblioteche emergono come temi di primo piano nella riflessione scientifica sull'impatto e le implicazioni della digitalizzazione e della digitizzazione massiva. Le analisi ISTAT evidenziano un progressivo declino delle pratiche di lettura, sia tradizionali che digitali, nonostante la preferenza per i libri stampati rimanga predominante, sottolineandone l'importanza simbolica e culturale. Le biblioteche, a lungo considerate pilastri della mediazione culturale, si trovano ad affrontare la sfida di adattarsi a questa nuova realtà che necessita del bilanciamento tra tradizione e innovazione. In questo contesto, tra le varie esperienze messe in campo negli ultimi anni, il modello phygital emerge come una soluzione promettente per integrare il valore sensoriale e simbolico dell'oggetto fisico con la flessibilità, la dinamicità e l'interattività dei formati digitali. Il paper esplora lo stato dell'arte di queste sfide e riflessioni nel contesto italiano. Mette in evidenza le tendenze emergenti e l'evoluzione socioculturale legata alla transizione verso nuovi modelli per coinvolgere le persone nel recupero della loro interazione con la conoscenza. Riflette sinteticamente sulle implicazioni etiche e socioculturali legate all'inclusione sociale delle comunità di utenti nella conoscenza della digitalizzazione, evidenziando la necessità di un approccio teoretico per definire e garantire il significato culturale alla dimensione digitale che oggi ci definisce e identifica.

Parole chiave: Digiteconomia; Ecosistema biblioteconomico; Mediazione culturale; Phygital; Inclusione sociale

ABSTRACT (ENGLISH)

Phygital Solutions and Cultural Mediation: Digithekonomic Reflections in the AI Era. The interaction with culture and its expressions is undergoing a deep transformation in the digital era, driven by digitalisation and innovative approaches. In this scenario, issues related to reading and interacting with books and libraries emerge as key topics in the scientific reflection on the impact and implications of digitalisation and massive digitisation. ISTAT analyses highlight a progressive decline in reading practices, both traditional and digital, despite the continued predominance of a preference for printed books, underscoring their symbolic and cultural importance. Libraries, long considered pillars of cultural mediation, face the challenge of adapting to this new reality, which requires balancing tradition and innovation. In this context, among the various initiatives implemented in recent years, the phygital model emerges as a promising solution to integrate the sensory and symbolic value of the physical object with the flexibility, dynamism, and interactivity of digital formats. This paper explores the state of the art of these challenges and reflections within the Italian context. It highlights emerging trend and the sociocultural evolution linked to the transition toward new models to re-engage people in their interaction with knowledge. It briefly reflects on the ethical and sociocultural implications of the social inclusion of user communities in understanding digitalisation, emphasising the need for a theoretical approach to define and safeguard the cultural significance of today's digital dimension that defines and identifies us.

Keywords: *Digitechnomos*; Libraries and information science; Cultural mediation; Phygital; Social inclusion.

* Nicola Barbuti ha curato i paragrafi 1, 4, 5; Mauro De Bari ha curato i paragrafi 2, 3, 5.

1. INTRODUZIONE

Per secoli, la comunicazione strutturata, variamente trasferita su supporti via via più stabili e raccolta e organizzata in ambienti resi disponibili all'accesso pubblico grazie all'elaborazione della dottrina biblioteconomica, ha rappresentato la sostanza culturale alla base dell'evoluzione e del progresso sociale, culturale ed economico degli individui e delle comunità antropiche (Mascia, 2023; Reading Agency, 2015). Tuttavia, nel nostro Paese tale presupposto ha iniziato a venire meno già a partire dalla seconda metà del secolo scorso, segnando un progressivo e irreversibile declino in atto anche in quello in corso.

Secondo una persistente opinione conservatrice popolare in ambito scientifico, una delle concause avocate come maggiormente responsabili di questo declino risiederebbe nella varietà di supporti che, a partire dalla fine del XIX secolo, avrebbe generato il concetto di multimedialità (Nardi, 2015). Un concetto che, a ben vedere, è invece nell'ordine di tutte le cosiddette "rivoluzioni" che hanno caratterizzato fin dalle ere storiche più antiche i media di trasmissione della comunicazione strutturata, segnando sempre l'innovazione quale elemento fondamentale di crescita sociale e culturale.

Ovviamente, secondo questa linea di pensiero, la rivoluzione computazionale e la trasformazione digitale in atto rappresentano ulteriori fattori intervenuti ad assestare il famoso colpo di grazia a un malato in agonia già da lungo tempo, accelerando un declino costante e dalle prospettive preoccupanti. Le fronde più fondamentaliste dei detrattori del digitale puntano l'indice accusatore sul progressivo cambiamento delle modalità di accesso e interazione degli utenti con gli strumenti di mediazione culturale, evidente nella preferenza oggi ormai quasi esclusiva per la comunicazione strutturata generata e veicolata nella dimensione digitale.

La situazione, già ipercomplessa prima della pandemia di Sars-CoV2 (Dominici 2019), si è ulteriormente complicata in seguito alle profonde destabilizzazioni socioculturali innescatesi durante le repentine restrizioni imposte per arginare il dilagare del virus (Dominici 2023). Infatti, l'apparente ri-scoperta durante le restrizioni delle mediazioni culturali disponibili in rete da parte di comunità di utenti fino a quel momento del tutto ignare avrebbe dovuto generare, in ricaduta, una crescita di consapevolezza delle opportunità di soddisfare bisogni percepiti come nuovi offerte dalla dimensione digitale.

A questo innesco repentino, il mondo delle biblioteche e la produzione editoriale hanno cercato di dare riscontro rispolverando e rilanciando tecnologie, tecniche e supporti digitali facilmente accessibili a diverse comunità di utenti, che in linea teorica avrebbero dovuto intercettare le istanze di rinnovamento e soddisfare i relativi bisogni.

Tuttavia, i dati sui numeri di utenti che nel nostro Paese, dopo la pandemia, annualmente accedono alla lettura di testi anche attraverso la rete o tramite supporti digitali, o agli OPAC delle biblioteche e alle digital libraries per la fruizione diretta online, mostrano impietosamente l'insuccesso dei risultati rispetto alle attese.

2. STATO DELL'ARTE: LO SCENARIO ITALIANO NEL POST-PANDEMIA

Nonostante il rilancio di valide iniziative intraprese già prima della pandemia per rendere pienamente accessibili le risorse digitali a tutte le fasce di utenti reali e potenziali (Comune di Bologna, 2022), inclusi quelli diversamente abili (Superabile, 2024), e gli annunci di nuove ambiziose imprese digitali, quale il Piano Nazionale di Digitalizzazione (PND) elaborato utilizzando anche i risultati di una consultazione pubblica (Digital Library Italia, 2022), dalla rilevazione ISTAT del 2022 solo il 39,3% della popolazione italiana di età superiore ai sei anni avrebbe letto almeno un libro per motivi non strettamente scolastici o professionali. Scende al 10,2% la popolazione che per leggere si è avvalsa di una biblioteca almeno una volta. Entrambe le rilevazioni mostrano un calo significativo rispetto anche a pochi anni prima, confermando il trend di progressiva riduzione dell'interesse per la lettura, soprattutto per quella fruita in biblioteca (ISTAT, 2022; CEPELL 2022).

I dati EUROSTAT rilevati lo scorso anno hanno confermato quelli del 2022, posizionando l'Italia al terzultimo posto in tutti i settori della comunicazione scritta, inclusi ebook, epub e media digitali vari. Non si è salvata neanche la lettura di svago, ultima roccaforte di resistenza fino a qualche anno fa (Agenzia DigiCult, 2024).

L'adozione di processi digitali ha offerto alle biblioteche l'opportunità di rilanciare il proprio ruolo di mediatore privilegiato tra cittadini e comunicazione strutturata. Tuttavia, gli sforzi compiuti per innovare più in profondità gestione e servizi, anche avvalendosi di alcune funzionalità di Intelligenza Artificiale (IA) senza però averne approfondito abbastanza le reali potenzialità, non paiono avere avuto il riscontro ipotizzabile (Barbuti & De Bari, 2022). L'ecosistema biblioteconomico è in ritardo rispetto alla velocità con cui sono cambiati gli approcci e i bisogni dei cittadini, e i metodi e i processi della digitalizzazione e della

digitizzazione. È in difficoltà nel concretizzare il ripensamento in corso su ruoli e funzioni che potrebbe esercitare nell'educare le comunità ad acquisire progressivamente consapevolezza sulle opportunità di crescita socioculturale, connesse con una transizione pur così rapida e pervasiva, qual è quella digitale. Gli istituti e i professionisti della biblioteconomia continuano a focalizzare riflessioni e dibattiti su quale sia la sorte destinata alle biblioteche nell'immediato futuro, anziché prendere coscienza della necessità inderogabile di convogliare le energie e le risorse intellettuali sull'elaborazione un sistema teoretico focalizzato sull'organizzazione, gestione, conservazione, trasmissione e valorizzazione dell'enorme sapere digitale sotteso alla produzione quotidiana di miriadi di dati digitali granulari.
Questa difficoltà dell'ecosistema biblioteconomico nel definire i requisiti teoretici su cui basare il processo di rimodellazione della propria identità storica, ormai irrinunciabile per affrontare consapevolmente le sfide peculiari all'ecosistema della digitalizzazione, rischia di precipitarlo in un collasso identitario che lo porterebbe a perdere definitivamente e irreversibilmente la propria vocazione più autentica, già da tempo in crisi: quella di farsi interprete privilegiato della mediazione tra le molteplici diverse comunità di utenti, sia potenziali che reali, e la conoscenza della cultura umana veicolata tramite comunicazione strutturata che l'ecosistema digitale mette già oggi a loro disposizione in molteplici formati.
La sfida è trovare un equilibrio tra questa secolare vocazione e il suo rinnovamento e riposizionamento nel contesto socioculturale e storico contemporaneo, focalizzando la riflessione sulla necessità di elaborare una nuova sistematizzazione teoretica. Pensiamo a una digiteconomia che, muovendo dalle conoscenze consolidate del sapere biblioteconomico, definisca i metodi e le strategie per generare, organizzare, gestire e disseminare al meglio un nuovo sapere focalizzato sulla conoscenza consapevole dell'ecosistema digitale, valorizzandone nel medio termine le istanze e le opportunità che esso ha, molte delle quali ancora oggi inespresse.
Allo scopo, la riflessione andrebbe articolata muovendo da un punto di osservazione prospettica, che abbia quale insostituibile e prioritario – ma non esclusivo – punto di fuga l'inclusione partecipativa e consapevole di tutte le comunità di utenti, potenziali e reali, nei processi di digitalizzazione e digitizzazione. Il fatto che, man mano che ci si avvicina, questo punto di fuga si allontani sfuggendo a ogni possibile velleità deterministica rappresenta il valore aggiunto di un approccio digiteconomico, che rimetta costantemente e diacronicamente in discussione i saperi transdisciplinari indispensabili a relazionarsi a contesti socioculturali via via sempre più ipercomplessi.

3. DISCUSSIONE: SFIDE E OPPORTUNITÀ

Nella rilevazione ISTAT del 2022 prima richiamata, il 69,8% dei non molti lettori italiani ha dichiarato di utilizzare solo libri cartacei (CEPELL, 2021) confermando la preferenza per l'esperienza sensoriale tangibile (ISTAT, 2022). A riguardo, riteniamo che a determinare questo dato concorrano istanze demo-etno-antropologiche non esclusivamente connesse con la lettura, tra le quali la dimensione simbolica dell'oggetto libro anche dal punto di vista strutturale, la sua accessibilità, la democraticità che evoca nell'immaginario della "conoscenza a portata di tutti". E, soprattutto, la confortevole certezza per l'utente che ogni lettura rappresenta un percorso ben definito da un numero di pagina iniziale sempre uguale, a un numero di pagina finale che, sebbene non si conosca a priori, si sa ci sarà sempre.
Tutti fattori entrati a far parte del nostro "dna" socioculturale in seguito a un processo evolutivo secolare lento e graduale. Infatti, l'"entità libro" – inteso nell'accezione di *biblíon*, di raccolta di parole organizzate in narrazione non casuale, veicolata tramite un artefatto stabile più o meno complesso (nostra definizione) – identifica una memoria collettiva rappresentativa del concetto stesso di "umano", in quanto manifestazione estetica (tangibile) e concettuale (non-tangibile) di un metodo di trasmissione della conoscenza, nel quale si sono riconosciute progressivamente quasi tutte le comunità antropiche succedutesi nel corso di diversi millenni. Dunque, come sottolineato da recenti studi (Esposito, 2021), la certezza dell'esistenza della memoria integrata nella fisicità dell'oggetto è un elemento chiave nella creazione del legame emotivo con il lettore: in una società sempre più digitalizzata (Gatti, & Colombo, 2023), per un'ampia fascia di popolazione il libro cartaceo rappresenta una sorta di "zona di conforto" nella quale attestarsi per resistere strenuamente e a oltranza alla smaterializzazione della cultura e alla granularizzazione informativa, che è prerogativa del digitale.
Nondimeno, nello stesso 2022 l'11,1% della popolazione ha letto solo e-book, o libri online, o ha utilizzato altri mediatori digitali di comunicazione testuale (ISTAT, 2022). Prioritariamente giovani, si intende, ma non solo. Riteniamo che la scelta sia determinata dalla non-tangibilità insita nella rappresentazione per immagini digitali dei contenuti testuali, in quanto è ormai un dato antropologicamente consolidato che le ultime generazioni preferiscono soddisfare i propri bisogni cognitivi accedendo a contenuti mediati per

immagini, possibilmente in movimento, piuttosto che leggendo testo (Brumberger, 2011). Inoltre, praticità di gestione dei dispositivi di lettura, facilità di accesso e di fruizione dei testi, costi via via sempre più contenuti degli hardware e delle edizioni elettroniche e digitali, democratizzazione dell'accesso alla conoscenza (Faggiolani, 2024) sono ulteriori plus conclamati soprattutto dai lettori più giovani.
D'altro canto, la materialità non-tangibile degli artefatti digitali risulta essere il principale fattore responsabile della percezione di "effimera" che ancora ne condiziona pesantemente una fruizione più estesa, in quanto privi di quella manifestazione tangibile che rende iconico e simbolico l'artefatto fisico, sebbene, come detto, anch'esso sia in progressiva dismissione. Va da sé che le differenze generazionali giocano un ruolo cruciale su questa percezione: i giovani nati durante la rivoluzione digitale, oggi abili a compulsare contenuti digitali ad alta risoluzione variamente (Mastrobattista, 2024), ma schematicamente interattivi, tendono a essere fisiologicamente più propensi alla lettura mediata dai supporti a loro familiari. Pur avendo altrettanta familiarità con i mediatori analogici, la maggior parte di loro li conosce e adopera solo perché costretta dai contesti scolastici.
Nonostante ampia parte della riflessione e il dibattito scientifico insista ancora oggi sulle apparentemente inconciliabili differenze, si sta progressivamente facendo strada una crescente convinzione che la dimensione fisica e quella digitale non siano opposti destinati nella migliore delle ipotesi a convivere senza riconoscersi, nella peggiore a prevalere una sull'altra determinandone l'estinzione. Anzi, alcune delle istanze più recenti identificano la chiave di volta proprio nel creare nuove esperienze che integrino la dimensione digitale nella fisicità dell'analogico (Meschini, 2024; Mastrobattista, 2024). La stessa antropologia suggerisce che il futuro delle relazioni con la comunicazione testuale non dipenderà esclusivamente dal formato tramite cui è veicolata, ma dall'abilità di integrare pratiche tradizionali e innovative (Ingold, 2007; Hine, 2000) in nuovi metodi e sistemi di interazione partecipativa e inclusiva.
Del resto, a ben vedere, oggi entrambe le dimensioni sono già parti integranti la contemporaneità che ci definisce, in quanto le istanze della digitalizzazione e della digitizzazione sono da tempo ben radicate negli interstizi della nostra relazione mentale con la quotidianità, che, di fatto, è dunque fisiologicamente già percepita quale phygital. Materializzare per scelta consapevole questa integrazione tramite processi di generazione di soluzioni interattivamente figitali (Giurickovic Dato, 2022) potrebbe rappresentare un passo importante per restituire alle entità testuali granulari, in qualsiasi manifestazione siano espresse, la vocazione più autentica di accessi a percorsi cognitivi, il cui "numero di pagina finale" sarebbe l'opzione di temporanea chiusura dell'esperienza che ciascun utente sceglie in base ai propri bisogni, non più vincolati alla consistenza predefinita e statica delle sole pagine cartacee (Barbuti e De Bari, 2022).
Recenti sperimentazioni di integrazione di ecosistemi biblioteconomici con la dimensione digitale – precisiamo che gli "ecosistemi biblioteconomici" di riferimento spaziano dal singolo libro a reti di biblioteche –, realizzate con il coinvolgimento inclusivo e partecipativo di utenti giovani, paiono confermare queste istanze, fornendo non pochi elementi da portare nella riflessione digiteconomica per trarne indicazioni utili a un'elaborazione teoretica in grado di affrontare sfide che si prospettano quanto mai impegnative (Mercanti, 2024).

4. L'EVOLUZIONE *PHYGITAL*

Il quadro sopra delineato dalla lettura dei dati ISTAT mostra un'ampia diversificazione delle esigenze e dei bisogni dei lettori, che spaziano dall'ostinato legame con l'oggetto fisico, alla necessità di accesso immediato ai contenuti e personalizzazione dell'esperienza di fruizione, fino alla ricerca di esperienze immersive capaci di replicare la relazione emotiva e simbolica con il libro nella sua rappresentazione classica, ma materializzata nella dimensione digitale. L'introduzione recente del modello phygital si propone quale potenziale riscontro a tale crescente richiesta.
A riguardo, recenti sperimentazioni di processi editoriali finalizzatisi nella stampa di volumi cartacei integrati con accessi a espansioni digitali che dilatano l'esperienza di interazione dei lettori e, nel contempo, li coinvolgono inclusivamente nell'arricchimento dei contenuti da associare ai testi, o di integrazione in volumi conservati in biblioteche di contenuti digitali che ne espandono l'interazione cognitiva (Barbuti e De Bari, 2024), attestano la possibilità di adottare lo stesso metodo utilizzando approcci differenti, ma sempre user-oriented in quanto basati sul coinvolgimento partecipativo e inclusivo degli utenti nell'esperienza di interazione con il testo mediato dagli artefatti figitali (Giurickovic Dato, 2022).
Tuttavia, queste iniziative, se per diversi aspetti lasciano intravedere prospettive di riattivazione dell'interesse delle comunità per la comunicazione testuale, devono essere valutate anche alla luce delle loro implicazioni più ampie. L'introduzione di approcci phygital, infatti, si traduce non solo in opportunità,

ma anche in sfide complesse e intricate che includono aspetti etici, socioculturali e tecnologici (Baker & Smith, 2020; Carr 2010). Le biblioteche e gli operatori culturali, infatti, sono chiamati a realizzare esperienze inclusive che, tenendo conto anche delle diversità culturali e sociali, evitino soluzioni standardizzate privilegianti determinati gruppi di fruitori e creino connessioni emotive e culturali tra tutti i possibili utenti e il testo in un contesto partecipativo (Hayles, 2012). Queste istanze hanno presieduto a una serie di sperimentazioni eseguite nell'ultimo quadriennio utilizzando la piattaforma *Expanded Endless Book* (*EXEBook*) nell'ambito di progetti editoriali e in biblioteche scolastiche (De Bari, Barbuti, & Azzella, 2022; Barbuti, 2024). L'ultima iniziativa in ordine di tempo, ancora in corso nell'ambito di un progetto TOCC finanziato da Invitalia, è una sperimentazione finalizzata ad applicare le funzionalità della piattaforma ai contenuti di una collana di edizioni anastatiche di opere a stampa antiche e storiche. Quale testo di apertura della collana è stata scelta la prima traduzione italiana de *Il Castello di Otranto* di Horace Walpole, stampata nel 1795 e riconosciuta come opera che inaugura il genere letterario gotico. Diversamente da quanto si potrebbe intendere dal titolo, è noto agli studiosi che Walpole non visitò mai l'Italia, tantomeno la Puglia. Tuttavia, la data di pubblicazione del romanzo originale e quella della sua prima traduzione italiana si collocano nel pieno della stagione dei *Grand Tour*, che vide numerosi viaggiatori stranieri, soprattutto inglesi, francesi e tedeschi, attraversare la penisola italiana percorrendola in tutta la sua estensione e scoprire la ricchezza di antiche vestigia che puntellavano soprattutto l'allora Regno di Napoli.
Alla luce di questa evidenza storica, si è formulata l'ipotesi che, nello scrivere il romanzo, Walpole possa essere stato in parte suggestionato dai racconti di qualcuno dei numerosi viaggiatori che avevano avuto modo di vedere di persona le *mirabilia* italiche. Infatti, è inevitabile pensare che, durante i lunghi viaggi affrontati, essi fossero entrati in contatto con le multiformi tradizioni culturali popolari che caratterizzavano le genti della nostra penisola, delle quali potrebbero avere restituito gli echi nei loro racconti una volta rientrati in patria.
Partendo da questa istanza, è stata eseguita la ricerca e l'analisi di fonti letterarie narrative del folclore dell'Italia Meridionale. Sono emerse connessioni tra l'opera di Walpole e la raccolta *Leggende di Puglia* di Saverio La Sorsa (1958), soprattutto con racconti a tema soprannaturale demoniaco e *monstra* fantastici. Queste connessioni sono state valorizzate marcando con le funzionalità di *EXEBook* parole, frasi e tavole dell'anastatica e collegandole a espansioni digitali riproducenti i relativi passi rilevati nell'edizione di La Sorsa, mettendo così a disposizione dei lettori la possibilità di esplorare in tempo reale le possibili relazioni tra culture apparentemente molto lontane tra loro all'epoca di edizione del romanzo di Walpole.
Nell'edizione, attualmente in stampa, le espansioni saranno accessibili e fruibili attraverso una app dedicata utilizzabile tramite dispositivi digitali, consentendo ai lettori di andare oltre le pagine del libro per immergersi in esplorazioni dinamicamente interattive.

5. SOLUZIONI PHYGITAL AL CROCEVIA DELL'*EVOLUZIONE IA*: PRIME SPERIMENTAZIONI

La corsa allo sviluppo di LLMs generativi e relative piattaforme di Intelligenza Artificiale (IA) iniziata alla fine del 2022, con i loro impianti teoretici e operativi in continua, rutilante evoluzione (Moriello, 2023), ha significativamente complicato uno scenario già denso di sfide irrisolte, non solo tecnologiche (Agenda Digitale, 2024). Nonostante già nel 2020 l'IFLA avesse pubblicato lo "Statement on libraries and artificial intelligence" (IFLA FAIFE, 2020), la comparsa di ChatGPT a novembre 2022 e la successiva, repentina proliferazione di sistemi di LLMs altrettanto performanti, quali Deep Seek e Gemini, hanno prodotto una vera e propria "sbornia IA".
Fortunatamente, la recente riflessione scientifica di settore ha posto le basi per avviare un confronto finalmente sistemico sulle implicazioni dell'applicazione dell'IA nell'ecosistema culturale, analizzandone il potenziale per organizzare, analizzare, gestire e rendere accessibili e interattive grandi collezioni di risorse sia fisiche, che digitali (Lamberti, Mellia, Vivarelli 2024; Meschini, 2024; Tammaro, Di Carlo e Zanichelli, 2024). Le prime valutazioni riconoscono ai LLMs efficacia nell'ottimizzare la catalogazione, nel recupero delle informazioni e nella personalizzazione dell'esperienza, a esempio, tramite "raccomandazioni" di lettura generate automaticamente (Lamberti, Mellia, Vivarelli 2024). Tuttavia, l'uso massivo, inconsapevole e acritico delle piattaforme di IA può generare una inutile quanto incontrollabile proliferazione di ciarpame digitale, contribuendo di fatto a spersonalizzare i processi di interazione e lettura (Floridi, 2020).
In questo scenario in rapidissima evoluzione, alcuni progetti stanno affrontando la sfida di sperimentare soluzioni phygital integrate con tool di IA, nell'ottica di migliorare ulteriormente l'interazione utente.

