

European Association for Digital Humanities 2026

"Linking Europe: Digital Humanities Without Borders"

Book of Abstracts



Jagiellonian University, Kraków, Poland
15-19 September 2026

Under the Honorary Patronage of the Rector
of the Jagiellonian University



JAGIELLONIAN UNIVERSITY
IN KRAKÓW

PAN



Instytut
Języka
Polskiego

Polskiej
Akademii
Nauk



Uniwersytet
Wrocławski



JAGIELLOŃSKIE CENTRUM
HUMANISTYKI CYFROWEJ



Digital Humanities Lab
Flagship Project at Jagiellonian University



RESEARCH
UNIVERSITY
EXCELLENCE INITIATIVE

Book of Abstracts

European Association for Digital Humanities 2026 "Linking Europe: Digital Humanities Without Borders"

15-19 September 2026, Jagiellonian University, Kraków, Poland

Co-organised by the Institute of Polish Language of the Polish Academy of Sciences and
the University of Wrocław

Layout and typesetting

Justyna Łukaszyk

Logo design

Patrycja Wojkowska

Edited by

Aleksandra Rykowska

Proofreading

Aleksandra Rykowska

ISBN 978-83-977695-1-9

DOI 22078032

Disclaimer

The content of each abstract is the sole responsibility of its author(s). The editors
and organizers are not responsible for the views expressed in the abstracts.

License

This publication is licensed under a Creative Commons Attribution 4.0
International License (CC BY 4.0)

Table of contents

17 Welcome

19 Introduction

Keynotes

22 Prof. Melissa Terras *Dreaming of Different DH Futures: Engaging with
Speculative Design Approaches*

23 Prof. Tomasz Majkowski *Digital Games and Fascist Pedagogy*

Workshops

25 A FAIR Flow: *Interactive Text Publication and Analysis with TextGrid and
MONAPipe*

31 Bridging text and meaning with CLARIN-PL: *Digital press, relational
lexicography and AI in SS&H*

39 Building Heritage Digital Twins with Open-Source Tools: *From Local
Collections to ECCCH, Europeana, and Wikimedia*

41 Graph-Based Text and Knowledge: *Modelling and Working with the
RAMEN Suite*

51 Is Everything Already Lost?! *Participative Workshop on Strategies for
Ensuring the long-term Hosting of Digital Scholarly Editions*

57 Knowledge Graphs and Linked Open Data made (quite) easy: *OntoME,
SHACL, Logre*

61 Network Analysis in Digital Humanities

63 No pain, no gain? *Rethinking research data management planning in DH:
A participative workshop on the use and benefits of data management
plans*

72 Research Data beyond Boundaries: *Semantic Linking of Collections
with WissKI*

79 Warsztaty przedkonferencyjne: *projektowanie zajęć, testów i egzaminów
wspierających SRL z dopuszczonym użyciem AI*

- Long papers**
- 84 All the World's a (Digital) Stage: Rethinking Presence in Digital Heritage**
- 85 Accessible Cultural Heritage: Digital Archives as a Creative Resource for Representing Presence in Contemporary Polish Off-Theatre
- 90 Digital Grassroots: Floppy Disk Magazines from Poland and the Challenge of Early Born-Digital Heritage
- 98 Staging Culture: Performativity, AI, and the Ethics of Digital Mediation

- Long papers**
- 105 Ask and You Shall Retrieve: LLMs, Libraries, and the Search for Meaning**
- 106 At the intersection of DH and LIS: a comparative analysis from Italy
- 113 Do you know what I'm looking for? LLM-supported query formulation for complex query languages
- 122 Integrating GenAI into Lexicographic Workflows: A Hybrid Approach to Semantic Classification

- Long papers**
- 131 Between the Lines: Rethinking Structure, Data, and Sequence**
- 132 A new approach for defining Research Data Management requirements – Development of a formal description model for the field of RDM
- 141 Beyond Pagination: Navigational Time and the Epistemology of the Digital Historical Text
- 147 Meta-modelling as a context system

- Long papers**
- 156 Building to Last: Infrastructures, Archives, and the Architecture of Memory**
- 157 From digital reproduction to cultural entity Structuring the historical medical records of the Former Psychiatric Hospital of Reggio Calabria (Italy)
- 166 Future Applications – The Text+ Registry as Reusable Technical Component Beyond Resource Cataloguing

- 175 Living Archives of the Present: Digital Humanities as Constitutive Infrastructure in the Edut 710 Initiative

- Long papers**
- 182 By the Book: Reengineering Polish Bibliography**
- 183 Bibliography, LLMs, and Polish Prints in 16th–19th Centuries
- 191 Curating a translation-centered set of 300 000 bibliographic metadata records with 60 humans in the loop: What are the gains for data-driven translation history?
- 198 Lem Knowledge Graph: A Project Overview