Tra i più recenti, si segnala il progetto "D_AEVUM 4.0 – The Phygital Ecosystem for the Enhancement of Unconventional Cultural Heritage"[1], finanziato dall'Università Ca' Foscari di Venezia nell'ambito del programma di attuazione INEST–Spoke 6 del Piano Nazionale di Ripresa e Resilienza (PNRR) a un'ATI composta dallo spin-off D.A.BI.MUS. S.r.l.[2] e dal DIRIUM di UNIBA. Il progetto mira alla rigenerazione partecipativa di beni culturali "non convenzionali", quali biblioteche storiche, archivi e musei minori. In questo contesto, l'IA non è utilizzata solo quale strumento per ottimizzare i processi di digitalizzazione e valorizzazione, ma fungerà da motore narrativo e adattivo capace di personalizzare l'esperienza culturale in tempo reale. Grazie all'integrazione in ambienti fisici di contenuti digitali interattivi, si stanno sperimentando dei veri e propri "portali esperienziali", che consentiranno agli utenti di accedere a interazioni aumentate che arricchiranno la relazione con i beni e gli ambienti di interesse. Il progetto, che sperimenta prototipi ad alto livello di maturità tecnologica (TRL7), potenzia la dimensione co-creativa già esplorata con *EXEBook*, offrendo esperienze modulabili, inclusive e sostenibili, capaci di attivare nuove forme di engagement tramite la combinazione tra IA, realtà mista e gamification.
Un ulteriore esempio di sperimentazione phygital è il progetto GAMEX Book (GaMing Approaches for Engaging Book eXperiences), finanziato a valere sul bando competitivo *ERC Seeds* dell'Università degli Studi di Bari Aldo Moro[3], con l'obiettivo di valorizzare i risultati di questa prima sperimentazione interna a UNIBA in un progetto da presentare su bando ERC europeo. Il progetto, attualmente in corso, si propone di implementare risorse phygital a stampa che integrino soluzioni di gaming digitale realizzate combinando funzionalità di IA, realtà aumentata (AR) e stampa tradizionale. Uno degli aspetti più sfidanti è la prospettiva di migliorare ulteriormente l'interazione dei lettori con le soluzioni phygital, arricchendole con edu-gaming che ne aumenterebbero sensibilmente il potenziale cognitivo e informativo. L'IA, in particolare, avrebbe un ruolo cruciale nel ridefinire la relazione tra lettori, contenuti e contesti, evolvendo dinamicamente il modello tradizionale di lettura lineare a favore di percorsi personalizzabili e *responsive*. L'approccio che si propone nel progetto favorisce il coinvolgimento partecipativo degli utenti attraverso un bilanciamento calibrato tra sfida e competenza, stimolando l'apprendimento esperienziale e favorendone l partecipazione attiva. La lettura diventa una dimensione aperta, in cui l'utente può interagire, esplorare e persino co-creare contenuti da integrare nella soluzione phygital in uso, attivando un processo di costruzione del significato che non è più esclusivamente personale, ma muove nella direzione di una condivisione collaborativa e collettiva. La personalizzazione resa possibile dall'IA, infatti, consente di adattare testi, linguaggi e percorsi narrativi in base a preferenze, livelli di comprensione e stili di apprendimento personalizzabili. Questa capacità di modellare l'esperienza in tempo reale dovrebbe rafforzare la comprensione e la memorizzazione del messaggio narrativo, rendendo la lettura accessibile, coinvolgente e modulabile in relazione alle specifiche esigenze di ciascun utente. I lettori potranno muoversi in una dimensione collaborativa costruendo, condividendo e scambiando nuovi saperi, dando vita a vere e proprie comunità di lettura collaborativa e creativa.
Tuttavia, alla luce di quanto finora delineato, emerge già oggi l'emergenza di organizzare, conservare, gestire e valorizzare culturalmente l'enorme massa critica di dati prodotta in questi scenari, siano essi generativi di soluzioni phygital, o espressione del progressivo ricorso all'IA anche per le più banali pratiche di interazione informativa da parte di comuni cittadini. Lo scenario in rapida e difficilmente controllabile evoluzione rilancia l'urgenza di avviare in concreto l'elaborazione teoretica una *digiteconomia*, che definisca metodi e regole in base alle quali classificare, organizzare, gestire, conservare, preservare e valorizzare la massa critica di dati digitali rinvenienti dai processi di digitalizzazione e di digitizzazione, garantendone nel contempo l'archiviazione e protezione secondo standard di sicurezza elevati, e fornendo riferimenti chiari sia a istituzioni e operatori culturali, che alle comunità di cittadini. Una digiteconomia che affronti anche gli interrogativi sulla capacità delle generazioni di nativi digitali di preservare la propria identità digitale in contesti futuri sempre più dominati dalle piattaforme globalizzate. Se le teorie di Clifford Geertz (1973) identificano la cultura in un insieme di significati condivisi che richiedono continuità per mantenere la loro rilevanza, la dimensione digitale, pur rappresentando di fatto la facies culturale contemporanea, rischia di granularizzare questi significati se non è progettata in modo inclusivo e rispettoso delle diversità socioculturali e ambientali.

[1] https://www.uniba.it/it/ateneo/rettorato/ufficio-stampa/comunicati-stampa/anno-2025/d-aevum-4-0-lintelligenza-artificiale-al-servizio-del-patrimonio-culturale
[2] https://www.dabimus.com
[3] https://www.uniba.it/it/ricerca/ricerca/finanziamenti-alla-ricerca/erc-seeds-uniba

BIBLIOGRAFIA

Agenda Digitale. (2024). Dalla biblioteca al database: Il futuro delle fonti storiche con IA e digitale. Retrieved from https://www.agendadigitale.eu/cultura-digitale/dalla-biblioteca-al-database-il-futuro-delle-fonti-storiche-con-ia-e-digitale/.

Agnoli, A. (2019). Le biblioteche come hub comunitari: Prospettive e opportunità. Bologna: Pendragon.

Baker, D., & Smith, L. (2020). The digital transformation of education: Technologies, perspectives and practices. London: Routledge.

Barbuti, N. (2024). Biblioteche in espansione. Digitizzazioni innovative per approcci cognitivi interattivi al libro e alla conoscenza. In Granata, G. (a cura di) (2024). I Convegno della Società italiana di scienze bibliografiche e biblioteconomiche (SISBB). Cagliari, 22-24 settembre 2022, Milano, Ledizioni, pp. 219-243.

Barbuti, N., & De Bari, M. (2024). Libri e biblioteche tra museabilità e musealizzazione digitale: sogno o realtà? In Atti del convegno AIUCD 2024: Il futuro del patrimonio digitale. Università di Catania. https://aiucd2024.unict.it/atti-del-convegno/.

Brumberger, E. (2011). Visual literacy and the digital native: An examination of the millennial learner. Journal of Visual Literacy, 30(1), 19–46. https://doi.org/10.1080/23796529.2011.11674683

Carr, N. (2010). The shallows: What the internet is doing to our brains. New York: W. W. Norton & Company.

CEPELL. (2021). Libro bianco sulla lettura e i consumi culturali in Italia. Retrieved from https://cepell.it/wp-content/uploads/2021/07/CEPELL-Libro-bianco-sulla-lettura-2021-digitale.pdf

Clifford, G. (1973). Interpretation of cultures. New York: Basic Books.

Comune di Bologna. (n.d.). Emilia Digital Library (EMILIB) - Utenti e Statistiche. https://opendata.comune.bologna.it/explore/dataset/emilia-digital-library-emilib-utenti-e-statistiche/information/ (rilasciato il 15 febbraio 2022).

De Bari, M., Barbuti, N., & Azzella, P. (2022). Letture, libri e testi in espansione: Il modello EXEbook. Per una nuova interazione con l'editoria a stampa. In Atti del II Congresso Beni Culturali in Puglia. Il patrimonio culturale pugliese. Ricerche, applicazioni e best practices (pp. 22-25). Bari.

Dominici, P. (2019). Dentro la società interconnessa. La cultura della complessità per abitare i confini e le tensioni della civiltà ipertecnologica. Milano: Franco Angeli.

Dominici, P. (2023). From emergency to emergence: Learning to inhabit complexity and to expect the unexpected. Salute e Società, 22(1), 135–151. https://doi.org/10.3280/SES2023-001010.

Esposito, A. (2021). La lettura tra tradizione e innovazione. Roma: Edizioni Cultura.

Floridi, L. (2020). The ethics of artificial intelligence. Oxford: Oxford University Press.

Gatti, L., & Colombo, F. (2023). Lettori nell'era digitale: Sfide e prospettive. Milano: Mondadori.

Giurickovic Dato, A. (2022), Figitale, parola di un presente che è già futuro. Treccani Magazine. 22 febbraio 2022 https://www.treccani.it/magazine/lingua_italiana/articoli/parole/Figitale.html.

Hine, C. (2000). Virtual ethnography. SAGE Publications.

IFLA FAIFE (Committee on Freedom of Access to Information and Freedom of Expression). (2020). IFLA Statement on Libraries and Artificial Intelligence. International Federation of Library Associations and Institutions. https://repository.ifla.org/items/8c05d706-498b-42c2-a93a-3d47f69f7646.

Ingold, T. (2007). Lines: A brief history. London: Routledge.

ISTAT. (2022). Rapporto annuale sulla lettura in Italia. https://www.istat.it.

Istituto Centrale per la Digitalizzazione del Patrimonio Culturale – Digital Library. (n.d.). Il Piano nazionale di digitalizzazione del patrimonio culturale. Ministero della Cultura. https://digitallibrary.cultura.gov.it/il-piano/.

Lamberti, F., Mellia, M. & Vivarelli, M. (2024). Biblioteche, lettura, intelligenza artificiale. Milano: Editrice Bibliografica.

Mascia, T. (2023). Pedagogia del leggere per piacere: Il ruolo della motivazione e l'identità del lettore. Pedagogia più Didattica, 9(1), 133-143. https://doi.org/10.14605/PD912309.

Mastrobattista, L. (2024). La lettura nell'era digitale: impatto sulla percezione del lettore. AIB Studi, 64(1), 27-39. https://doi.org/10.2426/aibstudi-13994.

Meschini, F. (2024). Biblioteche e Intelligenza artificiale: stiamo vivendo in tempi interessanti. Bibliotechae.it, 13 (2024), 1, 1-10. https://doi.org/10.6092/issn.2283-9364/19985. https://bibliothecae.unibo.it/article/download/19985/18180/81452.

Morriello, R. (2023). Dati e metadati bibliotecari per l'intelligenza artificiale. Biblioteche Oggi Trends, 1. http//doi.org/10.3302/2421-3810-202301-038-1; https://www.bibliotecheoggitrends.it/it/articolo/2882/dati-e-metadati-bibliotecari-per-l-intelligenza.

Nardi, A. (2015). Lettura digitale vs lettura tradizionale: implicazioni cognitive e stato della ricerca. Form@re - Open Journal per la formazione in rete, 15(1), 7-29. https://doi.org/10.13128/formare-15434.

Reading Agency. (2015). The impact of reading for pleasure and empowerment. https://readingagency.org.uk/wp-content/uploads/2024/06/The-Impact-of-Reading-for-Pleasure-and-Empowerment.pdf.

Tammaro, A. M., Di Carlo, S., Zanichelli, F. (2024). IA in biblioteca: quale impatto? I risultati di un'indagine in Italia. Biblioteche Oggi, 8. http://doi.org/10.3302/0392-8586-202408-003-1; https://www.bibliotecheoggi.it/it/articolo/16125/ia-in-biblioteca-quale-impatto-i-risultati

MeMo: Una mappa letteraria digitale per la memoria del Mezzogiorno

Laura Giurdanella[1], Giuseppe Palazzolo[2], Bernardo De Luca[3], Fara Autiero[4], Marco Gatto[5], Sabatino Peluso[6], Concetta Maria Pagliuca[7], Andrea Schembari[8]
[1] Università di Catania, Italia - laura.giurdanella@unict.it
[2] Università di Catania, Italia - giuseppe.palazzolo@unict.it
[3] Università degli Studi di Napoli Federico II, Italia - bernardo.deluca@unina.it
[4] Scuola Superiore Meridionale, Italia - f.autiero@ssmeridionale.it
[5] Università della Calabria, Italia - marco.gatto@unical.it
[6] Università della Calabria, Italia - pelusosaba@gmail.com
[7] Università degli Studi di Napoli Federico II, Italia - concettamaria.pagliuca@unina.it
[8] Università di Catania, Italia - andrea.schembari@unict.it

ABSTRACT (ITALIANO)
L'articolo presenta il progetto di ricerca PRIN 2022 PNRR MeMo che mira alla creazione di una mappa letteraria – digitale, interattiva e aderente ai principi FAIR – di testi della letteratura italiana dal secondo dopoguerra a oggi che abbiano tematizzato e fotografato l'incidenza di tre fenomeni specifici, quali la guerra, la (de)industrializzazione e l'emigrazione, nel contesto dell'Italia meridionale. In particolare, il contributo, partendo dalla necessità del progetto di preservare il patrimonio letterario del Meridione a partire da una rappresentazione spaziale digitale dei luoghi simbolici della memoria collettiva, intende presentare non solo le riflessioni teorico-critiche e metodologiche che si pongono alla base della creazione del *corpus* (principalmente secondo un approccio ecocritico e geoletterario), ma si sofferma anche sulle tecnologie utilizzate per la realizzazione della mappa digitale (mediante database online Open Source di raccolta dati).

Parole chiave: Preservazione del patrimonio letterario; Digitalizzazione e memoria culturale; Letteratura del Meridione; Ecocritica; Tecnologie per il patrimonio letterario

ABSTRACT (ENGLISH)
MeMo. *A digital literary Map for the Memory of Southern Italy*

The article presents the research project PRIN 2022 PNRR MeMo, which aims to create a literary map – digital, interactive, and adhering to FAIR principles – of Italian literary texts from the post-World War II period to the present. These texts have thematized and captured the impact of three specific phenomena: war, (de)industrialization, and migration within the context of Southern Italy.
Specifically, the contribution begins by addressing the project's necessity to preserve the literary heritage of the South through a digital spatial representation of the symbolic places of collective memory. It aims to present not only the theoretical, critical, and methodological reflections underpinning the creation of the corpus (primarily through an ecocritical and geographical-literary approach) but also focuses on the technologies employed in developing the digital map (using an Open-Source online database for data collection) and encoding the textual repertoire (in XML TEI language). These technologies are considered both tools and essential strategies for managing and visualizing large-scale corpora.

Keywords: Literary Heritage Preservation; Digitization and Cultural Memory; Southern Literature; Ecocriticism; Technologies for Literary Heritage

1. INTRODUZIONE
Il progetto MeMo si propone di costruire una mappa digitale delle esperienze letterarie del Meridione, in un arco temporale che va dal 1945 a oggi. Utilizzando gli strumenti dell'indagine geografico-letteraria e l'approccio materialista ed ecologico (ecocritica materiale), il progetto, con il suo impianto storiografico prevalente, privilegia l'analisi di luoghi simbolici e di testi legati a tre specifiche macroaree tematiche, relative tutte al rapporto tra fenomeni storico-culturali, di antropizzazione e l'ambiente, quali, il secondo conflitto mondiale, la (de)industrializzazione e l'emigrazione. I testi analizzati sono quindi visti alla luce del loro rapporto con il territorio, con i traumi della storia (De Luca, 2022), con lo spazio geografico e sociale (Gatto, 2021), con le comunità di riferimento e, in virtù del loro potenziale simbolico, come custodi di un valore memoriale e dell'interazione tra comunità antropiche e contesto spaziale (Palazzolo, 2021).

2. STATO DELL'ARTE

Negli ultimi decenni, l'attenzione critica verso il rapporto tra individuo, collettività e ambiente nei testi letterari è cresciuta, grazie alla sensibilità ecologica e allo *spatial turn* negli studi letterari (Iovino, 2006; Warf & Arias, 2009; Alfano, 2010; Iovino, 2016; Scaffai, 2017). Tuttavia, mancano studi sistematici non più egoriferiti (incentrati, cioè, sull'esperienza personale o soggettiva di un singolo autore o di una singola opera) ma piuttosto geocentrici (Bertrand, 2007), ovvero che consentano di fatto di partire dai luoghi per poi esaminare e analizzare quelle produzioni che hanno selezionato quei luoghi come spazi di rappresentazione letteraria. Non solo, sono assenti anche tentativi di accentramento degli studi critici attorno alla "questione meridionale" nella produzione letteraria. Difatti, se esiste una ricca bibliografia che, per un verso, si è interrogata criticamente sulle sorti del Mezzogiorno verso la fine degli anni Settanta (Abbate, 1968; Levy, 1996) e, per l'altro, più recentemente ha privilegiato l'analisi critica dei testi letterari attraverso gli strumenti dei cosiddetti *Trauma Studies* (Assmann, 2004), tali studi non sembrano, nel primo caso, aver trovato nuova linfa, esaurendosi, e, nel secondo caso, non sono stati capaci di coniugare le esigenze di mappatura geo-storica con contributi specifici sulla rappresentazione letteraria del rapporto tra comunità e ambiente nell'Italia meridionale.

Le prospettive di studio che allora si impongono e si integrano l'un l'altra in maniera innovativa sono almeno tre: geografica, culturale e storica. La centralità dei luoghi e degli ambienti, vista in prospettiva diacronica attraverso la lente della letteratura, implica dunque un approccio olistico che non ha ancora ricevuto un'adeguata attenzione. In tal senso, il progetto MeMo mira a colmare questa lacuna, privilegiando lo studio del rapporto tra i luoghi e le produzioni letterarie, con particolare attenzione ai processi simbolici e memoriali del Meridione. Ciò consentirà, al termine del progetto attualmente in corso, di ricostruire la complessa rete di fenomeni letterari e culturali che hanno preso forma dal secondo dopoguerra a oggi nelle regioni del Sud e delle isole, attraverso una prospettiva multipla, capace di integrare l'approccio diacronico allo studio letterario con quello geografico e culturale.

La ricerca mira infine a valorizzare il patrimonio letterario del Sud, spesso marginalizzato, e a recuperare testi legati alla rappresentazione del Meridione come riflessi di trasformazioni sociali e culturali.

3. IL PROGETTO

Il progetto MeMo mira alla creazione di (a) una mappa letteraria del Sud Italia dal 1945 a oggi, accompagnata da (b) un portale digitale interattivo, che consenta (c) uno studio integrato delle esperienze letterarie del Meridione, con (d) risvolti critico-ermeneutici e (e) didattici.

A fronte di questi obiettivi, la metodologia di indagine non può che essere multidisciplinare, per la necessità di combinare approcci storici, geografici e culturali allo studio di testi letterari che tematizzino l'influenza di tre fenomeni: gli effetti della Seconda Guerra Mondiale; l'industrializzazione e il suo eventuale declino (come nel caso di Ilva a Taranto, Italsider a Bagnoli e il Polo petrolchimico di Marina di Melilli); e i fenomeni migratori, inclusi lo spopolamento dei piccoli centri. Nello specifico, MeMo prevede:

a. La realizzazione di una mappa letteraria che unisca analisi geografica e studio di testi letterari[1] in prosa, poesia e no-fiction (biografie, saggi, inchieste, racconti odeporici, articoli di giornale…), in lingua italiana e nei vari dialetti, e che, data la centralità del dato territoriale, si focalizzi sul rapporto tra luoghi e processi di simbolizzazione memoriale, attraverso l'individuazione di aree puntuali o spazi interessati da conflitti e traumi storici che hanno attivato l'urgenza di una fissazione nell'immaginario culturale attraverso la rappresentazione letteraria (Moliterni, 2024);
b. L'allestimento di un portale web che, nel pieno rispetto dei principi FAIR (Findable, Accessible, Interoperable, Reusable) (Wilkinson, Dumontier, Aalbersberg, Appleton, Axton, Baak, Blomberg, et al., 2016) e secondo il paradigma dei Linked Open Data, contenga:
    - una mappa letteraria, realizzata mediante l'utilizzo di software di gestione dati Open Source, quale Heurist, in grado di mettere in relazione una quantità elevata di dati ricavati da ricerche in ambito umanistico, agevolmente gestibili;
    - un'interfaccia web intuitiva e *user-friendly*;
    - un ampio *corpus* testuale interrogabile mediante l'applicazione di filtri differenti (quali, luoghi e/o temi);
    - una visualizzazione dei dati letterari a partire da elementi geografici e spaziali con contributi di varia natura (materiali testuali, audio-visivi…).