- Long papers**
- 203 Computational Readings of Literary Texts: Cliffhangers, Curricula, and Crucibles**
- 204 Emotions in Feuilleton Cans: sensationalism in nineteenth-century Romanian novels
- 214 Genre Variation in Medieval Latin Alchemy: A Typological Approach
- 228 Using AI techniques to identify thematic topics in Greek literature curriculum

- Short papers**
- 236 Counting Verse: Computational Approaches to Poetry and Prose**
- 237 Colorfulness and Brightness in Karel Hynek Mácha's Novels in the Context of Quantitative Models and a RAG Agent
- 246 Exploring Nature in Large Visual Archives: The Case for Computational Environmental Humanities
- 252 "I/We" and Community in Nesmelov's Harbin Poems (1940)
- 258 Jakobson's „Poetic Function" in Vector Space: Parallelism, Cognitive Geometry, and Classical Chinese Verse
- 268 Modelling Differentiated Translationscapes: Network Dynamics of the Romanian Novel (1860–2025)
- 275 What Drives Change in Poetry: The Role of Cohort Effects in Poetic Feature Development Over Time

Short papers

282 **Curating Culture: Open Data and Infrastructure for Literature and Museums**

- 283 *Curation, circulation and visibility of Norwegian fiction in the digital literary sphere: platform-based literary dissemination of fiction from The Norwegian Purchase Scheme for Literature*
- 288 *Database of the Czech Literary Awards: From the Proprietary Solution to the FAIRified Dataset*
- 293 *JUHMP: A CIDOC-CRM-based Semantic Portal Prototype for Linking Jagiellonian University Heritage Across Archives, Libraries, and Museums*
- 298 *Museums as Open Science Infrastructure – Introducing the Museum Infrastructure Readiness Index (MIRI)*
- 305 *Reading in Digital Social Platforms: How Design Shapes Engagement*
- 310 *Visibility of Institutional Histories: A LOD-Based Approach to a Heterogenous Collection of GDR Museology*

Long papers

316 **Digital Editions Unbound: Graphs, APIs, and Networking**

- 317 *Bridging Isolated Projects: A Technical Approach to Interoperability in Digital Editions*
- 323 *Digital Editions for the Formal Sciences: The Case of Mathematics*
- 328 *LLM-Annotationen with mcp in the grapbased digital edition model ATAG*

Short papers

338 **Digital Fingerprints: Detecting Human and Machine Authorship**

- 339 *Availability of Indicators in a System for Using Rare Word Ngrams (Coulthard Tuples) to Determine Authorship*
- 344 *Basic linguistic measures help distinguish human and machine literary translation*
- 348 *Do Large Language Models have their specific styles? Stylometric analysis of machine generated texts*
- 356 *Evaluating BERTopic N-gram Configurations on Latvian Exile Press (1940s–2000s): the Coherence Paradox*

- 362 *Recognizing Authorial Style across English-to-Japanese Translation: Evidence from Nineteenth-Century American Narrative Prose*
- 368 *The Mysterious Case of Eva Bloch: Authorship Recognition in Translations*

Long papers

374 **Faithful Departures: Tracing Translation Strategy from Scripture to Censorship**

- 375 *Beyond Omission and Insertion: A Multi-Translation Corpus Study of „I Promessi Sposi“*
- 384 *Reading through the Iron Curtain: Translated Children’s Literature in the USSR during the Cold War*
- 392 *The targum Corpus: A Deep, Multilingual Resource for New Testament Translation Studies*

Short papers

399 **From Ink to Data: Automating the Digitization of Manuscripts, Maps, and Comics**

- 400 *From Image to Machine-Readable Text: Methodological Challenges of OCR and NLP in Historical Map Digitization*
- 405 *How to (Machine) Read 35,000 Chinese Comics? A Work-in-Progress Report on Digitization and Database Creation Efforts*
- 411 *Rebuilding Babel: AI-Mediated Interoperability for Working with Historical Sources*
- 416 *Towards Automated Processing of Late Medieval Manuscripts: A Case Study from Fifteenth-Century Poland*
- 420 *Tracing Parcel Change from Historical Maps to Built Form: A Land Law Oriented Spatial Humanities Workflow*
- 425 *Visualizing Gender Representation: Designing Bias-Aware Computer Vision Experiments for Chinese Comics*

Long papers

430 **From Scratch to Script: Training Models for Manuscripts, Newspapers, and Rare Scribal Hands**

- 431 An Open-Source AI Pipeline for Layout Segmentation and OCR of Early Chinese Periodicals
- 441 Atlas of Holocaust Literature – a digital infrastructure for mapping the Holocaust experience in literature
- 444 From Catalogue to Digital Resource: Digitizing and Reusing Manuscript Images from the Monastery of Konstamonitou (Mount Athos)