[1] Non è la provenienza degli autori a risultare derimente nella costruzione del *corpus* letterario, bensì il contesto geografico meridionale cui essi fanno riferimento.

c. Un approccio "integrato" nello studio del rapporto tra luoghi e memoria culturale che implica l'interazione di molteplici quadri teorici delle scienze letterarie e sociali, secondo i quali il ricorso a una cartografia letteraria costituisce un modo per ricostruire e osservare le metamorfosi della rappresentazione memoriale del Mezzogiorno. Tale mappa digitale, restituendo il legame tra i testi, i luoghi e gli eventi traumatici, consente di fatto una panoramica complessa delle interazioni tra luoghi e memorie storiche e rende accessibile e consultabile il patrimonio letterario del Sud per una valorizzazione culturale di ampio respiro. MeMo si distingue per integrare prospettive di ecocritica materiale, studi sul trauma, e *Cultural studies*, offrendo un nuovo approccio storiografico che vede i luoghi come depositari di conflitti sociali, ambientali e memorie culturali sedimentate. Questo metodo intende rinnovare l'attenzione verso il Meridione, spesso relegato ai margini della riflessione intellettuale contemporanea.
d. Da un punto di vista critico-ermeneutico, il progetto mira a comprendere come luoghi e comunità del Mezzogiorno siano stati trasformati dai tre fenomeni selezionati e simbolicamente rappresentati:
    - la guerra evidenzia l'improvvisa trasformazione del paesaggio (magari soggetto a una secolare rappresentazione tradizionale cristallizzata) in rovina;
    - l'industrializzazione e la successiva deindustrializzazione, invece, viene presentata come un fenomeno di medio periodo che agisce sui territori modificando profondamente la loro struttura ambientale e urbana;
    - i fenomeni di migrazione in uscita, invece, nel Sud hanno rappresentato per alcuni centri una vera e propria trasformazione fantasmagorica, che ha comportato lo spopolamento e lo svuotamento non solo di strutture fisiche, ma anche della memoria immaginaria.[2]
e. Concepita come strumento in grado di coniugare la valorizzazione dei luoghi e della memoria culturale ad essi associata, la mappa, avvalendosi anche di una serie di integrazioni multimediali, grafiche e di *Data Visualization*, si offre quale strumento didattico (Mattozzi, Rabuiti, Di Tonto, 2012) di agevole fruizione (*e-learning*), i cui vantaggi sarebbero numerosi:
    - l'adozione di un approccio interdisciplinare che stimoli l'apprendimento favorendo l'intersezione tra letteratura, geografia, storia e tecnologia, mediante l'utilizzo di dispositivi digitali e software al servizio degli studi umanistici;
    - l'offerta di una contestualizzazione non solo storico-culturale ma anche spaziale delle opere, possibile grazie alla visualizzazione delle esperienze letterarie che potrebbero anche divenire oggetto di studi per ulteriori percorsi e itinerari didattici virtuali e/o reali (De Lucia, Gallo, Marino, 2014);
    - la proposta di un approccio interattivo, in grado di favorire un più alto livello di coinvolgimento e di stimolare la motivazione degli studenti, configurando lo studio letterario come un'esperienza partecipativa e immersiva (Peterle, 2016) e rendendoli di fatto protagonisti del loro percorso di apprendimento, grazie alla possibilità di applicare filtri di ricerca personalizzati (es. per temi e/o contesti geografici) e selezionare in modo mirato le informazioni di interesse;
    - la personalizzazione e la flessibilità dell'apprendimento, mediante l'utilizzo di filtri d'interrogazione del *corpus* che consentono di adattare i contenuti alle esigenze della classe o del singolo studente, ad esempio in percorsi di apprendimento personalizzati o in lavori di gruppo (*cooperative learning*);
    - l'accessibilità a risorse di varia natura, grazie all'integrazione di contenuti multimediali (testi, immagini, video e audio) che, fornendo un'esperienza multisensoriale, migliorano la comprensione dei fatti letterari;
    - la valorizzazione del territorio e della memoria dei luoghi, con la realizzazione di progetti legati a opere ambientate in luoghi specifici, promuovendo la connessione tra l'essere umano e l'ambiente, tra abitanti e patrimonio storico-geografico-culturale (Marocco, 2014);
    - lo sviluppo del pensiero critico, a partire dall'analisi dei luoghi descritti nelle opere e dai temi indagati, grazie ai quali gli studenti imparano a interpretare e valutare elementi interni al testo e anche a riflettere sul significato culturale, storico e letterario dei luoghi.

[2] Nel caso di luoghi immaginari, non fisicamente localizzabili su una carta geografica, quindi non georeferenziati o corrispondenti a pseudo-toponimi, è prevista l'assegnazione di un'apposita etichetta classificatoria ("luogo immaginario"), accompagnata dalla possibilità di inserire un commento esplicativo volto a motivare e giustificare tale attribuzione.

4. UN DATABASE E MAPPA DIGITALE PER LA LETTERATURA DEL MEZZOGIORNO

Per la realizzazione del database e della mappa digitale del progetto MeMo è stata adottata la piattaforma Heurist, uno strumento open source concepito espressamente per la gestione e la pubblicazione di basi di dati nell'ambito delle scienze umane e sociali. La scelta di tale ambiente è stata determinata dalla sua elevata flessibilità nella modellazione dei dati, che consente di adattare la struttura informativa alle specificità del dominio umanistico, nonché dalla possibilità di definire e gestire relazioni complesse tra record.[3]

Nel caso di MeMo, è stato possibile creare una struttura modulare che integra informazioni su autori, opere, luoghi e categorie tematiche (guerra, industrializzazione/deindustrializzazione e migrazione), creando relazioni gerarchiche e trasversali tra queste entità, rappresentando accuratamente la complessità delle interazioni tra letteratura e spazio geografico e garantendo una gestione flessibile e integrata di dati eterogenei e interconnessi. La scelta del tool digitale è stata guidata dalla sua aderenza ai principi FAIR e dalla possibilità di pubblicare l'intero dataset su piattaforme web o di esportarlo in vari formati, tra cui XML e SQL, assicurando la sostenibilità del progetto a lungo termine.

Il database è stato progettato seguendo uno schema che rispecchia le esigenze di analisi del progetto. Le principali entità prese in considerazione sono:

a. *Luoghi*. Ogni luogo è fissato tramite coordinate geografiche precise e classificato secondo la tipologia (es. città, sito industriale, luogo immaginario). Per ogni sito sono registrate informazioni inerenti alla descrizione, ai riferimenti testuali (opere collegate), e sono associati dei prodotti multimediali (immagini o video).
b. *Opere*. Le schede relative alle opere includono informazioni su autore, titolo, edizione di riferimento, lingua, genere e collegamenti a luoghi e categorie tematiche. Si è deciso di introdurre in questo punto una doppia contestualizzazione temporale, distinguendo tra *Tempo reale* (anno di pubblicazione dell'opera) e *Tempo finzionale* (anno o arco temporale trattato nel testo).
c. *Autori*. Ogni autore è corredato da una scheda che include dati biografici, identificativi e rimandi a *authority file* nazionali e internazionali (es. ISNI, VIAF), elenco delle opere associate.
d. *Categorie tematiche*. Guerra, industrializzazione e migrazione sono le tre macroaree che collegano testi, luoghi e fenomeni storici.

La mappa digitale del progetto è stata sviluppata utilizzando il framework Leaflet, uno strumento Open Source ideale per visualizzare dati geospaziali in modo interattivo e dinamico. Questo strumento è stato scelto per la sua leggerezza, che consente performance rapide e un'ottima esperienza utente, nonché per la grande flessibilità alla personalizzazione, che lo rende adatto a soddisfare le esigenze specifiche del progetto MeMo. La mappa si integra perfettamente con Open Street Map, una soluzione Open Source che garantisce aderenza ai principi FAIR, facilitando l'accesso ai dati e la loro interoperabilità. L'interfaccia intuitiva di Leaflet permette agli utenti di esplorare le connessioni tra luoghi, opere e autori, rendendo la mappa uno strumento accessibile e altamente fruibile.

Le funzionalità della mappa sviluppata da MeMo includono la navigazione per luogo, che consente di visualizzare i dettagli principali dei siti selezionati, come descrizioni e immagini, con la possibilità di approfondire ulteriormente i dati correlati (vd. Fig.1).

[3] Tra i principali punti di forza della piattaforma si annoverano la capacità di aggiornare in tempo reale lo schema concettuale, la disponibilità di interfacce grafiche intuitive per la visualizzazione e l'analisi dei dati, e la predisposizione nativa alla pubblicazione online in ambienti dinamici e interattivi. Queste caratteristiche facilitano non solo la disseminazione dei contenuti, ma anche il loro riuso in contesti accademici e didattici.

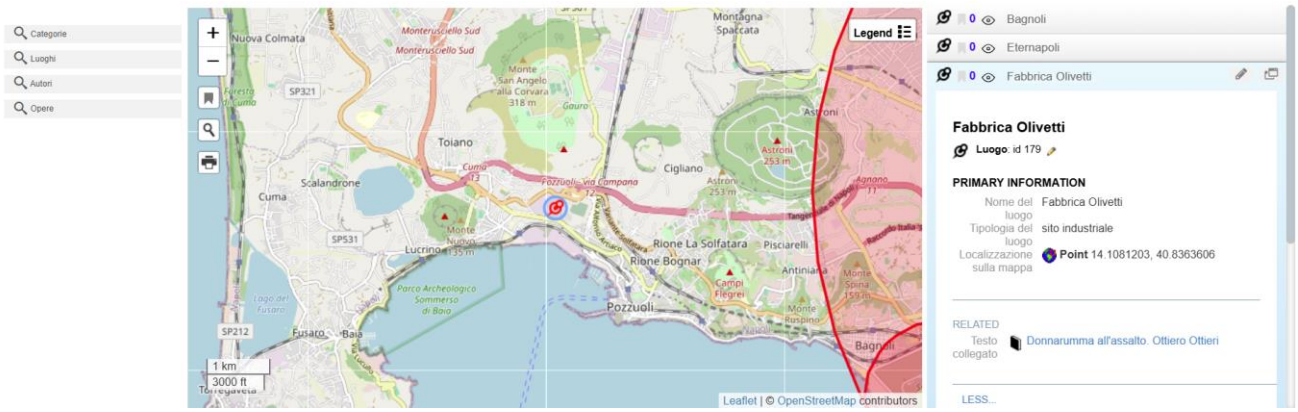


Figura 1. Navigazione per luogo

Gli utenti possono esplorare le opere letterarie associate ai luoghi e accedere direttamente alle schede descrittive, che includono estratti testuali e collegamenti ad altri elementi del database. Inoltre, la mappa offre una visualizzazione tematica attraverso layer distinti per le categorie principali di analisi – guerra, industrializzazione e migrazione – facilitando una consultazione mirata e approfondita.
Un elemento particolarmente interessante a corredo della mappa è rappresentato dalla presenza di filtri avanzati, che consentono agli utenti di affinare la loro ricerca in base a parametri specifici come il periodo storico, il genere letterario o le categorie tematiche. Questo sistema di filtraggio si rivela fondamentale per la ricerca scientifica, poiché permette di individuare *pattern* significativi e di analizzare le correlazioni tra i diversi dati. Per esempio, un utente interessato alla rappresentazione della deindustrializzazione potrebbe selezionare esclusivamente i luoghi e i testi collegati a questo fenomeno in una determinata finestra temporale, ottenendo una visione chiara e dettagliata dei risultati. Inoltre, i filtri aiutano a mettere in luce interconnessioni altrimenti difficili da individuare, offrendo nuove prospettive interpretative e facilitando la produzione di analisi critiche approfondite. Questo approccio rende la mappa un potente strumento non solo per la consultazione, ma anche per l'esplorazione e la scoperta di nuove linee di ricerca.

5. ULTERIORI SVILUPPI

Il progetto MeMo si propone di evolvere ulteriormente per ampliare il proprio impatto scientifico e culturale con una serie di sviluppi programmati. Innanzitutto, si punta a integrare la mappa con risorse digitali internazionali, come Europeana[4] e GeoNames,[5] per creare un ecosistema di dati interconnessi che favorisca una visione globale del patrimonio culturale. Questa integrazione permetterà di arricchire il *corpus* di opere e autori, includendo nuovi testi e materiali digitali, con particolare attenzione alla letteratura contemporanea e agli autori meno noti, per garantire una rappresentazione inclusiva e variegata.
In una seconda fase di espansione del progetto, si prevede inoltre l'implementazione del portale attraverso l'aggiunta di integrazioni puntuali di carattere critico-ermeneutico. Tali interventi saranno elaborati a partire dagli esiti del convegno conclusivo, previsto entro la fine dell'anno, il cui contributo sarà fondamentale per arricchire e approfondire i livelli di lettura della mappa. L'obiettivo è quello di guidare il lettore lungo percorsi di senso articolati e coerenti, capaci di valorizzare la complessità del *corpus* letterario e il suo ancoraggio al contesto geografico meridionale.
Un ulteriore obiettivo è l'arricchimento multimediale della mappa. Saranno integrati materiali audiovisivi, come interviste agli autori e documentari sui luoghi rappresentati, per offrire agli utenti un'esperienza più immersiva e multisensoriale. Parallelamente, si prevede lo sviluppo di percorsi didattici interattivi destinati a studenti e docenti al fine di stimolare l'interesse per la letteratura del Mezzogiorno, presentando contenuti personalizzabili e adattabili a diversi contesti educativi.
Il progetto mira anche a stabilire collaborazioni internazionali con istituzioni accademiche e culturali europee e mediterranee. Questi partenariati consentiranno di condividere metodologie e risultati, promuovendo uno scambio di conoscenze e un dialogo interdisciplinare. Infine, saranno sviluppate

[4] https://www.europeana.eu/it (cons. 25/01/2025).

[5] https://www.geonames.org (cons. 25/01/2025).

funzionalità avanzate di analisi dei dati, utilizzando tecniche di *machine learning* per identificare *pattern* nascosti e trend evolutivi nella rappresentazione letteraria dei luoghi.
Una delle principali direttrici di sviluppo futuro del progetto consisterà nell'integrazione di strumenti e tecnologie fondate sull'intelligenza artificiale, con particolare attenzione alle applicazioni del *Natural Language Processing* (NLP). In tale ambito, l'implementazione di modelli per il riconoscimento automatico delle entità nominali (*Named Entity Recognition*) e per l'estrazione delle relazioni (*Relation Extraction*) si configura come una valida opportunità per l'arricchimento del database. Tali strumenti potranno, infatti, contribuire in modo significativo alla normalizzazione dei nomi propri, alla disambiguazione delle varianti onomastiche e alla ricostruzione delle reti semantiche e relazionali tra persone, luoghi e opere favorendo una più efficace strutturazione e interoperabilità dei dati.
MeMo si fonda sui principi FAIR, promossi a livello europeo per una gestione sostenibile e responsabile dei dati della ricerca, garantendo per ogni entità registrata l'assegnazione di URL univoci e persistenti, l'accessibilità tramite un'interfaccia web pubblica corredata da documentazione esaustiva, l'interoperabilità attraverso uno schema dati aperto e leggibile, nonché la riusabilità grazie all'impiego di metadati espliciti e licenze d'uso chiare. Nell'ottica di una sempre maggiore apertura e interconnessione tra basi di dati, il progetto MeMo mira a sviluppare strategie di allineamento semantico con risorse esterne del Linked Open Data al fine di collegare le entità presenti nel database (persone, luoghi, opere) a nodi già esistenti del grafo della conoscenza globale, ampliando così le potenzialità di interrogazione, navigazione e interoperabilità (attualmente l'allineamento è attivo solo per le entità "persona" attraverso i riferimenti a VIAF e ISNI).
Questi sviluppi consolideranno il ruolo della mappa letteraria digitale come strumento di riferimento per lo studio interdisciplinare della memoria culturale del Mezzogiorno, coniugando innovazione tecnologica e rigore scientifico. La mappa non sarà solo un prodotto accademico, ma anche una risorsa dinamica e accessibile, capace di valorizzare e diffondere il patrimonio culturale del Sud Italia a un pubblico sempre più vasto.

BIBLIOGRAFIA

Abbate, M. (1968). L'alternativa meridionale. Roma-Matera: Basilicata Editrice.

Alfano, G. (2010). Paesaggi, mappe, tracciati. Cinque studi su letteratura e geografia. Napoli: Liguori.

Assmann, A. (2004). Ricordare. Forme e mutamenti della memoria culturale. Bologna: Il Mulino.

Bertrand, W. (2007). La Géocritique. Réel, fiction, espace. Paris: Éditions de Minuit.

De Luca, B. (2022). Il tempo diviso. Poesia e guerra in Sereni, Fortini, Caproni, Luzi. Roma: Salerno editrice.

De Lucia, S., Gallo, C., Marino, D. (Eds). (2014). Landscapes and mindscapes: metodologie di ricerca, percorsi geocentrati e poetiche dello spazio in una prospettiva comparata. Napoli: Marchese.

Gatto, M. (2021). Dal Sud si vede l'Italia. L'eredità politica e culturale di Alessandro Leogrande, L'Ospite Ingrato, 1-4.

Iovino S. (2016). Bodies of Naples: Stories, Matter, and landscape of porosity. London-New York: Bloomsbury.

Iovino, S. (2006). Ecologia letteraria. Una strategia di sopravvivenza. Milano: Ambiente.

Levy C. (ed). (1996). Italian Regionalism. History, Identity, and Politics. Oxford-New York: Berg.

Marocco, M. (2014). Un luogo esiste solo se raccontato. La dimensione letteraria come strumento di indagine dello spazio periurbano, Territorio, 68, 1, 117-124.

Mattozzi, I., Rabuiti, S., Di Tonto, G. (Eds). (2012). Il tema: Geostoria e Didattica, Il Bollettino di Clio, XIII, 0, maggio.

Moliterni, F. (2024). Finzioni meridionali. Il Sud e la letteratura italiana contemporanea. Roma: Carocci.

Palazzolo, G. (2021). Apocalisse e profezia. Franco Fortini critico e poeta. Roma: Carocci.

Peterle, G. (2016). Leggere, scrivere, abitare: proposte di ricerca e didattica tra geografia e letteratura. AST, LXI, 4, ottobre-dicembre, 26-30.

Scaffai, N. (2017). Letteratura e ecologia. Forme e temi di una relazione narrativa. Roma: Carocci.

Warf, B., Arias, S. (2009). The Spatial Turn. Interdisciplinary Perspectives. London: Routledge.

Wilkinson, M., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N. et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 3, 160018, https://doi.org/10.1038/sdata.2016.18.

The relationship between art and sound: An experiment on the engagement of the cultural tourist

Sara Benetti[1], Nicola Orio[2]
[1] Università di Padova – sara.benetti.11@studenti.unipd.it
[2] Università di Padova – nicola.orio@unipd.it

ABSTRACT (ENGLISH)

In recent years many cultural institutions have been moving towards the design of visit experiences that engage visitors through multisensory immersive experiences. These tours are constructed using cross-modal stimuli, as well as augmented and virtual reality, which allow the combination and overlaying of real and digital elements. Along the line of multisensoriality, the research presented in this paper addresses the relationship between sight and hearing during the enjoyment of a cultural site. The focus is on how their combination can influence visitors' emotions and perceptions. To this end, we designed a multisensory experience where participants listened to a narrative based solely on sounds and music (no speech) while visiting a real cultural site. We evaluated the effectiveness of the approach by measuring visitors' responses to three different stimuli. Results show that sound can be exploited as an aid to improve the visit of a cultural site.
Keywords: multisensory experience; emotional responses; acoustic stimuli; culture enjoyment

ABSTRACT (ITALIANO)

Analisi del rapporto tra arte e suono: un esperimento sul coinvolgimento del turista culturale

Negli ultimi anni le istituzioni culturali si stanno orientando verso la creazione di esperienze di visita in grado di coinvolgere il pubblico attraverso esperienze immersive multisensoriali. Percorsi di questo genere sono costruiti mediante l'uso di stimoli cross-modali, ma anche di realtà aumentata e virtuale che permettono di combinare e sovrapporre elementi reali e digitali. La ricerca descritta in questo lavoro segue la direzione della multisensorialità, focalizzandosi come la combinazione tra vista e udito possa migliorare la fruizione di un sito culturale, in particolare su come possa influenzare la percezione dei visitatori. A questo scopo, è stata progettata un'esperienza multisensoriale all'interno della quale i partecipanti ascoltavano una narrazione sonora basata esclusivamente su musica e rumori (senza l'utilizzo del parlato) durante la visita di un sito culturale. Le risposte dei visitatori a tre diversi stimoli sonori sono state valutate attraverso la somministrazione di un questionario. I risultati indicano che le tracce sonore possono essere utilizzate per aumentare il coinvolgimento dei visitatori.
Parole chiave: esperienza multisensoriale; risposte emotive; stimoli acustici; fruizione culturale

1. INTRODUCTION

In current museum environments, visitor's engagement is becoming increasingly relevant, and experiences are designed to be immersive on both emotional and sensorial aspects. Moreover, there is a growing interest towards inclusivity and accessibility for visitors with sensorial impairments. The common approach oriented towards 'seeing' is shifting to an approach oriented towards 'interacting' (Chivaran et al., 2022). The visitor's role is changing from being a passive consumer, to taking on an active role in the learning processes (Coppola & Zanazzi, 2020).

A widely used approach to multisensoriality involves blending of visual and auditory elements. Various studies showed that the combined presentation of these stimuli can intensify participants' emotional engagement. To this end, we designed an experiment to measure the effectiveness of three different acoustic experiences for visitors of a cultural site. The goal was to analyze to what extent the acoustic accompaniment interferes with the enjoyment of artworks.

2. RELATED WORK

A common tool in museums is the audio-guide: it allows visitors to improve their knowledge about a particular artwork or object through a voice that describes its appearance, history, and the context in which it was created (Mandarano, 2021). Sometimes the audio track includes music, which contribute to the construction of the narrative (Mildorf, 2024). Music and sounds in audio guides help setting the right mood and evoke sensations that increase the enjoyment of what is been said in the recording.

An experiment carried out by Miu et al. (2016) compared the emotions raised by watching a painting or listening to music by measuring the frequency of emotions and their relationship with other factors

(background, mood, kind of stimulus). Results showed that art and music tend to evoke different emotions when the Geneva Emotional Music Scale (GEMS) has been used. GEMS has been described as "the first model and instrument specifically designed to capture the richness of musically evoked emotions"[1]. The same scale was also chosen for the experiment in the present project.
Other studies investigated the relation between images/artworks and sounds/music and how it can condition the individual perception. For instance, the aim of the study conducted by De Benedetto et al. (2024) was to examine whether look at a painting while listening to a classical music piece with an emotionally incongruent content could alter the perception of the visual stimulus. The experiment was conducted in an anechoic chamber recording the brain activity of 25 participants. The results show that the emotional incongruence between music and painting alters the painting's emotional experience.
Baumgartner et al. (2006) aimed at understanding which brain regions were activated by the presentation of audiovisual stimuli first separately and then in combination. Also in this case, the experiment took place in a laboratory setting using various measurement parameters (e.g., EEG, EOG, HR, SCR) with 24 participants. The stimuli were chosen in association with three basic emotions (happiness, sadness, fear) from classical music pieces and images from the International Affective Picture System (IAPS). They observed a more intense emotional perception when images and music were presented together. It can be concluded that music can significantly amplify the emotional experience evoked by images. Another study on basic emotions raised by art and music (Braun Janzen et al., 2023) involved 142 participants who could admire a Kandinskij's painting in person at the Museo de Arte Contemporânea da Universidade de São Paulo while listening to musical excerpts by Schönberg through headphones. The surveys filled out by the participants reveal a tendency to rate the artwork generally more positively if the background music was appreciated. The comparison of these outcomes with the control group showed that the simultaneous presentation of emotionally congruent audiovisual stimuli intensified the emotional experience.
There studies can be extended also to the visit of museums and cultural sites. As theorized by Webb (1995), background music during a visit can influence mood, pleasantness and time perception. These observations are in line with the findings of a study carried out by Chen & Tsai (2015) at the Laiho Memorial Museum in Taiwan with 20 participants and two different music pieces. Brenner (2016) has done a similar experiment with 50 participants at the Renton History Museum. She compared the feedback of the experience with and without background music observing an influence of the music on length, comfort and engagement of the visit. In both cases only musical pieces were used and they served as background during the visit.