Long papers

- 450 **Higher Powers: Authority, Metaphor, and Sacred Space**
- 451 Modeling a Knowledge Graph of Religious Metaphors
- 461 Place-based Linking and Visualisation of Multidisciplinary Research Data: Demonstrators from the ATRIUM Project
- 468 Water-related professions, water-related cults: On the Spearman rank correlations of various Latin epigraphical and other data from the Roman Empire

Long papers

- 475 **Linked In: Communities, Taxonomies, and Rights in the DH Ecosystem**
- 476 DH as Linked Communities: A Comparison of Participants at Annual Conferences and in Lecture Series in Germany, Austria, and Switzerland
- 483 Heritage Digital Twins Beyond Modelling: Rights-Aware Architectures and Governance Patterns for European Cultural Data Infrastructures
- 491 Linking Research Activities: TaDiRAH as a Shared Taxonomy for Digital Humanities Research Practices

Long papers

- 498 **Making Their Mark: Vision, Iconicity, and the Automation of Cultural Traces**
- 499 Automated detection and transcription of figures in publications in humanities: the case of stemmata
- 510 Automated Maker's Mark Retrieval in Silverware

- 518 Beyond Description: Modelling Iconicity and Sensorimotor Grounding in Museum Audio Description

Short papers

- 525 **Modeling the Past: Events, Memory, and Space in Historical Data**
- 526 Data and Metadata Analysis of a Digital Humanities Holocaust Memory Project
- 530 Emotion-based approaches on personal accounts of the Holocaust and its aftermath
- 535 Linking Medieval Europe: collecting and connecting data on multilingual textual traditions
- 543 Serial Administrative Records as Relational Event Data – Modelling Historical Collection Infrastructures: The Case of Two Viennese Herbaria
- 549 The Theory of Servant Villages – Spatial Approach
- 552 „There is a land we love with shady beech-trees aspread”: Naturalizing the Nation in 19th Century Northern Europe

Long papers

- 561 **Networked Knowledge: From Verse to Vectors**
- 562 Linking the Unlinkable: Network Analysis and Fragmentation
- 569 Measuring the Shift: Interlinear Baselines for Quantifying Translator Intervention in Vector Spaces
- 578 'Stay lovely Boy, why fly'st thou mee': Network analysis and the circulation of early modern English poetry

Long papers

- 583 **Odzyskane, odtworzone, odegrane: odstępny polskiej humanistyki cyfrowej**
- 584 Język polski w Brazylii jako język oddziedziczony. Badania z wykorzystaniem narzędzi cyfrowych i AI polszczyzny prasy polonijnej: digitalizacja, budowa repozytorium i korpusu
- 590 Muza uczy się grać. Homeryckość w Assassin's Creed Odyssey
- 596 Witkacy i sztuczna inteligencja. Próba rekonstrukcji zaginionych dramatów

Long papers

615 Off the Main Path: Finding Style, Seams, and Shifts

- 616 Beyond the Close/Distant Divide: Digital Humanities Approaches to Style and Point of View in Translation
- 625 Cognitive Stylometry: A Computational Study of Defamiliarization in Modern Chinese
- 635 Neural Seismography: Detecting Redactional Seams in the Book of Isaiah with Dual-Model Transformer Perplexity

Poster Session

- 642 A corpus-based modelling of translator's style: stance, stylistics and repetition in English-to-Lithuanian translation of recurrent reporting verbs in „Brave New World” by A. Huxley
- 647 'Care Check' for Digital Platforms – Feminist Approaches to the Digitization of (Ethnographic) Data
- 650 Close to Jupyter: Empowering TextGrid Workflows with Jupyter4NFDI
- 654 Creating Space for New Ideas with OPERAS Innovation Lab
- 657 Designing a Collaborative Infrastructural Cluster for the Humanities: A Case Study of the Czech National Open Science II Project
- 660 Digital Critical Editing in Romanian Culture. Contributions to Research, Development and Innovation Using a Framework for Screens
- 665 Don't shy away from ethical and legal challenges—with ELSA4Memory
- 668 DoTS Suite: A Complete FAIR Solution for TEI Data Publishing and Searchable Web Application Development
- 672 e-ilustrace.cz: from static digitized artifacts into structured, interoperable research data
- 675 From Fragmented Heritage to Federated Digital Twins on ECCCH, Europeana, EOSC: Building the Finno-Ugric Data Sharing Space
- 678 Grammatical Information in the TEI Critical Apparatus
- 683 "Integrating Digitized Cultural Heritage into Spatial Redesign Using Digital Twins and Traditional Patterns for Immersive, Intercultural Environments"
- 691 Language in Expansion: Genre and Diachronic Variation in Danish Newspapers

- 702 Linking Early Modern Europe by means of an Applied Ontology of Place in a ResearchSpace Knowledge Contextualizing Environment
- 713 Linking Rumour Flows: A Gazetteer-Based Toolkit for Source and Reuse Detection in Historical Newspapers
- 716 Mapping the Science of Interwar Czechoslovakia: Digital Reconstruction of Scientific Networks and Methodological Challenges
- 720 Matching Papyri Fragments Using Computer Vision Methods and Graph Networks
- 723 Modelling Early Modern Reasoning for the Venetian Risposte: a Scalable Framework
- 732 Controlled Vocabulary for Content Notes on Historiographic Research Data
- 735 Open Tool Registries! Resolving the Directory Paradox with Wikidata
- 740 Operationalising Otherness: Evaluating Large Language Models in the Annotation of Literary Texts
- 745 Psalterium – nowy typ edycji dydaktycznej tekstu dawnego/Psalterium: A New Model of a Pedagogical Digital Scholarly Edition of Early Polish Texts
- 749 Recollectio: a framework for digital collection catalogs
- 752 RE-CREATIO – repozytorium cyfrowe poświęcone niezrealizowanym projektom architektonicznym
- 758 Stan prac nad „Korpusuj” – aplikacją do tworzenia i przeszukiwania automatycznie anotowanych tekstów
- 761 The ANR E-clesia project: Linking Texts and Images on Late Antique and Medieval Church Architecture
- 766 Theatre and Modernity. An AI-driven transformation of the 'Deutscher Bühnen=Spielplan' (1896-1910)
- 771 The GOLEM Annotated Corpus for Computational Literary Studies
- 779 The Great Unreferenced
- 784 The Lexical Database of Medieval Polish – its functionality and usefulness in research into Old Polish inflection
- 787 Thinking Differently About Academic Publishing: The Open Encyclopedia System in Humanities Research and Teaching
- 793 Towards Computational Analysis of 19th-Century Poetry in Galician Translation Spaces. Denationalizing translatorial metadata through digitization
- 801

- 804 Unified Open-Source Pipeline for Archival Data Analysis
- 808 Visual Reconstruction as Memory Negotiation: An Iterative Generative AI-Mediated Framework for Oral History
- 812 Weaving Conceptual Documentation with Operations: Connecting OntoME and WissKI through LOD4HSS

Long papers

- 815 **Reading the Credits: Authorship, Adaptation, and Reception Across Media**
- 816 Digital Titrology: A Data-Driven Analysis of Global Film Title Adaptation Rates (1950–2025)
- 822 Kafka in Paragame Discourse: Exploring Reception Patterns in Steam Reviews

Long papers

- 830 **Same Old Story? Cliché, Colonialism, and Community in the Age of AI**
- 831 AI and the Problem of Cliché
- 838 Data Colonialism and Indigenous Data Sovereignty in Generative AI for Endangered Languages
- 848 Modelling transhuman textual communities

Short papers

- 856 **Small Languages, Big Data: Annotation and NLP Beyond the Mainstream**
- 857 Building Infrastructure for Corpus-Driven Lexicography and Linguistic Annotation: The Middle-Persian Dictionary and Corpus (MPCD)
- 865 Data Infrastructure and LLM Benchmark for the Masurian Lect
- 871 Let's build the resources we need without the help we can't get! Wikidata as a communal union catalogue for materials in under-resourced languages
- 876 Lost Frequencies: Linguistic Disconnections in the Digital Recognition of Traditional Narratives
- 882 Morphological Annotation of Old Serbian in Universal Dependencies

- 890 Specific domain Named Entities Recognition for low resourced languages: a case study on Romanian fairytales

Short papers

- 898 **Structured Speech: Semantic Networks Across Administrative and Scholarly Genres**
- 899 CoMPaRS as a Cross-Linguistic Functionally-oriented Resource for LLM Evaluation
- 907 DiploBERT: A domain-specific transformer encoder for Diplomatic text understanding
- 913 Linking Speakers of the German Parliament to Wikidata: Scope and Coverage of Metadata
- 953 Quantifying Pseudoscience: Data-driven analysis of an astrological journal
- 957 To repeat or to shift? A corpus-based modelling of English-to-Polish translation of reporting verbs in EU press releases
- 968 Yet Another Death of Lexicography? Compiling Diachronic Latin Dictionary with LLMs

Short papers

- 973 **Słowo w słowo: leksykografia, norma i styl pod lupą danych**
- 974 Empiryczna weryfikacja autorskiej probabilistycznej koncepcji normy językowej (na materiale fleksyjnym i słowotwórczym)
- 979 Stylometryczne badania pism Fulcanelliego z użyciem pakietu Stylo
- 983 W głąb XVII-wiecznej polszczyzny – próba automatycznej dystrybucji znaczeń w hasłach przymiotnikowych z Thesaurusa Grzegorza Knapusza