3. EXPERIMENT AND METHODS

Most researchers focused on the relationship of visual art and music, while environmental sounds and noises made by objects and persons have been left aside. In our study we aim to investigate how the *soundscape* can interfere with enjoyment of a cultural place.
Thus, the goal of our experiment was to analyze if and how an auditory accompaniment to the visit of a cultural site impacts the elicited emotions and the way the experience is described. Based on the literature review, the expected result is that the combination of visual stimuli (seeing the interiors of a chapel) and auditory stimuli (hearing audio recordings of soundscapes) will allow participants to live an immersive visit and intensify their perceptions.
We aimed at developing an experiment in a real scenario, because a laboratory setting would not be suitable for providing the immersive experience that should be elicited by soundscapes. The chosen place is the Scoletta del Carmine in Padua (Figure 1), a chapel adjacent to the Parish of Santa Maria del Carmine. It was built in the 1300s and in the interiors are present frescoes from the 16th century realized by Giulio and Domenico Campagnola, Girolamo Tessari 'dal Santo' and Stefano dall'Arzere.[2] Despite its richness, it is not very well known by citizens and tourists.
We realized three audio tracks[3] that were related to the function of the chapel by providing an acoustic narrative of the place but were not specifically linked to the frescos. They aim to evoke moments when the place is *alive*, like during concerts or religious celebrations. All recordings are all less than three-minute long to avoid a loss of attention and concentration of the participants. Two registrations are constructed

[1] Geneva Emotional Music Scale (GEMS), https://musemap.org/resources/gems

[2] La Scoletta del Carmine in Padova, by Litardi, A., Parrocchia di Santa Maria del Carmine, https://carminepd.it/storia-e-arte/56-la-scoletta-del-carmine-in-padova

[3] The audio files are available at this link; https://www.game4culture.com/carmine/

through a sequence of ambient sound and music to create a narrative and allow listeners to imagine events happening around them. The third one in purely made of music.



Figure 1. Interiors of the Scoletta del Carmine

This should enable them to connect more deeply with their surroundings, experiencing an immersive environment. The sounds aim at bringing the chapel to life, extending beyond its function as a tourist destination.
A description of the audio recordings is provided in Table 1.

Stimulus	Description
Prayer	A moment of prayer begins in the empty chapel: a door closes, steps sound of a small group of people walking through the nave, noise of a book pages being flipped, a choir starts singing a Latin version of the 'Ave Maria' prayer that fades away at the end of the audio
Mass	Beginning of a mass in the chapel full of people: bells ringing, the noise of a crowd, a bell rings to start the celebration, steps sounds, the organ plays a lively piece that fades away at the end of the audio
Organ	A piece of music with a cheerful tone, as if played by an organist practicing

Table 1. Narratives associated to the three stimuli

The experiment took place in a single day with 143 participants, divided in small groups of no more than 20 persons. They were all students at the School of Human and Social Sciences and Cultural Heritage at the University of Padua. Instructions were to visit the Scoletta while listening to one of the three audio tracks, available through a QR-code, using the headphones of their personal device. A control group of participants did the visit without the audio. After the visit, participants were asked to fill a survey and then, if they liked, to listen to the other audio tracks and provide verbal feedback on which they considered the best accompaniment for the visit.
The survey was in Italian and included four sections. Firstly, participants indicated which audio they listened to during the visit (or no audio). The second section collected feedback about the pleasantness of the visit, its most significant aspects, and the quality of the audio. In the third section, participants were asked to indicate the intensity of various emotions experienced during the visit. The last section was dedicated to final comments and suggestions about the experiment. The control group did not answer the second section about the audio assessment.
According to the GEMS scale, the nine addressed emotions were: wonder, transcendence, nostalgia/longing, tenderness, tranquility, joy, power, tension, sadness[4].

4. RESULTS

The first output is that the majority of the participants considered the visit with audio accompaniment to be pleasant and would recommend it to a friend (Figure 2), according to a Likert scale from 1 ('not at all') to 7 ('a lot'). The figure shows that the acoustic narratives with environmental sounds were more appreciated than the music-only audio. The audio 'organ' was appreciated but the tendency to recommend it is lower than the others. In general, the audio 'organ' was generally less engaging than the other two.

[4] "9 great composers explained in 9 emotions | Marcel Zentner | TEDxInnsbruck", TEDx Talks, https://www.youtube.com/watch?v=_WPOsrRQfkk

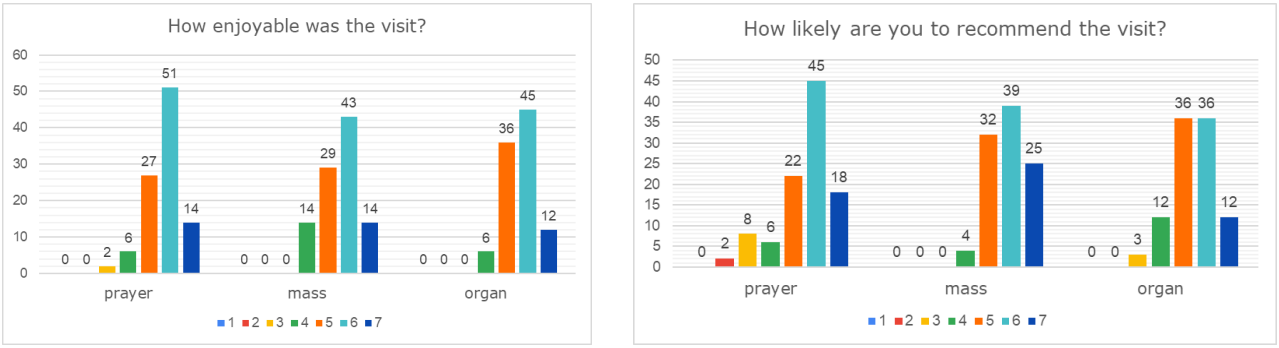


Figure 2. Distribution of responses to the questions about the appreciation of the visit

Using the same scale, participants were asked to evaluate the pleasantness of the audio tracks. The audio 'prayer' was appreciated by the 39% of the participants (considering responses with 6 and 7 points), the audio 'mass' by the 43%, and the 'organ' one by the 45%.
The survey also addressed the evocative power of the sound accompaniment. 44.6% of the participants answered affirmatively and some of them provided more details in an open-end question. As showed in the word cloud (Figure 3), frequent words are 'family' (famiglia), 'Sunday' (domenica), 'childhood' (infanzia) but also 'art' (arte), 'frescos' (affreschi).



Figure 3. Word cloud generated based on the question
'Did the visit bring to mind any particular memories, situations, or past experiences?'

As regards the emotional intensity experienced during the visit, it is possible to compare the responses between participants who did the experience with and without the audio. We used the same Likert scale from 1 ('not at all') to 7 ('a lot'). The emotions 'tension' and 'sadness' were in general rarely experienced and the modal value is 1 in both cases.
In the graph related to 'transcendence' (Figure 4), it can be observed a concentration of responses in the values 2 and 3 of rating scale for the audio 'prayer' and for the visit in silence (corresponding to the visit without the audio tracks). For the audio 'mass' the modal value is 6, whereas for the 'organ' one the responses are more distributed along the entire scale.

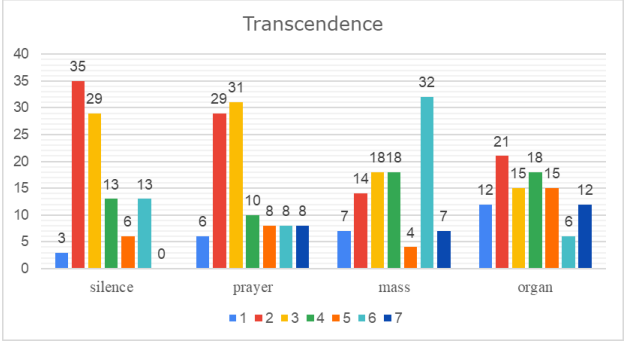


Figure 4. Distribution of responses on emotional intensity in relation to transcendence

Regarding ‘peace’, it is interesting to notice that the percentage of participants who experienced a strong emotional intensity (summing up the responses with values 6 and 7) is higher in relation to the audio 'mass' (36%), which is the loudest among the three. In the other cases, this emotion was experienced very intensely by 29% of the control group, 26% of the audio 'prayer', and 33% for the audio 'organ'.
The ‘nostalgia’ graph (Figure 5, left) shows that the range of responses for visit without audio goes from 1 to 5, whereas in presence of audio are present responses of high emotional intensity. From the ‘joy’ graph (Figure 5, right) it is possible to notice a similar distribution of responses for silence and ‘organ’. Similar results have been obtained also to ‘wonder’ and ‘tenderness’. This may highlight a possible difference and emotional engagement between the narration with than the visit with silence or music-only.

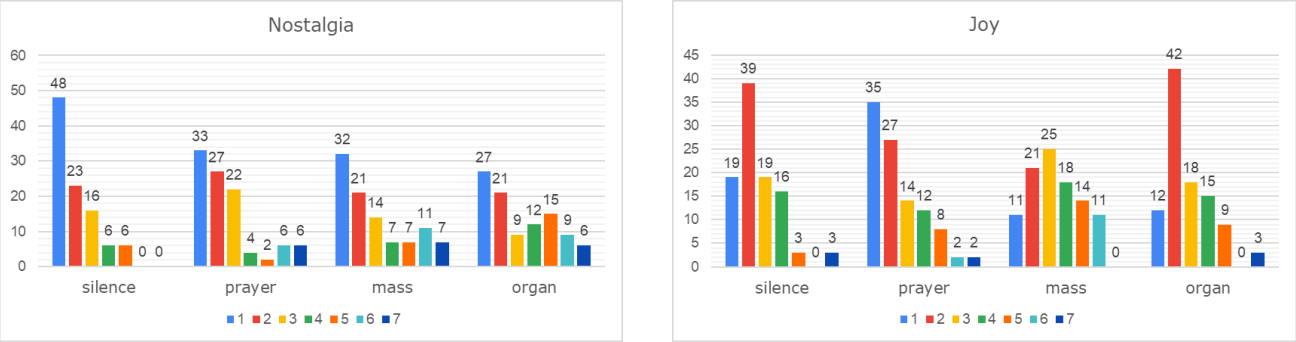


Figure 5. Distribution of responses on emotional intensity in relation to 'nostalgia' and 'joy'

Finally, ‘power’ was not experienced by participants and the responses are concentrated in the low values of the scale except for the ‘mass’ audio, where the responses are more distributed (showing more interference).
Participants could leave comments and suggestions about the experience both in the last part of the survey and verbally. In general, the experiment received positive feedback and considered able to bring people closer to places with historical and artistic importance. Several participants asked for longer audio tracks and the inclusion of information about the historic and artistic elements present in the environment. In additions, some suggested more connection between the music and sounds and the fresco scenes decorating the walls, that could be described through the audio.
At the end, participants were invited to verbally share their personal judgement on which of the three audio they considered the best and most suitable accompaniment for the visit of the Scoletta del Carmine. The responses were quite varied, but it is interesting to notice that around 47% of participants (who visit the place with the audio tracks) indicated as favorite the first one that they listened. This can be due to a priming effect, where exposure to a given stimulus influence the perception and the processing of subsequent information, even in the absence of explicit awareness of the influence.

5. CONCLUSIONS AND FUTURE WORK

The work presented in this paper explores how an acoustic narrative, based solely on environmental sounds and music, can improve the experience of visiting a cultural site. Results of the experiment we carried out show that participants appreciated the ability of acoustic narratives to elicit memories and increase the engagement during the visit. The effect on the emotions evoked by the audio tracks depended on both the stimulus and the kind of emotion, with some noticeable differences compared to the results obtained from a control group.
Our experiment was preliminary in nature. We are aware that the sample, consisting solely of students, is not representative of the entire population of potential visitors to the cultural site yet it can provide initial hints on how nonverbal audio can be exploited. The research is at its early stages, as results need to be confirmed by a larger number of acoustic narratives tested in different cultural sites. In particular, the fact that participating to a mass is a shared experience of many participants may explain the specific effect of one of the stimuli. Additional experiments should be carried out in different environments to verify this.
Another direction of research is the analysis of the contribution of music towards a pure soundscape, assuming that the former could be more likely to elicit emotions while the latter might be more effective to evoke memories. In conclusion, further research could also compare visit experiences similar to the one here experimented with more traditional ones, such as guided tours and audioguides.

REFERENCES

Bailey, K., Broackes, V., & de Visscher, E. (2019), "The longer we heard, the more we looked" Music at the Victoria and Albert Museum. Curator, 62, 327-341. https://doi.org/10.1111/cura.12334.

Baumgartner, T., Esslen, M., & Jäncke, L. (2006). From emotion perception to emotion experience: Emotions evoked by pictures and classical music, International Journal of Psychophysiology, 60(1), 34-43, ISSN 0167-8760, https://doi.org/10.1016/j.ijpsycho.2005.04.007.

Braun Janzen, T., de Oliveira, B., Ventorim Ferreira, G., Sato, J. R., Feitosa-Santana, C., & Vanzella, P. (2023). The effect of background music on the aesthetic experience of a visual artwork in a naturalistic environment. Psychology of Music, 51(1), 16-32. https://doi.org/10.1177/03057356221079866.

Brenner, B. (2016). Does music matter to museum visitors?: Understanding the effect of music in an exhibit on the visitor experience. ProQuest Dissertations Publishing. (Order No. 10138687). Available from Dissertations & Theses @ University of Washington WCLP; ProQuest Dissertations & Theses Global. (1804414286). Retrieved from https://search.proquest.com/docview/1804414286?accountid=14784

Chen, C.-L., & Tsai, C.-G. (2015). The influence of background music on the visitor museum experience: a case study of the Laiho Memorial Museum, Taiwan, Visitor Studies 18(2), 183-195.

Chivaran, C., Nappi, M. L., Capece, S., & Buono, M. (2022). Multisensory museum models for knowledge transfer, Advances in design and digital communication, II proceedings of the 5th international conference on design and digital communication, 4–6 November 2021.

Coppola, S., & Zanazzi, S. (2020). Experiencing art. The role of immersive technologies in museum education. Formazione & Insegnamento, 18(2), 36–49. https://doi.org/10.7346/-fei-XVIII-02-20_04.

De Benedetto, F., Ghiraldini, E., Capizzi, N., & Proverbio, A. M. (2024). Unraveling the interplay of emotions in art and music: An event-related potential investigation. Psychology of Music. https://doi.org/10.1177/03057356241245997.

Mandarano, N. (2021). Musei connessi: le nuove tecnologie nei musei romani. Rapporto sui Musei Romani redatto dal CROMA, Centro per lo studio di Roma dell'Università di Roma Tre.

Mildorf, J. (2024). Music and Narrative in Art Exhibition Audio Guides: "Beckmann & Amerika." In C. Van Nerom, A. Peeters, & B. Bouckaert (Eds.), Music and its Narrative Potential (1st ed., pp. 201–218). Brill Fink.

Miu, A.C., Pițur, S. & Szentágotai-Tătar, A. (2016). Aesthetic Emotions Across Arts: A Comparison Between Painting and Music. Front. Psychol. 6:1951. doi: 10.3389/fpsyg.2015.01951

Spallazzo, D., Spagnoli, A., & Trocchianesi, R. (2009). Il museo come organismo sensibile, Tecnologie, linguaggi, fruizione verso una trasformazione designed-oriented, Congresso Nazionale AICA, Associazione Italiana Informatica e Calcolo Automatico, Un nuovo "made in Italy" per lo sviluppo del Paese. ICT per la valorizzazione dei beni e delle attività culturali (Roma, 4-6 novembre), vol. 6.

Webb, R. (1995). Music, mood, and museums: a review of the consumer literature on background music, Visitor Studies 8(2), 15-29.

Analisi RTI delle iscrizioni runiche del Leone del Pireo (Arsenale di Venezia)

Paola Peratello[1], Elisa Corrò[2]
[1] Università Ca' Foscari Venezia, Italia - paola.peratello@unive.it
[2] Università Ca' Foscari Venezia, Italia - elisa.corro@unive.it

ABSTRACT (ITALIANO)

Questo poster presenta l'utilizzo del *Reflectance Transformation Imaging* (RTI) nel processo di analisi e digitalizzazione delle iscrizioni runiche del Leone del Pireo presso l'Arsenale di Venezia per pervenire a una lettura aggiornata dei testi iscritti attraverso una ricostruzione basata su dati nuovi. Questo lavoro intende cogliere il suggerimento di proseguire con lo studio testuale formulato da Snædal, autrice dell'ultima pubblicazione sulle iscrizioni (Snædal, 2014), e segue l'esempio di precedenti ricerche nell'ambito dell'epigrafia digitale (Akcay, 2015; Lech, Matera & Zakrzewski, 2021; Ponchio, Lamé, Scopigno & Robertson, 2018). Inoltre, il lavoro rientra nell'attività di ricerca "Patrimonio culturale digitalizzato e narrazione partecipata" (supervisor: Marina Buzzoni) e costituisce un obiettivo dello Spoke 9 - CREST (*Cultural Resources for Sustainable Tourism*), coordinato dall'Università Ca' Foscari Venezia all'interno del Progetto PNRR (2022-2025) CHANGES (*Cultural Heritage Active Innovation for Sustainable Society*). Scopo dell'attività di ricerca è lo sviluppo di modelli per la trasformazione digitale del patrimonio culturale italiano, con particolare attenzione alla cultura germanica, alla quale appartengono le iscrizioni runiche del Leone del Pireo. A questo si aggiunge un altro fine, divenuto particolarmente urgente anche per il Leone del Pireo, ossia l'impegno alla conservazione dei beni culturali in pericolo e soggetti a deterioramento: il monumento, posto all'esterno dell'Arsenale di Venezia, è infatti esposto ad erosione, quindi anche le iscrizioni risultano a rischio. I risultati raccolti saranno utilizzati nell'ambito di un'esposizione museale sul Leone del Pireo in programmazione all'interno del Museo Storico Navale. Con questo poster si intende porre l'accento sull'impatto positivo dell'uso della tecnologia nel rendere visibile e accessibile alla comunità ciò che a occhio nudo non lo è più nello studio delle varie categorie di beni culturali.
Parole chiave: Reflectance Transformation Imaging (RTI); Iscrizioni runiche; Leone del Pireo (Arsenale di Venezia); Accessibilità al patrimonio culturale

ABSTRACT (ENGLISH)

RTI analysis of the runic inscriptions on the Piraeus Lion (Venetian Arsenal).

This poster presents the use of Reflectance Transformation Imaging (RTI) in the process of analysing and digitising the runic inscriptions of the Piraeus Lion at the Venetian Arsenal in order to produce an up-to-date reading of its inscribed texts based on new data. This work puts into practice the suggestion to further analyse the texts proposed by Snædal, author of the last publication on the inscriptions (Snædal, 2014) and follows the example of previous research in the field of digital epigraphy (Akcay, 2015; Lech, Matera & Zakrzewski, 2021; Ponchio, Lamé, Scopigno & Robertson, 2018). Furthermore, this work is included in the research project 'Patrimonio culturale digitalizzato e narrazione partecipata' (supervisor: Marina Buzzoni) and constitutes one of the objectives of Spoke 9 - CREST (Cultural Resources for Sustainable Tourism), coordinated by Ca' Foscari University of Venice, and in one of the operational lines of the PNRR (2022-2025) Project CHANGES (Cultural Heritage Active Innovation for Sustainable Society). The aim of the research project is to develop models for the digital transformation of Italian cultural heritage, with a focus on Germanic culture, to which the runic inscriptions of the Piraeus Lion belong. In addition to this, there is another goal, which has also become particularly urgent for the Piraeus Lion, namely the preservation of cultural assets in danger and subject to deterioration: the monument, located outside the Venetian Arsenal, is in fact exposed to erosion, so the inscriptions are also at risk. The results collected will be included in a museum exhibition on the Piraeus Lion scheduled at the *Museo Storico Navale*. The aim of the poster is to emphasise the positive impact of the use of technology in making visible and accessible to the community what is now invisible in the study of various categories of cultural heritage.
Keywords: Reflectance Transformation Imaging (RTI); Runic inscriptions; Piraeus Lion (Venetian Arsenal); Accessibility to cultural heritage.

1. INTRODUZIONE

In Italia, le iscrizioni in caratteri runici rappresentano una percentuale nettamente inferiore rispetto alla produzione di iscrizioni latine: le prime si trovano principalmente a San Michele Arcangelo nel Gargano e a

Monte Sant'Angelo a Roma (Schwab, 2009). Le iscrizioni runiche del Leone del Pireo all'Arsenale di Venezia si distinguono dalle altre perché sono le uniche, ad oggi, tardo-medievali e di origine Variaga, quindi Scandinava. Come per i testi manoscritti e a stampa, anche i testi iscritti corrono il rischio di non essere più completamente o totalmente leggibili, in quanto non protetti da teche, per esempio, o perché spesso esposti ad intemperie o ad un illecito intervento umano. Tra questi casi, rientrano le tre iscrizioni runiche del Leone del Pireo. Esse sono incise sui lisci fianchi della statua marmorea greca del IV secolo: le superfici omogenee si prestavano, infatti, ad incisioni estese ed elaborate.[1] Sono, tuttavia, parti molto esposte alle intemperie e questo incide sul grado di leggibilità dei testi. Il metodo di analisi del *Reflectance Transformation Imaging* (RTI) non è mai stato applicato alle iscrizioni runiche del Leone, ma è stato suggerito dalla filologa Snædal che, a conclusione dello studio pubblicato nel 2014, auspica che i dettagli delle iscrizioni, con particolare riferimento all'iscrizione del fianco destro, possano essere analizzati con tecnologie avanzate come il RTI dome e lo scanning 3D. Si può leggere di più di quello che è già stato identificato senza l'utilizzo delle tecnologie sopra citate? Se sì, cosa? Queste sono infatti le domande di ricerca che guidano il presente progetto.

2. STATO DELL'ARTE

Reflectance Transformation Imaging (RTI)

Il *Reflectance Transformation Imaging* (RTI) è un innovativo metodo di esplorazione degli oggetti attraverso diverse angolazioni di illuminazione. Questa tecnica è particolarmente efficace per l'analisi di superfici prevalentemente piane con dettagli e variazioni localizzate. Il *viewer* offre un'esperienza interattiva che simula un effetto tridimensionale, consentendo di modificare facilmente la direzione e l'intensità della luce per esaltare diversi dettagli della superficie analizzata. A differenza delle tecniche di acquisizione tridimensionale, l'RTI genera immagini bidimensionali "aumentate", capaci di restituire dettagli invisibili a occhio nudo grazie alla variazione dell'angolo di incidenza della luce. Questa metodologia si basa su un set di immagini scattate con diverse condizioni di illuminazione, che vengono poi elaborate per creare una rappresentazione digitale interattiva. Il visualizzatore dedicato permette di modificare la direzione e l'intensità della luce in tempo reale, mettendo in risalto dettagli superficiali, incisioni e irregolarità del materiale analizzato.

Un aspetto distintivo dell'RTI è la possibilità di combinare diverse lunghezze d'onda – luce visibile, ultravioletta (UV) e infrarossa (IR) – offrendo così un'analisi multispettrale particolarmente efficace per lo studio di manufatti con stratificazioni pittoriche, iscrizioni o alterazioni superficiali. Questa tecnologia si dimostra infatti particolarmente adatta per la documentazione e la conservazione di oggetti con superfici prevalentemente piatte, come monete, tavolette incise, dipinti o manoscritti. Grazie alla sua capacità di esaltare dettagli spesso impercettibili, l'RTI si sta affermando come uno strumento essenziale per la ricerca e la valorizzazione del patrimonio culturale.

In Italia, l'RTI è stato impiegato con successo nello studio di reperti archeologici, epigrafi e opere d'arte, diventando uno strumento essenziale per la conservazione e valorizzazione del patrimonio culturale. Il *Visual Computing Lab* del CNR di Pisa ha sviluppato metodologie avanzate per l'acquisizione e l'analisi RTI, con applicazioni su manufatti di varie grandezze (Corsini, Dellepiane, Callieri, & Scopigno, 2006). Un contributo innovativo è quello di Cosentino, che ha combinato l'RTI con tecniche di macro fotografia multispettrale, migliorando la capacità di analisi di superfici complesse come manoscritti, dipinti e reperti archeologici (Cosentino, 2016). Questo approccio ha permesso di rivelare dettagli invisibili a occhio nudo, ampliando le possibilità di studio e documentazione digitale.

A livello internazionale, il *Cultural Heritage Imaging* (CHI) negli Stati Uniti d'America ha svolto un ruolo pionieristico nello sviluppo di software open-source per l'elaborazione RTI, promuovendone l'uso in musei e istituzioni accademiche (Malzbender, Gelb, & Wolters, 2001). Il *British Museum* e il *Victoria & Albert Museum* hanno applicato questa tecnologia per l'analisi di iscrizioni antiche e superfici decorate (Mudge, Malzbender, Schroer, & Lum, 2008), mentre l'Università di Oxford ha sviluppato progetti RTI in ambito epigrafico e papirologico (Harris, 2015). Il *Getty Conservation Institute* ha sperimentato il metodo nella conservazione di dipinti e superfici storiche, dimostrando l'efficacia dell'RTI nella rilevazione di dettagli nascosti (Earl, Basford, Bischoff, Bowman, Crowther, Dahl, Hodgson, Martinez, Isaksen, Pagi, Piquette, & Kotoula, 2011).

[1] Si veda la pietra runica U240 a Lingsberg in Svezia (https://www.runesdb.de/find/6306). Tra le numerose pietre runiche che riportano il motivo del *lindworm* (serpente, dragone), è quella che mostra più somiglianza con l'iscrizione runica del fianco destro del Leone. Questo motivo iconografico è molto frequente nelle iscrizioni scandinave, soprattutto in quelle svedesi.

Un riferimento chiave nella ricerca RTI è il volume "Multi-light Imaging: Highlight-Reflectance Transformation Imaging (H-RTI) for Cultural Heritage" (Duffy, Kennedy, Goskar, & Backhouse, 2018), che esplora le applicazioni avanzate dell'RTI e le sue potenzialità nel campo dei beni culturali. Questo lavoro contribuisce a consolidare il metodo come uno strumento fondamentale per la documentazione e l'analisi delle superfici, in particolare quando combinato con tecniche multispettrali e intelligenza artificiale per la classificazione automatizzata dei dati.
Nel contesto della documentazione e conservazione di manufatti culturali, l'uso del 3D *scanning* e del *Reflectance Transformation Imaging* offre risultati complementari. Entrambi i metodi offrono infatti strumenti avanzati per lo studio degli oggetti, ma con approcci e risultati differenti. Il 3D *scanning* è uno strumento potente per acquisire la forma tridimensionale di un oggetto, permettendo di ricostruirlo virtualmente e analizzarne il volume e le proporzioni. Questa tecnologia è particolarmente utile per monitorare eventuali deformazioni nel tempo, per riprodurre copie digitali di reperti e per integrarsi con strumenti di realtà aumentata o virtuale. Tuttavia, la capacità di catturare i dettagli superficiali più fini può risultare limitata, specialmente quando si tratta di incisioni molto leggere o testi scoloriti.
L'RTI, invece, non crea un modello tridimensionale dell'oggetto, ma lavora sulla variazione della luce per esaltare dettagli superficiali impercettibili. Questo lo rende estremamente efficace per lo studio di iscrizioni consumate, incisioni sottili e tracce di lavorazione che altrimenti potrebbero sfuggire all'occhio umano o a un semplice scanner 3D. La possibilità di modificare dinamicamente l'angolazione della luce permette agli studiosi di esaminare ogni minima variazione della superficie, rendendo questa tecnologia ideale per il monitoraggio del degrado di dipinti, monete, tavolette e altri manufatti con scritture o decorazioni in rilievo. Un altro aspetto rilevante riguarda la documentazione dei testi antichi. Se un'iscrizione è incisa profondamente o in rilievo, il 3D scanning può fornire una rappresentazione chiara e fedele. Tuttavia, quando il testo è poco profondo o eroso dal tempo, l'RTI si dimostra una soluzione più efficace, perché sfrutta la variazione dell'illuminazione per far emergere dettagli altrimenti invisibili. Questo è particolarmente utile nel caso di iscrizioni su monete, epigrafi lapidee e tavolette in argilla, dove la leggibilità è spesso compromessa dall'usura.
Lungi dall'essere alternative, queste due tecnologie sono fortemente complementari. Il 3D *scanning* permette di preservare la forma e il volume di un oggetto, mentre l'RTI consente di esplorare la sua superficie con un livello di dettaglio straordinario. Integrarle significa poter monitorare il deterioramento nel tempo in modo più accurato e ottenere una documentazione digitale più ricca, offrendo così nuove opportunità per la ricerca, la conservazione e la valorizzazione del patrimonio culturale. In riferimento al presente progetto, entrambe le tecnologie verranno applicate e messe a confronto durante la fase di analisi dei dati, per i motivi spiegati sopra.

Il Leone del Pireo

La statua marmorea greca, originaria di Atene e datata al IV secolo, era stata progettata per essere posta a guardia del porto (in greco antico, Πειραιεύς, "pireo") della stessa città, un crocevia di culture e lingue. Le iscrizioni runiche sono invece datate attorno l'XI secolo e sono opera di mercenari di origine vichinga, più precisamente Variaghi, che in quel periodo combattevano in nome dell'Impero bizantino. Il contenuto delle iscrizioni è legato alla volontà dell'incisore di fissare gli eventi straordinari affrontati in guerra, i compagni persi in battaglia, il compenso in denaro non riscosso a peritura memoria. Nel 1687, il Leone lasciò Atene per volontà di Francesco Morosini, che divenne successivamente Doge, a seguito di una vittoria di Venezia sugli Ottomani. Il Leone del Pireo oggi a guardia dell'Arsenale di Venezia risulta essere, di fatto, un oggetto culturale complesso, in quanto formato da più componenti: quella filologica e linguistica (iscrizioni e immagini), quella archeologica (statua unica nel suo genere), e spazio-temporale, che vede intrecciate le storie di tre paesi (Grecia, Scandinavia, Italia). Dal suo arrivo in Italia è stato oggetto di indagine da parte di studiosi provenienti da vari paesi in un arco temporale che si estende dalla fine del XVIII fino al XXI secolo: una panoramica comprensiva degli studi condotti fino al 2000 è disponibile in (Simone, 2001). Segue poi il lavoro di (Snædal, 2014), contenente un'interpretazione aggiornata delle iscrizioni. Le conclusioni tratte dalla runologa svedese lanciano un ponte verso ricerche che nel campo della runologia, e più in generale nell'epigrafia, si possono condurre grazie all'applicazione di avanzate tecnologie digitali, quali, per esempio, il RTI. L'applicazione di questa tecnologia richiede contesti normativi e ambientali specifici: deve essere innanzitutto richiesto il nulla osta da parte degli enti responsabili della tutela e conservazione dei beni culturali (Soprintendenza, Marina Militare); soddisfatto il primo requisito, il contesto fisico in cui è sito l'oggetto di studio deve essere predisposto con l'attrezzatura necessaria per poter procedere (vd. Fig. 1), aspetto che richiede l'intervento di manodopera specializzata.



Figura 1. Visuale dall'alto della statua del Leone del Pireo a sinistra del cancello d'ingresso all'Arsenale. La statua è recintata e confina sul lato destro il cancello di ingresso e sul lato posteriore con un canale. Foto scattata con drone DJI FC3582 (Novembre 2023, Peratello) e su concessione della Marina Militare di Venezia.

3. AMBITO NORMATIVO DI RIFERIMENTO

Il presente progetto rientra nell'attività di ricerca "Patrimonio culturale digitalizzato e narrazione partecipata" (supervisor: Marina Buzzoni) all'interno dello Spoke 9 - CREST (*Cultural Resources for Sustainable Tourism*), coordinato dall'Università Ca' Foscari Venezia, una delle linee operative del Progetto PNRR (2022-2025) CHANGES (*Cultural Heritage Active Innovation for Sustainable Society*). Lo Spoke 9 intende proporre delle soluzioni di gestione, conservazione e valorizzazione dei beni culturali, siano essi tangibili e intangibili, che considerino il cambiamento climatico e che includano l'integrazione e la partecipazione delle comunità locali. Il progetto qui presentato si muove infatti secondo queste direttive e un duplice quadro normativo di riferimento: a livello nazionale, secondo il "Codice dei Beni Culturali e del Paesaggio", tra i primi gli artt. 3(1) e 6(1); le "Linee Guida per la digitalizzazione del patrimonio culturale - AgID" (Agenzia per l'Italia Digitale), in particolare, gli artt. 2.1 "Perché: scopi e obiettivi della digitalizzazione" e 2.1.3 "Recupero di campagne di digitalizzazione pregresse";[2] a livello europeo, secondo la "Convenzione Europea di Faro", nell'art. 5 "Leggi e politiche sull'eredità culturale", comma (b) "mettere in luce il valore dell'eredità culturale attraverso la sua identificazione, studio, interpretazione, protezione, conservazione e presentazione"; nell'art. 12 "Accesso all'eredità culturale e partecipazione democratica" comma (d) "promuovere azioni per migliorare l'accesso all'eredità culturale, in particolare per i giovani e le persone svantaggiate, al fine di aumentare la consapevolezza sul suo valore, sulla necessità di conservarlo e preservarlo e sui benefici che ne possono derivare", e nell'art. 14 "Eredità culturale e società dell'informazione" comma (d) "[..] creazione di contenuti digitali relativi all'eredità culturale non dovrebbe pregiudicare la conservazione dell'eredità culturale attuale." Si evince quindi il tipo di impatto che le analisi possono avere sia a livello scientifico sulla base della tipologia di dati raccolti, della metodologia seguita, della rilevanza dei risultati, sia a livello di *public engagement*, per la sensibilizzazione dei cittadini ad oggetti culturali meno noti e, spesso, inaccessibili, grazie anche ai *digital twin* che facilitano la veicolazione del valore del patrimonio.

4. METODOLOGIA

L'applicazione della metodologia RTI produce una fotografia interattiva di un oggetto (Vanweddingen, Vastenhoud, Proesmans, Hameeuw, Vandermeulen, Van der Perre, Lemmers, Watteeuw, & Van Gool,

[2] Nell'art. 2.1 si specifica che progetti di digitalizzazione possano essere condotti per 1) conservare gli originali, 2) fruire e valorizzare i beni culturali, 3) studiare il patrimonio culturale, 4) recuperare di campagne di digitalizzazione pregresse, come viene suggerito anche da (Potter, Rönnlund & Wallensten, 2023).

2018). Le immagini si ottengono utilizzando un *light dome* al quale è fissata al centro una macchina fotografica professionale:[3] nello specifico, il dome RTI è disponibile presso il *Venice Centre for Digital and Public Humanities* (VeDPH) ed è collegato ad una macchina fotografica personalizzata Sony Alfa III, che cattura immagini con luce ultravioletta (UV, 950 nm), infrarossa (IR, 360 nm) e visibile con il medesimo *set up*.[4] Le analisi multispettrali, che permette di condurre il dome, sono applicabili ad una vasta gamma di materiali che comprende anche il marmo.
L'iscrizione sul fianco destro del Leone è la più complessa ma anche la più interessante da analizzare: è incisa all'interno di due linee che corrono parallele a formare il corpo di un dragone (*lindworm*). La macchina fotografica incorporata al dome scatta immagini di 20 cm, pertanto l'iscrizione può essere suddivisa come segue:

- iscrizione lato destro: 4 foto; altezza da terra al punto più alto 250 cm, AB 75 cm, CD 90 cm, BC 105 cm, DA 85 cm (vd. Fig. 2).

Ugualmente, l'area all'interno della quale le altre due iscrizioni (fianco sinistro e zampa posteriore sinistra) si sviluppano è così suddivisa:

- iscrizione lato alto sinistro (fianco e zampa anteriore sinistri): 10 foto ca.; altezza da terra 230 cm, lunghezza e larghezza dell'iscrizione: 70*7,5 cm
- iscrizione lato basso sinistro (zampa posteriore sinistra): 1 foto; 150cm di altezza da terra, lunghezza e larghezza dell'iscrizione: 90/100*90cm

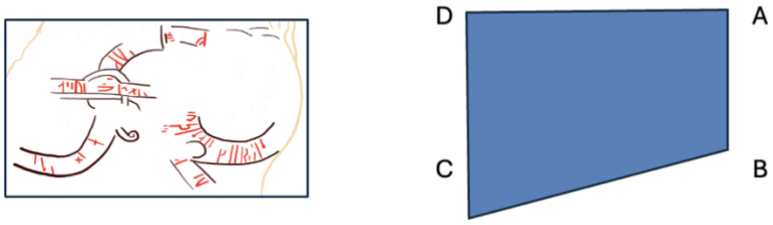


Figura 2. Da sinistra: la lettura del disegno e del testo dell'iscrizione sul fianco destro del Leone eseguita da Peratello; a destra: la ricostruzione dell'area in cui l'iscrizione è racchiusa.

5. RISULTATI ATTESI

I risultati attesi sono una maggior leggibilità delle linee ornamentali che corrono parallele e che contengono le rune iscritte, e una maggior leggibilità delle rune almeno dell'iscrizione sul lato destro: questa riporta che un certo Asmundr ha inciso le rune (il nome dell'incisore non è sempre dato in quest'ambito di studio); seguono poi altri nomi propri maschili, quali Æskell e Þorlæfr, probabilmente due figure importanti all'interno del gruppo di mercenari Variaghi. Del resto dell'iscrizione sono individuabili gruppi di rune o singole rune. È in programma l'inclusione delle immagini selezionate dal set delle foto multispettrali scattate col dome nella visualizzazione 3D delle iscrizioni (Peratello, De Bastiani & Fabbris, 2024). Infine, si segnala l'intenzione di estendere, in futuro, l'applicazione della metodologia qui presentata, e quindi delle stesse tecnologie, anche alle altre iscrizioni runiche in Italia.

RINGRAZIAMENTI

Il progetto di digitalizzazione delle iscrizioni del Leone del Pireo è supervisionato da Marina Buzzoni (Università Ca' Foscari Venezia) e l'analisi RTI non sarebbe stata possibile senza la collaborazione con Elisa Corrò (Università Ca' Foscari Venezia, VeDPH) e con tutto il gruppo di lavoro del VeDPH diretto e coordinato da Franz Fischer.

La pubblicazione è stata realizzata con il cofinanziamento dell'Unione europea – Next Generation EU – Codice identificativo Progetto PE_00000020 titolo "CHANGES - Cultural Heritage Active Innovation for Sustainable Society", CUP H53C22000850006. I punti di vista e le opinioni

[3] https://www.rti-dome.com/what-is-rti.
[4] Ad ogni scatto effettuato con la macchina fotografica collegata al dome corrispondono 36 immagini da cui si ottiene una rappresentazione bidimensionale del soggetto. I dettagli tecnici sono i seguenti: temperatura di colore 5000 K; 36 x LED bianchi; 36 x LED UV; 36 x LED IR.

espresse sono tuttavia solo quelli degli autori e non riflettono necessariamente quelli dell'Unione europea o della Commissione europea. Né l'Unione Europea né la Commissione Europea possono essere ritenute responsabili per essi.

BIBLIOGRAFIA

Akcay, A. (2015). RTI in Epigraphy: Phaselis Example. Soma2015: 19th Symposium on Mediterranean Archaeology. https://doi.org/10.13140/RG.2.2.14088.62728

Corsini, M., Dellepiane, M., Callieri, M., Scopigno, R. (2006). Reflection Transformation Imaging on Larger Objects: An Alternative Method for Virtual Representations. "British Archeological Reports - 2nd International Conference on Remote Sensing Archaeology".

Cosentino, R. A. (2016). Macro photography and multispectral imaging for cultural heritage documentation. E-Conservation Journal, 71-85. https://doi.org/10.18236/econs1.201310

Council of Europe Framework Convention on the Value of Cultural Heritage for Society (CETS No. 199), Faro 27/10/2005. https://www.coe.int/en/web/conventions/full-list?module=treaty-detail&treatynum=199

DECRETO LEGISLATIVO 22 gennaio 2004, n. 42. Codice dei beni culturali e del paesaggio, ai sensi dell'articolo 10 della legge 6 luglio 2002, n. 137. (Gazzetta Ufficiale n.45 del 24-2-2004 - Suppl. Ordinario n. 28). https://www.gazzettaufficiale.it/sommario/codici/beniCulturali

Duffy, S.M., Kennedy, H., Goskar, T., Backhouse, P. (2018). Multi-light Imaging: Highlight-Reflectance Transformation Imaging (H-RTI) for Cultural Heritage. https://doi.org/10.5284/1110911

Earl, G., Basford, P., Bischoff, A., Bowman, A., Crowther, C., Dahl, J., Hodgson, M., Martinez, K., Isaksen, L., Pagi, H., Piquette, K., & Kotoula, E. (2011). Reflectance Transformation Imaging systems for ancient documentary artefacts. "EVA London 2011: Electronic Visualisation and the Arts." Bowen, J.P., Dunn, S., & Ng, K. (Eds), BCS, The Chartered Institute for IT, 147-154. https://doi.org/10.14236/ewic/EVA2011.27

Harris, S. (2015). Reflectance Transformation Imaging (RTI) for visualising leather grain surface morphology as an aid to species identification: a pilot study, vol. 42.

Lech, P., Matera, M., & Zakrzewski, P. (2021). Using Reflectance Transformation Imaging (RTI) to document ancient amphora stamps from Tanais, Russia. Reflections on first approach to their digitalisation. Journal of Archaeological Science: Reports, 33. https://doi.org/10.1016/j.jasrep.2021.102839

Linee guida AgID per la digitalizzazione del patrimonio culturale. (2022). Piano nazionale di digitalizzazione del patrimonio culturale, versione 1.0. https://docs.italia.it/italia/icdp/icdp-pnd-digitalizzazione-docs/it/v1.0-giugno-2022/index.html

Malzbender, T., Gelb, D., & Wolters, H. (2001). Polynomial texture maps. "Proceedings of the 28th annual conference on Computer graphics and interactive techniques (SIGGRAPH '01)." Association for Computing Machinery, New York, 519–528. https://doi.org/10.1145/383259.383320

Mudge, M., Malzbender, T., Schroer, C., & Lum, M. (2008). New reflection transformation imaging methods for rock art and multiple-viewpoint display. "VAST06: The 7th International Symposium on Virtual Reality, Archaeology and Intelligent Cultural Heritage". Journal of Archaeological Science, 93. https://doi.org/10.2312/VAST/VAST06/195-202

Peratello, P., De Bastiani, & C. Fabbris, G. (2024). Reinventing approaches to accessibility within the Digital Paradigm: some Italian case studies. DH2024 "Reinvention & Responsibility" Book of Abstracts. Karajgikar, J., Janco, A., Otis, J. (Eds), Zenodo, 124-127. https://zenodo.org/records/13761079

Ponchio, F., Lamé, M., Scopigno, R., & Robertson, B. (2018). Visualizing and Transcribing Complex Writings through RTI. 2018 IEEE International Congress on Information Science and Technology (CiSt), Marrakech, Marocco, 227-231. https://doi.org/10.1109/CIST.2018.8596602

Potter, R., Rönnlund, R., & Wallensten, J. (2023). An Evaluation of Substance Painter and Mari as visualisation methods using the Piraeus Lion and its runic inscriptions as a case study. Heritage Science, 11, 226. https://doi.org/10.1186/s40494-023-01071-7

RTI-Dome, https://www.rti-dome.com/what-is-rti

Schwab, U. (2009). Le rune in Italia. Alessandria: Edizioni dell'Orso.

Simone, G. (2001). Il Leone del Pireo. Testo e immagine nel Medioevo germanico. Saibene, M. Grazia & Buzzoni, M. (Eds), Milano, Cisalpino, 204-211.

Snædal, T. (2016). Runes from Byzantium: reconsidering the Piraeus Lion. Byzantium and the Viking World, 187-244.

Spoke 9 – CREST, Progetto PNRR (2021-2026) CHANGES. https://pric.unive.it/progetti/spoke-9-changes/home

U240, Lingsberg. https://www.runesdb.de/find/6306

Vanweddingen, V., Vastenhoud C., Proesmans, M., Hameeuw, H., Vandermeulen, B., Van der Perre, A., Lemmers, F., Watteeuw, L., & Van Gool, L. (2018). A Status Quaestionis and Future Solutions for Using Multi-light Reflectance Imaging Approaches for Preserving Cultural Heritage Artifacts. Digital Heritage. Progress in Cultural Heritage: Documentation, Preservation, and Protection. EuroMed 2018. Ioannides, M. et al. (Eds), Springer, Cham., 204-211. https://doi.org/10.1007/978-3-030-01765-1_23

Torino anni Ottanta. Digitalizzazione del patrimonio documentario e ricostruzione virtuale delle mostre negli spazi pubblici e privati.

Filippo Yahia Masri
[1] Università di Genova, Italia – filippoyahia.masri@edu.unige.it

ABSTRACT (ITALIANO)

Questo contributo si propone di definire il panorama storico-artistico della Torino degli anni Ottanta e di ricostruire digitalmente le principali mostre di arte contemporanea del periodo. Il lavoro si concentra sullo studio di due nodi fondamentali: la chiusura della Galleria Civica di Arte Moderna (GAM) e le attività espositive delle altre istituzioni. Per la GAM, sono state analizzate le acquisizioni del museo operate durante il decennio e la presenza attuale di opere degli anni Ottanta nelle collezioni. Il progetto prevede di creare un catalogo digitale delle opere prese in esame, analizzando un patrimonio utile per comprendere le dinamiche del periodo. L'attenzione verrà inoltre posta sull'attività del Castello di Rivoli e sulle mostre di cinque gallerie protagoniste del decennio. La ricostruzione virtuale delle mostre consentirà una mappatura dei movimenti di galleristi, artisti e collezionisti attivi nel capoluogo piemontese, digitalizzando fonti primarie e secondarie ad oggi inedite o comunque di difficile reperibilità. Il sistema artistico della Torino degli anni Ottanta verrà ricostruito digitalmente, archiviando le fonti raccolte su database dedicato e creando ambienti digitali, permettendo la fruibilità di un patrimonio ad oggi ancora insondato. Mediante l'utilizzo di ReCap Pro di Autodesk, sarà possibile processare la nuvola di punti generata da un rilievo fotografico così da ottenere una triangolazione fotogrammetrica che consenta di ricostruire il modello tridimensionale di spazi e allestimenti artistici. Il progetto permetterà di ricreare mostre ancora non analizzate, consentendo la fruibilità e l'accessibilità di materiale inesplorato, digitalizzando spazi, allestimenti e opere. Queste ultime, opportunamente elaborate tramite SolidWorks, verranno ricostruite virtualmente e, nel caso di performance, anche animate. Grazie a UNITY, le mostre potranno essere esplorate in realtà aumentata – con visori –, per un'esperienza immersiva.
Parole chiave: digitalizzazione; ricostruzione virtuale; realtà aumentata.

ABSTRACT (ENGLISH)

Turin in the eighties. Digitization of documentary heritage and virtual reconstruction of exhibitions in public and private spaces.