Short papers

- 988 **The Cost of Doing DH: Sustainability, Law, and Infrastructure Governance**
- 989 CLARIAH-VL+ as a structural node in the European SSH Research Infrastructure: Towards a Flemish Open Science Cloud
- 993 From Records to Events: Epistemic and Technical Dimensions of Transforming the PAESE Database into an Event-Oriented Structure

- 997 Geographic Citation Insularity in Digital Humanities: A Bibliometric Analysis Using OpenAlex and The Index of DH Conference
- 1005 Infrastructure Without Borders: Hosting the First ERIC in Poland
- 1010 When a Database Dies: Legal Instruments for Preventing the Loss of Knowledge

Long papers

- 1016 **The Stylometric Impostor: AI Translations, Copycats, and Pivot Spaces**
- 1017 *Homage, Copycat or Brilliantly Different? Christina Dalcher and The 'Handmaid's Tale Bandwagon'*
- 1026 *Is Textual Similarity Invariant under Machine Translation? Evidence Based on the Political Manifesto Corpus*
- 1036 *The Stylometry of Literary Translation by ChatGPT: a Bilingual Experiment*

Short papers

- 1046 **Unsearchable, Unseen, Unheard: Barriers to Visibility in Digital Humanities**
- 1047 *Digital Accessibility Infrastructures and Deaf Audiences in Italian Theatre*
- 1052 *Disabled Languageing in the Digital World: Applying Crip Linguistics to the Study of Online Language Practices Among Individuals with Disabilities*
- 1057 *From Black Box to Glass Box: The Hermeneutics of Vulnerability and the Ethics of Care in Medical Artificial Intelligence*
- 1062 *What Platforms Retrieve: Romanisation Conventions and the Digital Invisibility of Chinese Migrants in European Archives (1900–1950)*

Short papers

- 1066 **Weaving the Web: Linking Data Across DH Projects**
- 1067 *A Collaborative Bibliography on Digital Humanities and Artificial Intelligence: Practices, Infrastructure, and Knowledge Organization*
- 1072 *Modelling Early Modern Institutional Authority: A TEI-Based Pilot on Cross-Border Foundations*

- 1075 *Polilex. Linking Lexical Resources for Polish*
- 1082 *Put it in the SCRINIUM and it's never lost: a multisearch application for DH projects*
- 1086 *Zotero as Semantic RDF Platform for an Entire Humanities Research Workflow*

Long papers

- 1094 **Words in Numbers: Sound, Synonymy, and Statistical Semantics**
- 1095 *Creating a Historical Pre-trained Ottoman Language Model Using BERT*
- 1101 *High-Dimensional Phonosemantics: Modeling Sound–Meaning Mappings in English and Polish lexis*
- 1108 *Synonymy and Information Theory: Quantifying Contextual Predictability and Information Gain in Language*

AI Statement

In the preparation of this abstract, ChatGPT (OpenAI) was used to assist in rephrasing and restructuring portions of the text in order to improve readability, clarity, and overall flow. The tool was employed solely for linguistic refinement and stylistic optimization. All substantive content, technical descriptions, research framing, and scholarly claims originate from the authors. The authors reviewed and revised all AI-assisted output and take full responsibility for the final version of the text.

References

- Abitbol, Roy, Ilan Shimshoni, and Jonathan Ben-Dov. 2021. "Machine Learning Based Assembly of Fragments of Ancient Papyrus." *Journal on Computing and Cultural Heritage* 14 (3): 1–21. <https://doi.org/10.1145/3460961>.
- Ostertag, Cecilia, and Marie Beurton-Aimar. 2021. "Using Graph Neural Networks to Reconstruct Ancient Documents." In *Pattern Recognition. ICPR International Workshops and Challenges*, edited by Alberto Del Bimbo, Rita Cucchiara, Stan Sclaroff, et al., vol. 12667. Springer International Publishing. https://doi.org/10.1007/978-3-030-68787-8_3.
- Pirrone, Antoine, Marie Beurton Aimar, and Nicholas Journet. 2019. "Papy-S-Net: A Siamese Network to Match Papyrus Fragments." *Proceedings of the 5th International Workshop on Historical Document Imaging and Processing* (Sydney NSW Australia), September, 78–83. <https://doi.org/10.1145/3352631.3352646>.
- Pirrone, Antoine, Marie Beurton-Aimar, and Nicholas Journet. 2021. "Self-Supervised Deep Metric Learning for Ancient Papyrus Fragments Retrieval." *International Journal on Document Analysis and Recognition (IJDAR)* 24 (3): 219–34. <https://doi.org/10.1007/s10032-021-00369-1>.
- Vu, Manh Tu, and Marie Beurton-Aimar. 2024. "ViT-ED: Transformer Network for Image Similarity Measurement." In *Document Analysis and Recognition – ICDAR 2024*, edited by Elisa H. Barney Smith, Marcus Liwicki, and Liangrui Peng. Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-70546-5_18.