This contribution aims to define the historical-artistic panorama of Turin in the eighties and to reconstruct digitally the main contemporary art exhibitions of the period. The work focuses on two fundamental issues: the closure of the Galleria Civica di Arte Moderna (GAM) and the exhibition activities of other institutions. For the GAM, the museum's acquisitions during the decade and the current presence of works from the eighties in the collections were analysed. The project aims to create a digital catalogue of the works under consideration, analysing a heritage useful for understanding the dynamics of the period. The focus will also be on the activity of the Castello di Rivoli and on the exhibitions of five galleries protagonists of the decade. The virtual reconstruction of the exhibitions will allow a mapping of the movements of gallerists, artists and collectors active in the capital of Piedmont, digitizing primary and secondary sources currently unpublished or difficult to find. The artistic system of Turin in the eighties will be digitally reconstructed, storing the sources collected on a dedicated database and creating digital environments, allowing the use of a heritage still unexplored today. Using Autodesk's ReCap Pro, it will be possible to process the point cloud generated by a photographic relief in order to obtain a photogrammetric triangulation that allows the three-dimensional model of spaces and artistic arrangements to be reconstructed. The project will allow to recreate exhibitions not yet analyzed, allowing the usability and accessibility of unexplored material, digitizing spaces, settings and works. These are processed by SolidWorks and will be reconstructed virtually, or even animated in the case of performances. Thanks to UNITY, the exhibitions can be explored in augmented reality - with visors -, for an immersive experience.
Keywords: digitisation; virtual reconstruction; augmented reality.

1. STATO DELL'ARTE

La GAM, le gallerie private e l'Arte Povera sono state protagoniste dell'importante successo artistico di Torino nel periodo tra l'inizio degli anni Sessanta e la fine degli anni Settanta. La lettura del momento subito successivo risulta però complessa e anche gli studi più recenti ne hanno trattato solo in maniera collaterale. Il contributo permetterà di evidenziare non solo i principali avvenimenti, ma anche premesse e ripercussioni delle attività artistiche nel capoluogo piemontese durante il decennio preso in esame.
La chiusura della GAM, avvenuta tra il 1981 e il 1993, rappresenta un caso significativo per la lettura del periodo nel contesto cittadino. La corposa bibliografia riguardante la storia del museo ha spesso trascurato, se non in rare eccezioni (Bertolino, Christov-Bakargiev & Passoni, 2016), il dinamismo di tale istituzione negli anni Ottanta. Seppur precluso ai più, il museo è stato comunque attivo attraverso importanti acquisizioni, operate grazie all'intervento della Fondazione De Fornaris e della Fondazione CRT e documentate in maniera parziale (Villa, 2003; Castagnoli & Grana, 2008; Castagnoli & Volpato, 2008). La difficoltà dello studio nasce dall'assenza di fonti di prima mano che testimonino questi movimenti, avvenuti sottotraccia. Gli studi sul ruolo del museo dalla fondazione sino agli anni Ottanta, la letteratura sulla GAM dal momento della riapertura e i documenti d'archivio presenti nella Biblioteca d'Arte della Fondazione Torino Musei verranno utilizzati per comprendere il ruolo culturale e sociale del museo nel momento della sua assenza e in quello subito successivo. Non esiste ad oggi uno studio specifico sul tema, nonostante diversi approfondimenti sui decenni precedenti. Il periodo di chiusura della GAM costituisce un caso di studio prezioso per comprendere l'importanza del dibattito sulla museologia sorto negli anni Ottanta: un esempio di variazione netta tra chiusura e riapertura, e non di graduale adattamento alle nuove ambizioni culturali come successo per gli altri musei. In questo senso risultano fondamentali gli studi sugli allestimenti fino al 1981 e gli approfondimenti sull'inaugurazione del museo rigenerato nel 1993 (Maggio Serra, 1993: Maggio Serra & Passoni, 1993). Per la ricerca riguardo l'arricchimento delle collezioni avvenuto durante gli anni Ottanta e l'analisi delle opere degli anni Ottanta presenti negli spazi della GAM, sono imprescindibili i cataloghi degli acquisti e delle collezioni del museo. Il sito del museo non prevede ad oggi una digitalizzazione completa delle proprie opere, ma rende comunque accessibile agli studiosi diversi documenti presenti nella Biblioteca d'Arte della Fondazione Torino Musei.
Il contributo si concentra inoltre sulle mostre di quelle che sono state individuate come le gallerie più rappresentative del periodo: la Galleria Stein, la Galleria Tucci Russo, la Galleria Persano, la Galleria Guido Carbone e la Galleria In Arco. Tutte attive durante il decennio, hanno tuttavia data di apertura e di chiusura diverse. Mentre la Galleria Stein era già affermata negli anni Sessanta, la Tucci Russo e la Persano iniziano la propria attività negli anni Settanta, consolidando il proprio lavoro negli anni Ottanta. Le gallerie Guido Carbone e In Arco, invece, nascono rispettivamente nel 1985 e nel 1987. Gli spazi privati citati risultano ancora attivi e dispongono di archivi ricchi di documenti, oltre che di siti web in alcuni casi parzialmente aggiornati anche sulle mostre degli anni Ottanta (Viale, 2008). L'assenza di approfondimenti riguardo il decennio non risparmia però la storia delle gallerie. Dunque, data l'assenza di studi aggiornati, la ricerca si concentra sul materiale bibliografico coevo. Nello specifico, i cataloghi e il materiale editoriale delle mostre organizzate nel decennio saranno la base di partenza per il successivo studio comparato sull'attività dalle gallerie. Per comprendere la rilevanza e l'impatto delle mostre, è stato necessario lo spoglio delle principali riviste di settore[1].
Il progetto rende fruibile anche materiale riguardo gli artisti attivi nel periodo, ad oggi non ancora oggetto di studi, come Bruno Zanichelli, Maurizio Vetrugno, Raffaello Ferrazzi, Pierluigi Pusole ma anche esponenti della Transavanguardia, i cui lavori iniziano timidamente ad essere esposti nelle mostre torinesi, come Mimmo Paladino e Nicola De Maria. Questi artisti cercarono di imporre il proprio lavoro in un contesto dominato dall'ormai consolidata Arte Povera. Il peso di quest'ultima ha oscurato l'attività di artisti che provarono a trovare un proprio spazio nel capoluogo piemontese. Per questo motivo, si attesta l'assenza di materiale bibliografico aggiornato: i cataloghi delle mostre e i pochi cataloghi ragionati permetteranno comunque un primo approfondimento del tema.
Risulta invece utile e completo il materiale riguardante la storia e le collezioni del Castello di Rivoli (Beccaria & Gianelli, 2005), che servirà per confrontare i movimenti delle istituzioni private e quelli dell'istituzione pubblica, nata nel 1984, con progettualità ed ambizioni diversi rispetto a quelli della GAM (Bertolino, Christov-Bakargiev & Passoni, 2016).

[1] Le principali riviste di riferimento saranno "Bolaffiarte", "DATA", "DOMUS", "Flash Art", "Il Giornale dell'Arte", "NAC" e "Segno".

Per la parte di digitalizzazione, non esiste ad oggi una piattaforma che renda disponibili i documenti necessari per l'analisi e lo studio del periodo preso in esame.

2. OBIETTIVI DEL PROGETTO

Il contributo prevede una ricostruzione quanto più accurata del contesto di sviluppo delle esperienze trattate e del dibattito critico sorto attorno ad esse. Tali informazioni sono utilizzate per una corretta lettura degli eventi artistici più significativi del panorama torinese negli anni Ottanta. Durante questa fase, ci si concentrerà sul ruolo della GAM nel tessuto culturale della città, approfondendone il caso singolare di chiusura. Le problematiche allestitive nel momento della riapertura e le scelte riguardanti le variazioni strutturali del museo, operate durante la ristrutturazione, rappresenteranno il primo approfondimento. Si procederà poi all'analisi delle principali mostre organizzate dalle gallerie private, approfondendo la presenza di artisti già affermati e le nuove proposte. La ricerca si concentra sul materiale presente nelle principali biblioteche torinesi, tra cui spicca per ovvia importanza la Biblioteca d'Arte della Fondazione Torino Musei, e negli archivi di gallerie e artisti attivi in quegli anni: tra i più fruibili e completi, sicuramente l'archivio Giulio e Anna Paolini, il cui materiale è stato digitalizzato in maniera puntuale. Questi documenti saranno riferimento per ogni fase di studio e primo nucleo di raccolta per il materiale archiviato nel database dedicato. Lo studio si orienta poi sul materiale editoriale relativo alle mostre delle gallerie private e del Castello di Rivoli, portando avanti la ricerca all'interno degli archivi di gallerie e artisti attivi nel periodo. Fonti primarie e secondarie verranno anch'esse archiviate digitalmente (Spallone, Bertola & Ronco, 2019).
Attraverso i cataloghi e la ricerca d'archivio, verrà ordinato digitalmente il corpus delle opere acquisite dalla GAM durante la chiusura dello spazio e quello delle opere degli anni Ottanta presenti nelle attuali collezioni del museo, cercando, in entrambi i casi, di comprendere le motivazioni delle scelte operate. Sul fronte delle gallerie private e del Castello di Rivoli, verranno ricostruiti digitalmente gli spazi e gli allestimenti delle mostre, rendendo fruibili attraverso realtà virtuale le opere esposte. La ricostruzione del panorama storico-artistico negli spazi pubblici e privati verrà realizzato attraverso l'utilizzo della fotogrammetria e della ricostruzione virtuale, sfruttando planimetrie e materiale fotografico riguardo gli allestimenti (Lo Turco, 2019; Luigini & Panciroli, 2018). In particolare, mediante l'utilizzo di un software quale ad esempio ReCap Pro di Autodesk, sarà possibile processare la nuvola di punti generata da un rilievo fotografico – ottenuto ad esempio tramite un laser scanner o, più semplicemente, mediante un supporto dotato di fotocamera – in modo tale da ottenere una triangolazione fotogrammetrica che consenta di ricostruire il modello tridimensionale di spazi e allestimenti artistici. L'utilizzo di un software facente parte dell'ambiente Autodesk permette una facile interoperabilità con altri software in ambiente CAD o BIM. Il progetto permetterà di rendere ricostruibili luoghi e mostre ancora non analizzate, consentendo di rendere fruibile e accessibile materiale ancora inesplorato, digitalizzando spazi, allestimenti e opere. La necessità della ricostruzione digitale risulta imprescindibile sia per la fruibilità di mostre ambientate in spazi ancora visitabili, ma anche per la creazione di modelli relativi alle sedi precedenti delle gallerie. La stessa tecnologia verrà applicata alle opere esposte, rilevate in prima persona o attraverso il materiale fotografico e audiovisivo raccolto durante le ricerche. Le opere bidimensionali e tridimensionali, opportunamente elaborate e, se necessario, modificate tramite software come SolidWorks, verranno ricostruite virtualmente e, nel caso di azioni performative, anche animate (Böhler & Marbs, 2004;Jasink & Dionisio, 2016). Tramite il software UNITY, le mostre potranno essere esplorate in realtà aumentata – grazie all'uso di un visore –, per un'esperienza immersiva e sensoriale (Pecchinenda, 2003; Antinucci, 2004; Antinucci, 2007; Verdiani, 2017)[2].
L'ultima fase del progetto tratterà le principali ricadute dell'assenza del museo sul panorama cittadino durante gli anni Ottanta e il ruolo sociale, didattico e ludico della GAM dopo la riapertura del 1993. Il materiale sulle mostre delle gallerie private e del Castello di Rivoli permetterà di elaborare uno studio riguardo il clima culturale della città, evidenziando manifestazioni e artisti dimenticati o comunque ancora trascurati. Verrà inoltre reso accessibile a tutti il sito creato durante la ricerca, con l'ambizione di creare uno standard di riferimento per successive ricerche – anche su ambiti cronologici e geografici diversi – e la possibilità di renderlo implementabile. Il lavoro sviluppato digitalmente verrà reso incorporabile per le istituzioni pubbliche e private oggetto della ricerca.

[2] UNITY risulta ad oggi la soluzione migliore per gli scopi del progetto e vanta diversi utilizzi nei principali musei internazionali.

3. PROSPETTIVE

- Una ricostruzione accurata del contesto di sviluppo delle esperienze storico artistiche nella Torino degli anni Ottanta e del dibattito critico che le ha circondate. Ciò permetterà una comprensione più approfondita del panorama artistico dell'epoca, ad oggi ancora insondato.
- Un'analisi dettagliata del ruolo della GAM nel tessuto culturale di Torino, con particolare attenzione alla sua chiusura e alle scelte operate durante ristrutturazione e la riapertura.
- La creazione del catalogo delle acquisizioni della GAM operate durante gli anni di chiusura e quello delle opere degli anni Ottanta nelle attuali collezioni del museo.
- La ricostruzione dell'attività del Castello di Rivoli e delle principali gallerie del periodo: La Galleria Stein, la Galleria Persano, la Galleria Tucci Russo, la Galleria Guido Carbone e la Galleria In Arco. L'analisi della storia degli spazi allestitivi, insieme alla raccolta del materiale relativo ad essi, inclusi cataloghi, interviste e articoli, permette di creare una raccolta di fonti primarie e secondarie riguardanti i movimenti artistici del periodo.
- L'identificazione e la catalogazione delle principali mostre della Torino degli anni Ottanta, tenutesi presso il Castello di Rivoli e le gallerie oggetto della ricerca, con la ricostruzione di allestimenti, cataloghi e del dibattito critico sorto attorno ad essi. Il progetto permette di comprendere le dinamiche di un periodo ancora insondato e rilevante per i successivi sviluppi artistici.
- La creazione di un database dedicato, nel quale verrà inserito il materiale raccolto durante la ricerca, proveniente dalle biblioteche, dagli archivi delle gallerie, dalle interviste ai protagonisti del panorama artistico e da altre fonti primarie e secondarie. Questo database rappresenterà un prezioso strumento per la ricerca e la consultazione futura, consentendo la fruibilità di materiale inedito e di difficile reperibilità.
- Archiviazione digitale su database delle opere della GAM studiate durante la ricerca e del materiale relativo alle mostre degli spazi pubblici e privati oggetto dello studio. Questa soluzione permetterà la mappatura delle principali esperienze artistiche della Torino degli anni Ottanta.
- La costruzione di modelli tridimensionali virtuali delle mostre torinesi degli anni Ottanta, con la ricostruzione digitale degli spazi espositivi, degli allestimenti e delle opere d'arte utilizzando l'uso di fotogrammetria e software dedicati. Questi modelli consentiranno una fruizione virtuale delle mostre, rendendo accessibile materiale ancora inesplorato e offrendo un'esperienza immersiva e sensoriale.
- La creazione di un sito web che renderà accessibili i risultati della ricerca, con l'obiettivo di diventare uno standard di riferimento per ricerche future e di essere implementabile per istituzioni pubbliche e private coinvolte nella ricerca.

BIBLIOGRAFIA

Antinucci, F. (2004). Comunicare nel museo. Roma-Bari: Laterza.

Antinucci, F. (2007). Musei Virtuali. Roma-Bari: Laterza.

Beccaria, M., & Gianelli, I. (2005). Castello di Rivoli: 20 anni d'arte contemporanea. Milano: Skira.

Bertolino, G., Christov-Bakargiev, C., Passoni, R. (2016). Dalle bombe al museo, 1942-1959: La rinascita dell'arte moderna: l'esempio della GAM di Torino. Cinisello Balsamo: Silvana Editoriale.

Böhler, W., Marbs, A. (2004). 3D scanning and photogrammetry for heritage recording: A comparison. 12th International Conference on Geoinformatics: Geospatial Information Research: Bridging the Pacific and Atlantic, University of Gävle, Sweden, 7-9 June.

Castagnoli, P. G., Grana, F. (2008). GAM: La Galleria Civica d'Arte Moderna e Contemporanea. Torino: Allemandi.

Castagnoli, P. G., Volpato, E. (2008). Dieci anni di acquisizioni per la GAM di Torino: 1998-2008. Torino: Allemandi.

Jasink, A. M., Dionisio, G. (2016). MUSINT 2. Nuove esperienze di ricerca e didattica nella museologia interattiva. Firenze: Firenze University Press.

Lo Turco, M. (2019). Digital & Documentation. Digital Strategies for Cultural Heritage. Pavia: Pavia University Press.

Luigini, A., Panciroli, C. (2018). Ambienti digitali per l'educazione all'arte e al patrimonio. Milano: FrancoAngeli s.r.l.

Maggio Serra, R. (1993). Arte moderna a Torino. Opere d'arte e documenti acquisiti per la Galleria Civica d'Arte Moderna e Contemporanea di Torino 1986-1992. Vol. II. Torino: Allemandi.

Maggio Serra, R., Passoni, R. (1993). Il Novecento: Catalogo delle opere esposte. Galleria Civica d'Arte Moderna e Contemporanea. Milano: Fabbri Editori.

Pecchinenda, G. (2003). Videogiochi e cultura della simulazione. Roma-Bari: Laterza.

Spallone, R., Bertola, G., Ronco, F. (2019). SfM and digital modelling for enhancing architectural archives heritage. 2019 IMEKO TC-4 International Conference on Metrology for Archaeology and Cultural Heritage, Firenze, Dicembre.

Verdiani, G. (2017). Retroprogettazione. Metodologie ed esperienze di ricostruzione 3D digitale per il patrimonio costruito. Firenze: Dida Press.

Viale, L. (2008). Questo mondo è fantastico. Vent'anni con Guido Carbone. Milano: Electa.

Villa, G. C. F. (2003). La Galleria d'arte moderna di Torino: Cronaca di un'istituzione. Cinisello Balsamo: Silvana Editoriale.

Linguistica dei corpora e informatica umanistica per la valorizzazione plurilingue del patrimonio culturale: implementazione del progetto UniVOCIttà

Rita Gramellini[1], Valeria Zotti[2]
[1] Alma Mater Studiorum Università di Bologna, Italia - rita.gramellini2@unibo.it
[2] Alma Mater Studiorum Università di Bologna, Italia - valeria.zotti@unibo.it

ABSTRACT (ITALIANO)

La fruizione del patrimonio culturale da parte di un pubblico internazionale rappresenta un'esigenza sempre più rilevante nella società contemporanea. In tale prospettiva si colloca il progetto *UniVOCIttà*, sviluppato dall'unità di ricerca LBC-CeSLiC del Dipartimento di Lingue, Letterature e Culture Moderne dell'Università di Bologna. Il progetto si inserisce tra le ricerche nel campo dell'informatica umanistica e propone una metodologia innovativa volta alla valorizzazione plurilingue del patrimonio artistico e culturale di Bologna attraverso la creazione di un corpus comparabile plurilingue annotato, composto da scritti di viaggio di personaggi illustri, dal Rinascimento all'epoca contemporanea.

In collaborazione con specialisti di *Web Technologies* dell'Università di Bologna, è stato realizzato il prototipo di un'applicazione web mobile destinata al turismo culturale, che consente la consultazione della banca dati plurilingue tramite visualizzazioni sia lineari che dinamiche. Poiché le testimonianze letterarie sono attualmente disponibili solo nella lingua originale, è in corso un'implementazione della app nell'ambito di un progetto di dottorato finalizzato all'integrazione delle traduzioni dei contenuti in lingua italiana e nelle altre lingue del progetto. A tal fine, si prevede di testare un sistema di traduzione automatica neurale (DeepL), e un chatbot generativo di ultima generazione (ChatGPT 4o). Tuttavia, considerati i limiti di questi strumenti di intelligenza artificiale nella traduzione di lessemi e termini relativi alle belle arti, gli *output* verranno sottoposti a una fase di revisione manuale mediante la consultazione di un sotto-corpus comparabile di scritti di viaggio in lingua italiana, attualmente in fase di realizzazione. L'obiettivo finale è di offrire a utenti provenienti da diversi orizzonti linguistici una risorsa digitale che favorisca l'accesso e la valorizzazione del patrimonio artistico e culturale bolognese in una prospettiva plurilingue.

Parole chiave: patrimonio culturale digitale; corpus comparabile; applicazione web; turismo culturale; traduzione automatica

ABSTRACT (ENGLISH)

Corpus Linguistics and Digital Humanities for the Multilingual Valorisation of Cultural Heritage: Implementation of the UniVOCIttà Project. The enjoyment of cultural heritage by an international audience represents an increasingly relevant need in contemporary society. Within this perspective, the *UniVOCIttà* project was developed by the LBC-CeSLiC research unit of the Department of Modern Languages, Literatures and Cultures at the University of Bologna. The project is situated within the field of digital humanities and proposes an innovative methodology aimed at the multilingual valorization of Bologna's artistic and cultural heritage through the creation of an annotated multilingual corpus collecting travel writings by illustrious personalities from the Renaissance to the present day.

In collaboration with Web Technologies specialists from the University of Bologna, a prototype of a mobile web app for cultural tourism was developed. This application enables the consultation of the multilingual textual database through both linear and dynamic visualizations. As the literary testimonies are currently only available in the original languages, an ongoing implementation of the application within a doctoral project intends to integrate translations of all content into Italian and the other project languages. To this end, the project plans to test a neural automatic translator (DeepL) and a state-of-the-art generative chatbot (ChatGPT 4o). However, due to the limitations of these artificial intelligent systems in the translation of lexemes and terms related to the fine arts, the outputs will undergo a manual revision phase through the consultation of a comparable Italian sub-corpus of travel writings, currently under development. The ultimate goal is to provide users from diverse linguistic backgrounds with a digital resource that facilitates the valorization and access to Bologna's artistic and cultural heritage from a multilingual perspective.

Keywords: digital cultural heritage; comparable corpus; web application; heritage tourism; machine translation

1. INTRODUZIONE

Nella società contemporanea, l'ampia fruizione del patrimonio culturale in formato digitale costituisce un'esigenza di crescente rilevanza. In questo contesto, l'informatica umanistica si configura come un ambito promettente, in quanto in grado di offrire strumenti e metodologie efficaci per la realizzazione di contenuti culturali digitali fruibili in formato elettronico (Barbuti, 2018; Münster et al., 2019). Affinché le pratiche tecnologiche non si limitino a una semplice transizione in digitale dei beni culturali (Donati, 2019: 326), risulta imprescindibile adottare approcci orientati alla diffusione e fruizione del patrimonio digitale da parte di un pubblico eterogeneo e internazionale.

Il presente contributo illustra una nuova fase di implementazione del progetto *UniVOCIttà*, condotto nel biennio 2022-2023 dall'unità di ricerca LBC-CeSLiC[1] (Centro di Studi Linguistico-Culturale) del Dipartimento di Lingue, Letterature e Culture moderne (LILEC) dell'Università di Bologna. Il progetto si propone di promuovere la conoscenza e la diffusione internazionale del patrimonio artistico e culturale della città di Bologna e della sua provincia, mediante la creazione di una banca dati multilingue di scritti di viaggio, consultabile attraverso una piattaforma interattiva per il turismo culturale.

Nonostante l'applicazione sia stata ideata come una risorsa plurilingue destinata a un pubblico internazionale, la fruizione dei testi è attualmente circoscritta agli utenti in grado di comprendere le lingue originali dei documenti. L'attuale fase del progetto prevede, pertanto, la traduzione dei segmenti testuali in lingua italiana e nelle altre lingue del progetto, attraverso la sperimentazione di due strumenti di intelligenza artificiale: il traduttore automatico neurale DeepL, e il chatbot generativo di ultima generazione ChatGPT 4o. Considerati i limiti comprovati di tali "sistemi artificiali intelligenti" (Andler, 2023: 23) nella traduzione di lessemi e termini afferenti al dominio delle belle arti, gli *output* verranno sottoposti a una revisione linguistica e concettuale, supportata dalla consultazione di un sotto-corpus comparabile italiano di letteratura di viaggio, attualmente in fase di elaborazione. L'obiettivo finale del progetto è la realizzazione di una risorsa digitale in grado di facilitare la fruizione del patrimonio culturale e artistico bolognese da parte di utenti provenienti da diversi contesti linguistici e culturali.