Modelling Early Modern Reasoning for the Venetian Risposte: a Scalable Framework

Emmanuela Carbé Ca' Foscari University of Venice – VeDPH

Federico Boschetti ILC CNR Pisa – VeDPH, Italy

Isabella Cecchini ISEM CNR Rome, Italy

emmanuela.carbe@unive.it, federico.boschetti@ilc.cnr.it

Policy reasoning, Data Modelling, Early modern Venice, Archival series, HTR

Abstract

The documentary series *Risposte* (replies), preserved in the Archivio di Stato di Venezia, forms an almost continuous collection of advisory opinions drafted by the Cinque Savi alla Mercanzia for the Venetian Senate (1540-1733). The paper presents a pilot study that treats the *Risposte* as structured units of policy reasoning that can be modelled and analysed at scale. Focusing on register 142 (1607-1610), for which a complete digitisation and a subset of manual diplomatic transcriptions are available, we define the single *Risposta* as the primary decision unit and design a two-layer data model: a minimal documentary core plus an analytical layer with controlled vocabularies (policy domain, decision orientation, triggers, actors, geography). Institutional variants are encoded as explicit relations rather than treated as exceptions. To support scalability, we report a first evaluation of an automated text recognition (ATR) workflow bridging Transkribus and eScriptorium through the XML-PAGE interchange format, and we prototype lightweight rule-based procedures (deontic formula extraction, monetary expressions, similarity measures) for semi-automatic population and quality control. The pilot operationalizes two directions of scalability: a stable decision-unit model and an interoperable transcription pipeline to populate it, enabling future large-scale

analysis of institutional coherence, policy drift, and trade oversight in early modern Venice.

Introduction

The documentary series *Risposte* (replies), preserved in the Archivio di Stato di Venezia, constitutes an almost continuous collection of advisory opinions drafted by the Cinque Savi alla Mercanzia for the Venetian Senate between 1540 and 1733 (Borgherini-Scarabellin 1925). Established as a permanent magistracy in 1517, the Cinque Savi operated at the intersection of mercantile practice and governmental decision-making, addressing matters of trade regulation, citizenship privileges, consular tariffs, monetary stability, and fiscal enforcement.

The period between approximately 1570 and 1670 witnessed profound geopolitical transformations that reshaped Mediterranean trade networks and challenged Venice's commercial position. Economic historians have long drawn on the *Risposte* to document these changes (Fusaro 2015; Pezzolo 2013; Sella 1968, 1994); Borgherini-Scarabellin (1925) remains the only monograph devoted to the magistracy, and the series itself still awaits a dedicated modern study and representation as structured data. Rather than treating the *Risposte* solely as narrative evidence of economic change, this project approaches them as structured units of policy reasoning and investigates whether the series can be modelled as a coherent governance system (de Vivo 2007).

This defines a specific data-modelling problem (Flanders & Jannidis 2019). Recent computational work on Venetian serial sources operates on records whose structure the source itself supplies: the 1808 cadastre pairs every parcel with fixed fields for number, owner, function and area (di Lenardo et al. 2025; Karch et al. 2025), and the *Garzoni* corpus of some 55,000 apprenticeship contracts records master, apprentice, trade and term in equally regular slots (Colavizza 2017). A *Risposta* offers far less: a date, a marginal note, the signatures of the Savi, and between them an argued opinion. The note is a scribal finding aid, not a classification: it points to a subject, while the decision itself, its orientation and its conditions, exists only in the prose, in formulas such as *stimiamo* (we deem), *debba* (must) or *non possa* (may not) and in the argument built around them. That problem is

the difference between structure supplied by the source and structure read into it.

Proof of Concept and Data Model

The proof of concept is built on register 142 (1607-1610), for which a complete photographic digitisation and a subset of manual diplomatic transcriptions are available. On this material the pilot tests two complementary directions of scalability: model scalability, through a stable decision-unit model with a minimal schema, and quantity scalability, through an ATR workflow that populates the model at scale.

The pilot comprises four components: the decision-unit data model with its controlled vocabularies; a structured dataset of 61 *Risposte*, with missing folio ranges modelled as first-class gap rows (register coverage $\approx 51\%$); the manual diplomatic transcriptions, which constitute the ground truth for ATR alignment and evaluation; and the rule-based extraction and quality-control procedures. Beyond the pilot lie the enlargement of the manually corrected ground truth, the training of a recognition model for the untranscribed portions, and the extension of the workflow to the full series.

The register presents three recurrent documentary markers: a date (including the intra-register *detto*, *ditto*), a marginal topic note (often present), and concluding signatories. These regularities ground rule-based segmentation and high-yield extraction. We define the *Risposta* as the primary decision unit and encode its recurring institutional variants (dissent, continuations) as explicit, typed relations, placing the model within the event- and relation-centric tradition of cultural-heritage data modelling (Doerr 2003).

The pilot also shows the limits of treating the unit as atomic. A single signed text may resolve several requests, quote its trigger (*supplica*: the petition), attach conditions to a grant, or speak with an individual rather than a collegial voice. The unit is nonetheless retained, on two grounds: the signed *Risposta* is the institutional act the Senate received, and in the domain expert's assessment it addresses a single matter in the large majority of cases. Composite units are therefore recorded as flagged exceptions within the unit; the planned refinement of the schema separates the documentary container from the deliberative acts it hosts, each with its own orientation and signatories.

The schema separates two layers. The minimal documentary layer captures source-near features required for segmentation, alignment, and automated population: *register_id*; *unit_id* *folio_start/end*; *date_original* and *date_iso*, *marginal_note* (with presence flag); *signatories_raw/norm* (with count/completeness); *text_diplomatic*; and *transcription_status* (*manual_full* / *manual_partial* / *regist* / *not_transcribed*). The relation vocabulary provides for *dissent_to*, *continuation_of* and *repeated_supplication*: a dissenting opinion forms a unit of its own, linked to the unit it contests, and a continuation points back to the block it extends. In register 142 the relation attested so far is *repeated_supplication*, linking the two entries of the Federici citizenship dossier. An analytical layer adds controlled vocabularies for research: *policy_domain*, *decision_orientation*, *document_trigger*, plus actors and geography for aggregation and network views. Values in this layer are hypotheses proposed by the analyst, recorded with an explicit uncertainty state and subject to expert validation. The two layers are released as two distinct datasets, joined on the unit identifier: the documentary dataset is stable and citable, while the analytical dataset is versioned separately, so that hypotheses can evolve without altering the record of the source.

Semi-automatic population and quality control rely on rule-based procedures in Python, a choice grounded in the register's language: stable and formulaic, it combines a Tuscan administrative base, a Venetian lexical component, and a Latin formulaic crust, a configuration that Venetian legal and administrative writing retains into the early eighteenth century (see Tomasin 2001, 2010). This stability makes high-yield extraction tractable with transparent rules. The pilot extracts dates, folio references, signatories, and monetary amounts and percentages by regular expression; institutional actors against a magistracy gazetteer; deontic formulas (*debba*: must, *non possa*: may not, *sotto pena*: under penalty, *stimiamo*: we deem) and a fiscal lexicon (*datij*: customs duties, *gravezze*: levies, *decima*: ten-per-cent tax, *tansa*: assessed levy, *ducatti*: ducats) by controlled-vocabulary matching; and the cited norms and precedents, the *parti* (the resolutions) and statutes the Savi invoke, which the pilot is extracting as the basis for a future relation graph. TF-IDF and cosine similarity, preferred to neural embeddings for their transparency and their adequacy to a corpus of this size, support exploratory clustering and quality control, surfacing candidate groupings

and outliers; they recover, for example, the two-entry Federici citizenship dossier as a most-similar pair. Reliability varies with the target: closed-class features (dates, monetary expressions, folio references) are high-yield targets, though orthographic variation and imperfect boundary detection still produce false positives, and all matches enter the dataset as candidates; interpretive features are structurally ambiguous, since bare modal verbs (*possa*: he may, *possano*: they may) can express normative permission or prohibition as well as epistemic possibility, and therefore enter the dataset as flagged hypotheses awaiting expert validation. The lexica, the controlled vocabularies, and the extracted hypotheses (each recorded with its rule identifier and its textual evidence) are available in the project's public repository; the extraction scripts are being consolidated for a subsequent release.¹

Scaling Transcription: ATR, Interoperability, and Model Training

The second aim of the pilot is to assess the feasibility of a scalable transcription methodology, based on the services provided by H2IOSC, for the series in its entirety, thereby enabling large-scale analysis of institutional coherence, policy drift, and trade oversight in early modern Venice.

The Humanities and Cultural Heritage Italian Open Science Cloud (H2IOSC) federates the national consortia of the principal European research infrastructures for Cultural Heritage (E-RIHS-IT), Digital Humanities (DARIAH-IT), Computational Linguistics (CLARIN-IT), and Open Science (OPERAS-IT). H2IOSC is currently developing and testing a textual workflow built upon the services provided by each infrastructure. The pipeline originates from digital images delivered via the IIIF protocol (<https://iiif.io>), proceeds through Automated Text Recognition (ATR – a unified framework encompassing both HTR and OCR) and linguistic analysis via UDPipe, and concludes with the integration of bibliographic references through Zotero.

¹ Project repository: <https://github.com/vedph/risposte-savi>; project site: <https://vedph.github.io/risposte-savi/>.