Dopo aver delineato la genesi del progetto *UniVOCIttà*, il presente contributo ne illustra in primo luogo la metodologia adottata e i risultati finora conseguiti, con un focus specifico sul processo di costruzione del corpus testuale plurilingue, sull'annotazione linguistica dei dati e sullo sviluppo di un'applicazione web mobile destinata al turismo culturale. Nella parte conclusiva, si presentano le prospettive di sviluppo futuro, che includono l'integrazione di traduzioni multilingue dei contenuti dell'app mediante il test di due sistemi di intelligenza artificiale.

2. GENESI DEL PROGETTO

Tra le diverse iniziative di informatica umanistica finalizzate alla diffusione del patrimonio tangibile e intangibile italiano, si distingue il progetto "Lessico dei Beni Culturali"[2] (LBC), avviato nel 2013 da Annick Farina presso l'Università di Firenze. Tale progetto, che attualmente coinvolge diverse università italiane e straniere, nasce dall'esigenza di sviluppare risorse linguistiche digitali plurilingui (banche dati testuali, corpora comparabili, corpora paralleli, lessici) volte a favorire l'accesso internazionale alla conoscenza del patrimonio culturale italiano, iniziando da quello toscano (Farina, 2016; Farina & Martínez, 2020: 8).

Nell'ambito di questo progetto interuniversitario, un gruppo di linguiste del Dipartimento di Lingue, Letterature e Culture moderne (LILEC) dell'Università di Bologna ha costituito nel 2018 l'unità di ricerca LBC-CeSLiC (Centro di Studi Linguistico-Culturali), responsabile della costruzione di un sotto-corpus comparabile in quattro lingue (inglese, francese, spagnolo, russo) dedicato al patrimonio architettonico, artistico, museale e naturalistico della città di Bologna e della regione Emilia-Romagna, denominato BER. Questa risorsa raccoglie testi appartenenti a diverse tipologie testuali (diari di viaggio, romanzi, blog turistici, manuali tecnici, saggi di critica d'arte), i quali vengono suddivisi in tre macrocategorie definite dal progetto madre fiorentino: testi letterari, divulgativi e tecnici (Perotto & Zotti, 2022: 168).

Sulla scia delle attività condotte nell'ambito del progetto LBC, l'unità di ricerca LBC-CeSLiC ha avviato il progetto di disseminazione *UniCittà: Riscoprire Bologna attraverso l'Alma Mater Studiorum*, finanziato tra il 2019 e il 2021 dalla Fondazione Cassa di Risparmio di Bologna all'interno del bando "RiScopriamo la città". *UniCittà* è stato concepito con l'obiettivo di diffondere a un pubblico internazionale la conoscenza del patrimonio immobile (musei, biblioteche, archivi, palazzi storici) e mobile (opere d'arte, collezioni librarie) dell'Università di Bologna mediante un corpus plurilingue digitale di testimonianze presenti e passate che

[1] L'unità di ricerca è composta da Valeria Zotti (referente, francese), Ana Pano Alamán (spagnolo), Monica Turci e Antonella Luporini (inglese), Monica Perotto (russo).

[2] https://www.lessicobeniculturali.net/ (cons. 10/04/2025)

attestano la ricchezza storica e culturale dell'Alma Mater (Zotti, 2023: 199). I testi selezionati sono stati dapprima sottoposti a un processo di annotazione con etichette (*tags*) che identificano otto categorie rappresentative del patrimonio universitario[3]. Successivamente, sono stati resi disponibili tramite una banca dati "annotata", consultabile gratuitamente in versione web[4] e sul repository istituzionale dell'Università di Bologna.[5]
La metodologia e i risultati ottenuti nell'ambito di *UniCittà* hanno posto le basi per un progetto derivato condotto dall'unità di ricerca LBC-CeSLiC, volto alla valorizzazione del potenziale turistico-culturale dell'intero patrimonio bolognese: il progetto *UniVOCIttà*.

3. METODOLOGIA

Il progetto di ricerca *UniVOCIttà: Voci digitali sull'unicità del patrimonio bolognese*, finanziato con fondi NextGenerationEU nell'ambito del bando AlmaIdea 2022 e coordinato da Valeria Zotti, si propone di preservare e promuovere il patrimonio tangibile (luoghi di culto, piazze, mura, torri) e intangibile (parlata, cucina, tradizioni) di Bologna e della sua provincia attraverso un corpus di testimonianze letterarie in lingua originale redatte da personaggi illustri (letterati, artisti, scienziati, politici) che hanno visitato il capoluogo emiliano nel corso dei secoli, il corpus BER-LET.
Come menzionato da Zotti (2025), la selezione dei testi è stata effettuata seguendo i criteri definiti da L'Homme (2004: 126-127) per la costituzione di un corpus specialistico, così articolati:
- Ambito di specialità: patrimonio culturale tangibile e intangibile della città di Bologna e della sua provincia;
- Lingue: inglese, francese, spagnolo, russo;
- Livello di specializzazione: specializzazione media;[6]
- Tipo di documento: testi appartenenti al genere della letteratura di viaggio, quali resoconti, diari di viaggio, corrispondenze epistolari, romanzi;
- Supporto: testi scritti, in formato cartaceo o digitale;
- Periodo di pubblicazione: dal XVII al XXI secolo.

Per garantire una copertura esaustiva delle fonti, la ricerca dei testi è stata principalmente guidata dalle numerose ricerche disponibili sul fenomeno del *Grand Tour*, concentrandosi sugli studi relativi ai viaggiatori che hanno esplorato l'Emilia-Romagna (Zotti, 2025). Tra le opere di riferimento si distinguono l'antologia *Viaggi e viaggiatori del Settecento in Emilia e in Romagna* curata da Giorgio Cusatelli (1986) e i contributi di Attilio Brilli (2005; 2014). Le opere di pubblico dominio sono state reperite dai principali archivi e biblioteche digitali internazionali.[7] Quando disponibili esclusivamente in formato cartaceo, i testi sono stati digitalizzati mediante la tecnologia OCR (Riconoscimento Ottico dei Caratteri) dai membri dell'unità e da un gruppo di tirocinanti.
Al termine del progetto, il corpus risulta composto da oltre 200 testi, di cui circa il 45% redatti in lingua francese, il 35% in inglese, il 15% in spagnolo e il 10% in russo. Nonostante l'intento di garantire una rappresentazione equilibrata delle quattro lingue coinvolte, la distribuzione effettiva dei materiali testuali riflette le diverse tradizioni di viaggio nel corso dei secoli. Non sorprende, ad esempio, la predominanza delle opere di viaggiatori inglesi e francesi, che furono tra i principali protagonisti del *Grand Tour*.
I documenti raccolti sono stati successivamente annotati manualmente con il software di analisi qualitativa Atlas.ti[8] (Muhr, 1991) seguendo un approccio *bottom-up*. L'osservazione empirica dei testi nelle quattro lingue ha consentito di identificare le macrocategorie più rilevanti del patrimonio artistico e culturale dell'area di interesse, come Luoghi del Patrimonio ("LU"); Opere d'arte ("OPA"); Usi e costumi ("USIeCOSTUMI"), e le sotto-aree ("Museo"; "Zona Universitaria"; "Portico", "Pittura"; "Architettura"; ecc.). Complessivamente, sono stati annotati 1860 frammenti testuali per le quattro lingue del progetto.

[3] Artisti ("ART"); Libri ("LIB"); Luoghi ("LU"); Oggetti ("OGG"); Opere d'arte ("OPA"); Periodo storico ("PE"); Personaggi storici ("PERS"); Tematica ("TEM"); Titoli di libri ("TIL").
[4] https://site.unibo.it/unicitta/it (cons. 24/01/2025)
[5] https://amsacta.unibo.it/id/eprint/7166/ (cons. 24/01/2025)
[6] "La pertinenza di un testo a livello alto, medio o basso è determinata dalle caratteristiche del pubblico, del mezzo e dello scopo. Ad esempio, un testo prodotto da uno specialista per gli studenti può essere definito di "livello medio" (Cabré, 2007-2008: 40, trad. mia). Nel caso del corpus BER-LET, i testi di letteratura di viaggio raccolti sono stati scritti da autori colti e destinati a un pubblico di lettori di cultura medio-alta (Zotti, 2025).
[7] British Library, Library of Congress, LexisNexis, Internet Archive, HathiTrust, BNF Gallica, Bibliothèque et Archives nationales du Québec, Persée, Biblioteca russa di stato, Biblioteca presidenziale B. Eltsin, Biblioteca Digital Hispánica della Biblioteca Nacional de España, Biblioteca Virtual Miguel de Cervantes.
[8] https://atlasti.com/ (cons. 24/01/2025)

4. RISULTATI

I frammenti annotati sono stati resi disponibili tramite un'applicazione web mobile multilingue[9] sviluppata in collaborazione con il Dipartimento di Ingegneria e Scienze Informatiche (DISI) e il Dipartimento di Filologia Classica e Italianistica (FICLIT) dell'Università di Bologna, sotto la supervisione di Fabio Vitali, professore ordinario di Informatica ed esperto di *Web Technologies*. Questa piattaforma, disponibile in tutte e quattro le lingue del progetto, intende mettere i testi digitalizzati al servizio del grande pubblico e degli operatori coinvolti nell'industria del turismo, al fine di promuovere Bologna come una meta di turismo culturale esperienziale (Apostolakis, 2003: 799; Zotti, 2023: 385).

L'applicazione si articola in quattro sezioni principali ("Progetto", "Su di noi", "Corpus" e "Mappe") e offre due modalità distinte di esplorazione del corpus letterario. Nella sezione "Corpus", l'utente ha accesso ai frammenti testuali attraverso una visualizzazione lineare, con la possibilità di effettuare ricerche incrociate mediante l'impiego dei metadati etichettati come filtri di ricerca (vd. Fig. 1) (Zotti, 2025). Al fine di garantire una migliore fruizione del patrimonio letterario digitalizzato, ogni citazione è accompagnata da un collegamento ipertestuale al testo integrale in formato elettronico, permettendo la consultazione approfondita delle fonti originali.

La sezione "Mappe" propone invece una visualizzazione dinamica di itinerari culturali georeferenziati che consentono ai visitatori odierni di esplorare in tempo reale i luoghi del patrimonio bolognese seguendo le "voci" dei viaggiatori che li hanno preceduti nel corso dei secoli (vd. Fig. 2). L'applicazione UniVOCIttà contribuisce così ad arricchire l'esperienza del territorio mediante lo *storytelling* virtuale, una strategia narrativa che collega racconti ed emozioni personali alla visita dei siti culturali (Roque, 2022: 26-27).



Figura 1. La visualizzazione lineare nella sezione "Corpus"

[9] https://univocitta.github.io/UniVOCItta/ (cons. 24/01/2025). Poiché sviluppata nell'ambito di un progetto finanziato con durata annuale, l'applicazione web mobile del progetto è stata concepita come prototipo per testare l'approccio metodologico. Data la natura sperimentale di questa fase, sono previsti sviluppi futuri per migliorare l'accessibilità e conformare la piattaforma agli standard previsti dalle direttive UE e alle linee guida definite dall'Agenzia per l'Italiana Digitale (AgID).



Figura 2. La visualizzazione dinamica nella sezione "Mappe"

5. IMPLEMENTAZIONI IN CORSO E FUTURE

Sebbene l'applicazione sia stata concepita come una risorsa plurilingue destinata a un pubblico internazionale, l'accesso ai frammenti testuali del corpus comparabile BER-LET risulta attualmente limitato a coloro che conoscono le lingue di partenza. Le testimonianze letterarie nelle sezioni "Corpus" e "Mappe" sono infatti disponibili esclusivamente in lingua originale (es. Dickens in inglese, Stendhal in francese, Cervantes in spagnolo, Tolstoj in russo), come si evince dalla Figura 2.

Per ovviare a questo limite, una nuova fase è attualmente in corso nell'ambito di un progetto di dottorato e prevede l'integrazione delle traduzioni delle citazioni in lingua italiana e nelle altre lingue del progetto. Tale implementazione mira a favorire una maggiore fruibilità dei contenuti dell'applicazione, rendendo questa esperienza inedita del patrimonio disponibile alla cittadinanza locale, al pubblico italiano, e a turisti e studiosi di diversi orizzonti linguistici e culturali.

In mancanza di traduzioni edite della maggior parte dei testi del corpus e quindi dell'impossibilità di costituire un corpus parallelo, è stato stabilito di testare l'efficacia di un software di traduzione automatica neurale, DeepL, e di un chatbot generativo di ultima generazione, ChatGPT versione 4o. Questi sistemi presentano delle differenze significative sia nella loro architettura che nei dati utilizzati per l'addestramento (Hendy et al., 2023). I modelli di traduzione automatica neurale, come DeepL, si basano tipicamente su un'architettura *encoder-decoder* e vengono addestrati su corpora paralleli di grandi dimensioni allineati a livello di frase (Koehn, 2020: 125; Pérez-Ortiz et al., 2022: 143; Yvon, 2023: A19). Diversamente, GPT, così come altri modelli generativi di grandi dimensioni (*Large Language Models* - LLM), si fonda su un'architettura *decoder-only* e apprende in maniera automatica come generare porzioni di testo coerenti e ben strutturate attraverso l'esposizione a enormi quantità di dati principalmente monolingui, estratti da fonti eterogenee disponibili sul Web (Ciotti, 2023: 299). Sebbene non sia stato specificatamente progettato per svolgere compiti di traduzione, recenti studi dimostrano che ChatGPT è in grado di produrre traduzioni efficaci e accurate, pur presentando ancora diverse criticità (Jiao et al., 2023; Hendy et al., 2023). Questi risultati suggeriscono dunque la necessità di condurre ulteriori ricerche che mettano a confronto e valutino le prestazioni di diversi modelli di intelligenza artificiale, soprattutto in settori specialistici tradizionalmente meno esplorati, come quello delle belle arti.

Alcuni test preliminari condotti sul sotto-corpus francese BER-LET hanno evidenziato numerosi limiti di ChatGPT 4o e DeepL nella resa in italiano di lessemi e termini relativi al patrimonio artistico e culturale a causa della mancanza di addestramento su dati di dominio specializzati. A titolo esemplificativo, entrambi i sistemi hanno mostrato delle approssimazioni nella traduzione di combinazioni lessicali specializzate, come nel caso di *piliers carrés et ronds*, che è stato tradotto letteralmente come "pilastri quadrati e rotondi", senza cogliere la distinzione semantica e concettuale propria della lingua di arrivo. In italiano esistono infatti due termini distinti nella terminologia architettonica: i "pilastri" sono a sezione quadrata o poligonale, mentre le "colonne" sono a sezione circolare (Zotti & Gramellini, 2025).

Per correggere gli errori generati da questi sistemi artificiali intelligenti nella traduzione del linguaggio specialistico dei beni culturali (Zotti & Turci, 2022), è stata avviata la costituzione di un sotto-corpus comparabile italiano di letteratura di viaggio che sarà consultabile tramite l'applicazione. Questa risorsa, originariamente non prevista nel progetto e attualmente in corso di realizzazione, rappresenta un'implementazione significativa volta al miglioramento delle prestazioni delle tecnologie di traduzione automatica. Il sotto-corpus sarà dapprima impiegato come riferimento per la revisione e la correzione manuale degli *output* inaccurati generati dai sistemi di intelligenza artificiale (*post-editing*). Successivamente, verrà utilizzato per l'apprendimento supervisionato di questi strumenti attraverso una fase di *fine-tuning* finalizzata a ottimizzare la qualità traduttiva in relazione all'ambito di specialità del progetto. Il sotto-corpus italiano faciliterà inoltre la conservazione digitale di un patrimonio letterario attualmente poco noto, data la scarsità di resoconti di viaggiatori e viaggiatrici italiane in Emilia-Romagna digitalizzati e consultabili su biblioteche e archivi online. In questo modo, l'attuale implementazione del progetto di ricerca consentirà a un pubblico sempre più ampio di conoscere il patrimonio artistico e culturale bolognese.

BIBLIOGRAFIA

Andler, D. (2023). Intelligence artificielle, intelligence humaine : la double énigme, Paris: Gallimard.

Apostolakis, A. (2003). The convergence process in heritage tourism. Annals of Tourism Research, 30(4), 795-821. https://doi.org/10.1016/S0160-7383(03)00057-4.

Barbuti, N. (2018). From Digital Cultural Heritage to Digital Culture: Evolution in Digital Humanities. DTUC '18: Proceedings of the 1st International Conference on Digital Tools & Uses Congress, Paris 3–5 Ottobre 2018. Reyes, E., Szoniecky, S., Mkadmi, A., Kembellec, G., Fournier-S'niehotta, R., Siala-Kallel, F., Ammi, M. & Labelle, S. (Eds.). 1-3. https://doi.org/10.1145/3240117.3240142.

Brilli, A. (2005). Bologna nei taccuini e negli scritti dei viaggiatori stranieri dell'Ottocento. Banca Etruria. ISBN: 201-81-0100-168-8.

Brilli, A. (2014). Viaggiatori stranieri fra Romagna e Marche: 19.-20. Secolo. Bologna: Minerva Edizioni. ISBN: 978-88-7381-717-8.

Ciotti, F. (2023). Minerva e il pappagallo. IA generativa e modelli linguistici nel laboratorio dell'umanista digitale. Testo e Senso, 26, 289-315. https://doi.org/10.58015/2036-2293/671

Cusatelli, G. (1986). Viaggi e viaggiatori del Settecento in Emilia e in Romagna. Bologna: Il mulino.

Donati, D. (2019). La digitalizzazione del patrimonio culturale. Caratteri strutturali e valore dei beni, tra disciplina amministrativa e tutela opere d'ingegno. P. A. Persona e Amministrazione, 2, 323-337. https://doi.org/10.14276/2610-9050.2093.

Farina, A. & Martínez, M. C. N. (2020). La banca dati LBC. In Billero, R., Farina, A. & Martínez, M. C. N. (Eds.), I Corpora LBC. Informatica Umanistica per il Lessico dei Beni culturali. Firenze: Firenze University Press. 7-17. https://doi.org/10.36253/978-88-5518-253-9.

Farina, A. (2016). Le portail lexicographique du Lessico plurilingue dei Beni Culturali, outil pour le professionnel, instrument de divulgation du savoir patrimonial et atelier didactique. Publif@rum, 25, https://riviste.unige.it/index.php/publifarum/article/view/1811.

Hendy, A., Abdelrehim, M., Sharaf, A., Raunak, V., Gabr, M., Matsushita, H. & Awadalla HH. (2023). How good are GPT models at machine translation? A comprehensive evaluation. arXiv preprint arXiv:2302.09210.

Jiao, W., Wang, W., Huang, JT., Wang, X. & Tu, Z. (2023). Is ChatGPT a good translator? A preliminary study. arXiv preprint arXiv:2301.08745.

Koehn, P. (2020). Neural Machine Translation. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108608480.

L'Homme, MC. (2004). La terminologie: principes et techniques. Presses de l'Université de Montréal. https://doi.org/10.4000/books.pum.10693.

Muhr, R. (1991). ATLAS/ti — A prototype for the support of text interpretation. Qualitative Sociology, 14(4), 349-371.

Münster, S., Apollonio, F. I., Bell, P., Kuroczynski, I., Di Lenardo, I., Rinaudo, F. & Tamborrino, R. (2019). Digital Cultural Heritage meets Digital Humanities. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XLII-2-W15, 813-820. https://doi.org/10.5194/isprs-archives-XLII-2-W15-813-2019.

Pérez-Ortiz, JA., Forcada, ML. & Sánchez-Martínez, F. (2022). How neural machine translation works. In Dorothy Kenny (Ed.), Machine translation for everyone: Empowering users in the age of artificial intelligence. Berlin: Language Science Press, 141-164. https://doi.org/10.5281/zenodo.6653406.

Perotto, M. & Zotti, V. (2022). Il patrimonio culturale e artistico di Bologna nel corpus LBC (Lexikon of Cultural Heritage) e nel corpus UniCittà (UniCity). In Conterno, C. & Sabba, F. (Eds.), Il patrimonio culturale della Biblioteca Universitaria di Bologna e della città allo specchio dei viaggiatori europei. Esplorazioni tra la prima modernità e l'era contemporanea, Bologna: Bologna University Press, 161-165.

Roque, M. I. (2022). Storytelling in Cultural Heritage: Tourism and Community Engagement. In Campos A. C. & Almeida, S. (Eds.), Global Perspectives on Strategic Storytelling in Destination Marketing. IGI Global. 22-37. ISBN: 978-16-6843-436-9.

Yvon, F. (2023). La traduction multilingue : analyse d'une prouesse technologique. MediAzioni, 39. A17-A34. https://doi.org/10.6092/issn.1974-4382/18785.

Zotti, V. & Gramellini, R. (2025). Intelligenza umana e intelligenza artificiale per la traduzione del lessico artistico: valutazioni preliminari sul confronto tra DeepL e ChatGPT 4.o . Atti della XXa Giornata Scientifica REALITER e della XXIIIa Giornata AETER, San Millán de la Cogolla: Cilengua. In fase di revisione.

Zotti, V. & Turci, M. (Eds.). (2022). Nuove strategie per la traduzione del lessico artistico. Da Giorgio Vasari a un corpus plurilingue dei beni culturali. Firenze: Firenze University Press. https://doi.org/10.36253/979-12-215-0061-5.

Zotti, V. (2023). Didactique de la méthodologie de corpus et applications pratiques dans le tourisme patrimonial et en lexicographie bilingue spécialisée : le projet UniVOCIttà. Actes des 11èmes Journées Internationales de la Linguistique de Corpus, Université de Grenoble Alpes, Lidilem. 384-389. https://jlc2023.sciencesconf.org/data/pages/abstracts_JLC_2023.pdf

Zotti, V. (2025). Lessico artistico plurilingue e valorizzazione del patrimonio culturale italiano: i progetti UNICittà e UniVOCIttà. XXXIII Convegno dell'Associazione Italiana per la Terminologia (Ass.I.Term), Napoli 7 novembre 2023, Peter Lang, in stampa.

Zotti, V. (2023). Traduire en français le lexique du patrimoine artistique de la ville de Bologne : le sous-corpus comparable BER du projet LBC. In Zotti, V. & Turci, M. (Eds.), Nuove strategie per la traduzione del lessico artistico. Da Giorgio Vasari a un corpus plurilingue dei beni culturali. Firenze: Firenze University Press. 197-229. https://doi.org/10.36253/979-12-215-0061-5.14.

“Il mio sommario dunque è tutto qui?” Per Franco Fortini

Emmanuela Carbé[1], Mariangela Giglio[2], Pietro Orlandi[3], Jacopo Maria Romano[4], Giulio Quaresima[5]
[1] Università di Siena, Italia – emmanuela.carbe@unisi.it
[2] Università di Bologna, Italia – mariangela.giglio2@unibo.it
[3] Università di Siena, Italia – p.orlandi@student.unisi.it
[4] Università di Siena, Italia – jacopomaria.romano@student.unisi.it
[5] Università di Siena, Italia – giulio.quaresima@unisi.it

ABSTRACT (ITALIANO)
Il poster esamina l’approccio adottato dal Centro di ricerca Franco Fortini nello sviluppo di progetti digitali dedicati allo scrittore, che fu docente all’Università di Siena dal 1971 al 1989. L’aspetto centrale delle iniziative riguarda il contributo scientifico nella gestione e valorizzazione del suo archivio ibrido, conservato presso la Biblioteca Umanistica dell’ateneo. In questa prospettiva, il poster si focalizza su alcuni progetti chiave, e in particolare la pubblicazione di un catalogo semantico che integri la bibliografia di e su Fortini alle descrizioni dei materiali archivistici, nonché l’impiego di tecniche forensi per lo studio dei materiali digitali. La ricerca mostra come questi interventi integrati favoriscano nuove modalità di interazione con il corpus fortiniano, stimolando ulteriori indagini sulle tecniche compositive e sulla complessa storia editoriale delle sue opere. Il progetto intende offrire uno spunto di riflessione sull’importanza del dialogo interdisciplinare, portando un esempio concreto di collaborazione e sinergia tra studiose e studiosi del Centro afferenti a diverse discipline.
Parole chiave: archivi letterari digitali nativi; bibliografia; filologia digitale; Franco Fortini

ABSTRACT (ENGLISH)
“Then is my summary all here?” For Franco Fortini. The poster examines the approach adopted by the Franco Fortini Research Center in developing digital projects dedicated to the writer, who was a professor at the University of Siena from 1971 to 1989. The focus of the initiatives concerns the scientific contribution to the management and enhancement of Fortini's hybrid archive, preserved at the Biblioteca Umanistica of the University. The poster highlights key projects, particularly the publication of a semantic catalog that integrates the bibliography of/about Fortini with descriptions of archival materials, and the application of forensic techniques for the study of born-digital materials. The research shows how these integrated interventions promote new modes of interaction with the Fortini’s corpus, encouraging further investigations into the compositional techniques and the complex editorial history of his works. The project aims to provide a point of reflection on the importance of interdisciplinary dialogue, offering a concrete example of collaboration and synergy among scholars of the Center from various disciplines.
Keywords: Born-Digital Literary Archives; Bibliography; Digital Philology; Franco Fortini

1. INTRODUZIONE
Prima della sua scomparsa, avvenuta nel novembre del 1994, Franco Fortini donò il proprio archivio all’Università di Siena, dove aveva insegnato dal 1971 al 1989. Grazie alla donazione fu istituito, nel 1995, il Centro di Ricerca Franco Fortini, che si dedica alla promozione di studi e ricerche sull’autore e su altri importanti scrittrici e scrittori i cui fondi sono oggi conservati presso la Biblioteca Umanistica dell’ateneo.[1]
Il Fondo Fortini comprende una vasta e articolata raccolta di materiali: carteggi, manoscritti, dattiloscritti, stampe da computer, diari, agende e taccuini; raccolte di materiale bibliografico annotato, ritagli e fotocopie (Lenzini, 2021); disegni, incisioni e dipinti realizzati da Fortini stesso (Fortini, 2001); floppy disk e audiocassette. All’archivio si aggiunge la biblioteca privata, composta da circa 5600 volumi, perlopiù punteggiati da annotazioni e postille.
Il Centro si caratterizza per il suo approccio interdisciplinare volto allo studio e alla valorizzazione del materiale archivistico. Negli ultimi anni, anche grazie a un gruppo di giovani studiose e studiosi, sono state avviate iniziative che hanno incluso progetti digitali e di informatica umanistica con un’impostazione graduale e ad ampio raggio. Tra i progetti più rilevanti si ricorda la pubblicazione della rivista “L’Ospite ingrato” su una piattaforma OJS di USiena Press – Firenze University Press;[2] e la fondazione, con lo stesso editore, della collana ad accesso aperto “Carteggi e Materiali”,[3] che accoglie edizioni scientifiche e critiche

[1] Il sito web del Centro di Ricerca Franco Fortini è disponibile all’indirizzo: https://centrofortini.unisi.it/ (cons. 14/04/2025).
[2] https://oaj.fupress.net/index.php/oi (cons. 13/04/2025).
[3] https://books.fupress.com/series/carteggi-e-materiali-del-centro-interdipartimentale-di-ricerca-franco-fortini/149 (cons. 13/04/2025).

delle corrispondenze e delle opere di Franco Fortini basate sui materiali e sugli avantesti conservati presso l'archivio. Si è inoltre dato avvio a una valorizzazione dei primi numeri della rivista cartacea "L'Ospite ingrato", oggi di difficile reperimento: il progetto, in collaborazione con l'editore Quodlibet, prevede l'uscita di un ebook del primo numero (1998), accompagnato da una nuova nota introduttiva di Luca Lenzini.
Il Centro ha avviato laboratori, seguito tirocini e tesi di master di informatica umanistica: tra le iniziative si segnalano dei primi carotaggi nel campo delle edizioni scientifiche digitali di testi fortiniani (Tancredi, 2020), prove di concetto di repository digitali (e.g. una campagna di digitalizzazione e di pubblicazione online – con relativa metadatazione – dei libretti "Mal'aria", di cui una collezione completa è conservata all'interno del fondo Alessandro Parronchi), di percorsi virtuali divulgativi dedicati all'opera di Fortini e alle sue traduzioni.[4] In questo contributo ci si concentra sugli aspetti che si ritengono metodologicamente più significativi, illustrando alcuni progetti che mirano a un approccio omnicomprensivo all'opera di Fortini: la pubblicazione di un catalogo semantico che integri la bibliografia di e su Fortini alle descrizioni dei materiali archivistici, l'impiego di tecniche forensi per lo studio dei materiali nati digitalmente, i primi approcci di studio critico del fondo ibrido, e le prospettive future con l'impiego di strumenti di analisi quantitativa della sua opera.

2. IL PORTALE ONLINE BIGRAFO

Nell'ambito dei progetti più strettamente legati alle Digital Humanities, "Biblio-grafo. Un catalogo semantico per il Centro di ricerca Franco Fortini" (BiGraFo) rappresenta un'iniziativa avviata nel 2023 per mappare e analizzare la produzione intellettuale e artistica di Franco Fortini. In questa prima fase si è pubblicata una piattaforma digitale che ospita esclusivamente uno strumento di ricerca bibliografica basato sul volume *Bibliografia degli scritti di Franco Fortini* (Bassi & Nencini, 2022), che costituisce il più recente repertorio a disposizione.[5] È stata sviluppata un'ontologia a partire dal modello FaBio (FRBR-aligned Bibliographic Ontology),[6] che si è ritenuto un buon punto di partenza per la rappresentazione dei dati complessi oggetto di questo lavoro, e sufficientemente adatto per avviare almeno una prima riflessione sulle relazioni tra l'informazione bibliografica e il dato archivistico (Antonietti, Sanfilippo, Carbé, 2024). La piattaforma è attualmente supportata da Omeka-S (vd. Fig. 1), un CMS sviluppato dal Roy Rosenzweig Center for History and New Media e oggi gestito dalla Corporation for Digital Scholarship. Ospitata sui server dell'Università di Siena, BiGraFo mira a essere la base per la pubblicazione dei progetti futuri del Centro.
La fase di sviluppo riguarda tre direttrici: ampliare la base di conoscenza mettendo in relazione la bibliografia di Fortini con la descrizione del fondo archivistico cartaceo e della biblioteca (interrogabili tramite API nel catalogo OPAC dell'Ateneo), nonché con la descrizione della parte born-digital contenuta in floppy disk (vd. par. 3); pubblicare nello stesso ecosistema un repertorio della bibliografia della critica, esplicitando le connessioni possibili con l'archivio e la bibliografia di Fortini; infine, integrare nella piattaforma strumenti di indagine quantitativa dell'intero *corpus* della sua opera.

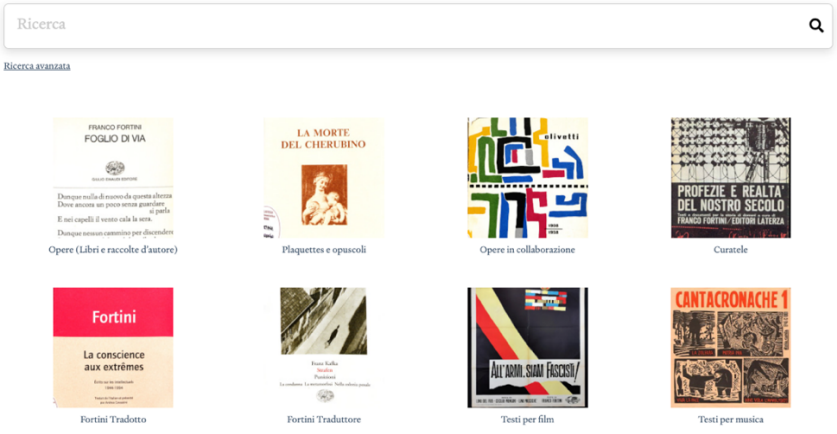


Figura 1. Schermata di accesso del portale BiGraFo

[4] Si vedano le pagine dedicate all'iniziativa Fortini24, realizzate per il trentennale della sua scomparsa: https://dfclam.github.io/fortini24 (cons. 13/04/2025).
[5] La piattaforma è disponibile nella sua versione beta all'indirizzo https://digilet.unisi.it/s/fortini (cons. 13/04/2025).
[6] http://www.sparontologies.net/ontologies/fabio (cons. 13/04/2025).

3. IL FONDO IBRIDO: IPOTESI DI LAVORO PER IL NATIVO DIGITALE

Fortini iniziò a usare il computer già negli anni ’80 (Lenzini, 2016) e nel corso della sua vita sono attestati diversi modelli di Macintosh, compreso un portatile PowerBook 1000 (Marrucci & Tinacci, 2005). Tra le sue poesie è celebre *Durable 5168*, pubblicata nella raccolta *Composita solvantur* (Fortini, 1994), in cui l’autore pone al centro della poesia due floppy disk contenenti il suo lavoro di poeta (“il mio sommario dunque è tutto qui?”, v. 3). Il fondo Fortini, uno dei principali e più ricchi archivi letterari ibridi italiani di uno scrittore del Novecento, rappresenta un esempio emblematico di come le risorse nate digitalmente e quelle cartacee possano essere integrate efficacemente. Oltre ai materiali cartacei, la Biblioteca conserva infatti floppy disk contenenti backup del suo ultimo computer.

Lo studio dei floppy ha evidenziato le sfide connesse alla conservazione e all’analisi di materiali custoditi su supporti obsolescenti, che impongono di elaborare nuove ipotesi di lavoro sul piano archivistico e filologico. Per garantire la conservazione dei materiali digitali, l'Università di Siena nel 2022 ha avviato una campagna di messa in sicurezza dei floppy, accanto alla rimediazione di numerose audiocassette conservate in archivio. Per i floppy è stata effettuata una copia forense (bitstream copy), ovvero una replica esatta e completa di tutti i dati presenti sul supporto digitale. La copia include non solo i file visibili all'utente, ma anche i dati nascosti, cancellati, e lo spazio non allocato, rappresentando quindi una duplicazione fedele dell’intero contenuto del dispositivo (Kirschenbaum, 2010; Boddington, 2016).

Le copie forensi sono diventate successivamente l’ambito di indagine privilegiato di numerosi progetti digitali, contribuendo a definire delle ipotesi di lavoro per la preservazione dei materiali stessi e per la a consultazione in sicurezza (Kirschenbaum, 2012; Ries, 2017; Ries, 2022). Parallelamente, il Centro ha lavorato alla realizzazione di un indice di consultazione dei materiali contenuti nei floppy, al fine di offrire alle studiose e agli studiosi una prima finestra sul metodo di organizzazione dei testi dello scrittore. Infine, è stata avviata l’elaborazione di un software che, basandosi sulla lettura dell’esadecimale dei file, permette di identificare e datare file legacy di tipo testuale. Questo strumento, ancora in fase di sviluppo, consentirà non solo la lettura dei file in esadecimale e la loro trasformazione in un testo leggibile, ma anche forme di annotazione e di estrazione automatica dei *timestamp*.

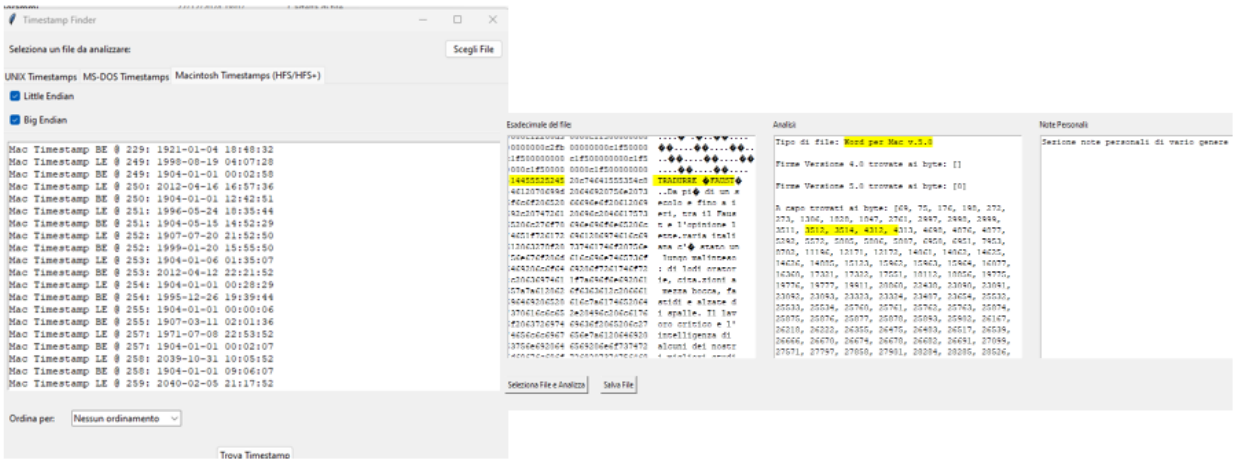


Figura 2. Schermate di esempio del software di analisi dell’esadecimale

4. UN DIALOGO TRA ANALOGICO E DIGITALE

Le tecniche forensi stanno permettendo delle prime analisi dei file contenuti nei floppy disk, rivelando non solo i metadati e le tipologie di file presenti in archivio, ma anche dati cancellati o corrotti. Stanno emergendo ad esempio lacerti di testi e composizioni che l’autore aveva copiato e incollato in vari documenti per rielaborare parti specifiche. Questi residui, memorizzati ma invisibili attraverso i comuni programmi di videoscrittura, offrono una prospettiva unica sulle pratiche compositive dell’autore. Si stanno inoltre rintracciando disposizioni grafiche alternative rispetto ai testi a stampa, che danno prova di uno sperimentalismo compositivo legato all’uso del medium digitale fino a oggi sconosciuto nelle opere di Fortini. Tramite analisi approfondite si stanno infine appurando manipolazioni e corruzioni di file o parti di file. Un esempio è la presenza di stringhe testuali residuali nell’esadecimale: apparentemente inutili, si rivelano invece strumenti preziosi per uno studio filologico dei testi, e in particolare per dimostrare filiazioni e relazioni tra testimoni. In diversi casi è stato possibile accedere a lacerti di componimenti che

erano stati eliminati o parzialmente sovrascritti, e che permettono di individuare fasi compositive e varianti d'autore (Ries, 2018).
L'urgenza di lavorare sull'archivio ibrido di Fortini – includendo quindi all'archivio cartaceo anche la parte nata digitalmente – è particolarmente stringente se riflettiamo sull'ultima stagione creativa dello scrittore, caratterizzata da un'intensa progettualità e soprattutto da un metodo di lavoro biforcato tra supporti digitali e tradizionali: un vero e proprio «continuo pendolarismo tra cartaceo e digitale» (Marrucci & Tinacci, 2005), che si concretizza in un regime contaminato e ibrido. Nel suo fondo archivistico, d'altronde, le stampe da computer convivono con i quaderni e gli appunti manoscritti, gli articoli di giornale e le fotocopie dei contributi in volume, fitte di rimaneggiamenti, con i primi brogliacci di lavoro, prolungandosi fino al suo archivio digitale e toccando addirittura il suo fondo librario.
Tale oscillazione tra cartaceo e digitale, in un dialogo costante, dai confini spesso sovrapposti e sovrapponibili, ben si evince dai numerosi "lavori in corso", a cui l'autore si stava dedicando poco prima della morte. Il suo archivio digitale entra in comunicazione con le serie archivistiche del fondo d'autore, gettando ulteriore luce su Fortini quale «metodico selezionatore» di sé stesso (Conti, 2022: 4): le aggregazioni digitali, difatti, permettono di comprendere l'organizzazione del materiale che lo scrittore operava all'interno del suo archivio personale, le preferenze di ordinamento, i criteri classificatori applicati ai documenti (ora più tipologici, ora più contenutistici), le titolature e le denominazioni aggregative dei *file* coinvolti nelle operazioni di trascrizione tramite *word processor*.
Alla stregua di un mezzo di contrasto, le strutturazioni all'interno della dimensione digitale si rivelano estremamente preziose al fine di comprendere non solo *cosa* Fortini decidesse di riversare nel suo personal computer, ma altresì *come* lo realizzasse. Un limpido esempio per comprendere queste dinamiche conservative, e a un tempo creative e ideative, è il lavoro di trascrizione, ordinamento e aggregazione che Fortini stava svolgendo per una raccolta inedita dedicata alle letterature straniere, intitolata nelle fasi genetiche più avanzate *Saggi stranieri*, e di cui tra l'altro offre testimonianza «un piano per l'*Opera omnia*» allegato nell'autunno del 1993 in una lettera a Ernesto Franco, all'epoca redattore Einaudi.
Tenendo a mente la tendenza, tutta fortiniana, all'indicizzazione, già illustrata da Luca Lenzini (2024), è necessario far dialogare le procedure selettive e coesive testimoniate dagli indici cartacei con le suddivisioni strutturate all'interno della dimensione digitale. Proprio sulla base dei contributi scelti all'interno della raccolta in lavorazione è possibile osservare contrastivamente e interrogativamente le aggregazioni per così dire "analogiche" con quelle digitali: evidenziando in primo luogo le eventuali inclusioni o esclusioni, e secondariamente le proiezioni e le ricadute esegetiche sulla costruzione macro-testuale della raccolta. L'obiettivo ultimo, dunque: fornire un'analisi delle intersezioni tra il Fortini "costruttore di sillogi" e il Fortini conservatore e archivista di sé stesso.

5. CONCLUSIONI E PROSPETTIVE FUTURE

Nel contesto dello studio del lascito di Fortini, il Centro sta dunque sviluppando una serie di progetti interconnessi e complementari che mirano alla valorizzazione del suo patrimonio. Si è sviluppato un flusso di lavoro che propone di integrare più progetti paralleli, ciascuno focalizzato su specifiche aree di ricerca: la consultazione e l'attraversamento dei materiali nati digitalmente, anche con l'applicazione di metodi forensi, la creazione di un catalogo semantico, pensato per espandersi e aggiornarsi con nuove iniziative integrate. Tra queste, l'obiettivo più ambizioso è offrire all'interno della piattaforma un'indagine quantitativa dell'opera di Fortini, da mettere in relazione con il catalogo bibliografico e archivistico. Sono state avviate alcune prime prove, basate su tecniche NLP, limitatamente all'opera poetica di Fortini, grazie a un lavoro di digitalizzazione e OCR dei testi e alla definizione di un modello per il caso fortiniano.
Queste iniziative sono progettate per coprire un ampio spettro del lavoro filologico e critico di Fortini, introducendo nuovi strumenti e metodi di analisi che rispondano alla complessità e alla varietà del materiale e della produzione letteraria di un intellettuale del Novecento. Il gruppo di lavoro sostiene la propria operatività attraverso un dialogo continuo e multidisciplinare tra specialisti di diverse aree, che integrano reciprocamente le loro competenze specifiche. Il Centro si propone dunque come un polo culturale attivo, e in continuo dialogo con altre realtà. Questa sinergia permette di approfondire e ampliare le metodologie di studio dell'archivio, facilitando l'analisi e la comprensione dell'opera: la vitalità e sostenibilità dei progetti digitali può essere assicurata solo grazie a un mosaico di voci, in cui tutti gli elementi coinvolti possono contribuire armonicamente a una visione progettuale d'insieme.

BIBLIOGRAFIA

Antonietti, L., Sanfilippo, E.M., Carbé, E. (2024). L'ontologia BiGraFo: verso un modello semantico per l'opera di Franco Fortini. In Me.Te. Digitali: Mediterraneo in rete tra testi e contesti: Proceedings del XIII Convegno Annuale AIUCD, 423-428.

Bassi, E., & Nencini, E. (2022). Bibliografia di Franco Fortini. Macerata: Quodlibet.

Carbé, E. (2023). Digitale d'autore. Macchine, archivi, letterature. Firenze-Siena: FUP University Press. 10.36253/979-12-215-0023-3

Conti, A. (2022). Franco Fortini al vaglio della militanza. L'avventura della poesia tra "Verifica dei poteri" (1965), "Una volta per sempre" (1963) e "Questo muro" (1973). Finzioni, 1-16. 10.6092/ISSN.2785-2288/14153

Fortini, F. (1994). Composita solvantur. Torino: Einaudi.

Fortini, F. (2001). Franco Fortini. Disegni, incisioni, dipinti, a cura di E. Crispolti e G. Nava. Macerata: Quodlibet.

Fortini, F. (2006). Un giorno o l'altro, a cura di M. Marrucci e V. Tinacci. Macerata: Quodlibet.

Kirschenbaum, M.G. (2012). *Mechanisms: New media and the forensic imagination* (First MIT Press paperback edition). Cambridge, Massachusetts: The MIT Press.

Lenzini, L. (2016). Don't Save! Fortini, un Mac e le ultime cose. L'Ulisse. Rivista di poesia, arti e scritture, 19, 97–101.

Lenzini, L. (2021). L'Archivio Fortini: Consuntivo provvisorio. Quaderni del PENS, 4, 7–12. 10.1285/i2611903xn4p7

Marrucci, M., & Tinacci, V. (2016). Governare il magma. Un giorno o l'altro a dieci anni dall'edizione. L'Ulisse. Rivista Di Poesia, Arti e Scritture, 19, 102–106.

Marrucci, M., & Tinacci, V. (2005). L'edizione di uno scritto a testimonianza plurima, cartacea e informatica: Un giorno o l'altro di Franco Fortini. Filologia italiana, 2, 215-221.

Ries, T. (2017). Philology and the digital writing process. *Cahier voor Literatuurwetenschap*, 9, 129–158.

Ries, T. (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.

Ries, T. (2022). Digital history and born-digital archives: The importance of forensic methods. *Journal of the British Academy*, *10*, 157–185. 10.5871/jba/010.157

Tancredi, G. (2020). Al di là della speranza. Una prova di filologia digitale d'autore (edizione diplomatica e critica). Tesi del Master in Informatica Umanistica ed Edizione elettronica, relatore Roberto Rosselli del Turco, Università di Siena.



UNIVERSITÀ di VERONA
Dipartimento di LINGUE E LETTERATURE STRANIERE

daih

ASSOCIAZIONE per l'INFORMATICA UMANISTICA e la CULTURA DIGITALE