CLARIN-IT provides ATR services through the eScriptorium web platform (Kiessling et al. 2019; <https://gitlab.com/scripta/escriptorium>), originally developed by EPHE and now maintained with contributions from INRIA and other partners, as well as through new APIs integrated into the H2IOSC marketplace. To expand the range of ATR baseline models openly available to the scientific community, CLARIN-IT is implementing an interoperability strategy between Transkribus (Mühlberger et al. 2019; <https://www.transkribus.org>), accessed via a Team licence with a medium-scale credit allocation, and eScriptorium, connecting both platforms by routing outputs and inputs through the XML-PAGE interchange format. The two platforms play complementary roles. Transkribus currently offers higher out-of-the-box performance on generic documents through its supermodels (the Text Titan family, most recently Titan II), and is therefore a practical way to bootstrap a new transcription for manual correction; its models, however, are closed. eScriptorium, built on the open Kraken engine, allows project-specific models to be trained on the corrected material and released openly.

For this pilot the Text Titan I ter supermodel was employed; its output was imported into eScriptorium, where the automated transcription was aligned with the manual diplomatic transcription used as ground truth (Figure 1). The line-by-line segmentation model displayed an accuracy value of approximately 97%. For text recognition, a first project-specific model was created in eScriptorium from the aligned ground truth, which currently comprises twenty pages of diplomatic transcription: the model was trained on eighteen pages and evaluated on two held-out pages drawn from the same set, with no page used for both training and hyperparameter selection. The observed CER of $\approx 15\%$ should be read as a provisional, indicative figure rather than a stable estimate of model performance, due to the small size of the sample: extending the manually corrected ground truth is the immediate priority for more reliable results.

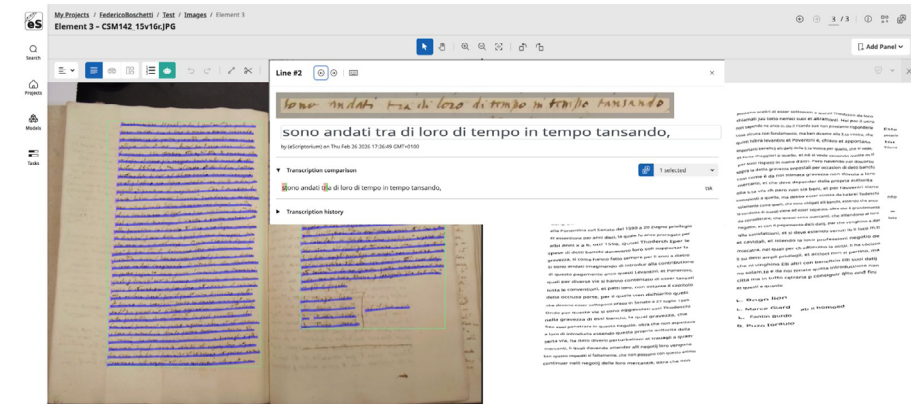


Figure 1: manual transcription merged with the automatic recognition (eScriptorium alignment view)

A new model will then be trained to recognise the portions of the manuscripts not yet manually transcribed. The model will be deposited in both the ILC4CLARIN repository and Zenodo, and its metadata will be made available through both the Virtual Language Observatory (VLO) of CLARIN and HTR-United.

Conclusions

The pilot delivers a documented model for segmenting and representing *Risposte* as decision units, a structured dataset comprising 61 records, built on manual diplomatic transcription as ground truth, a first evaluation of the ATR workflow (segmentation error $\approx 3\%$; text-recognition CER $\approx 15\%$), and the rule-based extraction and quality-control procedures released with the project repository. The next steps are the enlargement of the manually corrected ground truth, the training of a recognition model for the untranscribed portions, the act-level refinement of the schema, and the extension of the workflow to further registers. Together these operationalise the two directions of scalability: a stable decision-unit model (model scalability) and an interoperable transcription pipeline to populate it (quantity scalability).

Preliminary observations suggest recurring patterns across heterogeneous domains. The pilot generates a testable historical hypothesis: that decisions concerning citizenship, consular regulation, currency management, and maritime trade may share recurrent appeals to fiscal protection, harm containment, and procedural accountability. If confirmed on a larger and

independently validated corpus, these patterns would support an interpretation of the *Risposte* as components of a distributed governance architecture regulating economic flows. Extending the workflow to additional registers will make this hypothesis testable at scale, by quantifying segmentation accuracy, transcription quality and alignment effort, and precision and recall for high-yield extractions (dates, signatories, monetary expressions, deontic formulas).

Acknowledgements

The authors thank the Archivio di Stato di Venezia, CLARIN-IT and its national coordinator Monica Monachini for her support of the ATR workflow. The workflow relies on services developed within H2IOSC (Humanities and Cultural Heritage Italian Open Science Cloud), funded by the European Union – NextGenerationEU. Author contributions (CRediT): Emmanuela Carbé: conceptualization, methodology, data curation, software (project website and data-processing pipeline), writing – original draft, writing – review and editing. Federico Boschetti: methodology (ATR workflow), software, writing – original draft. Isabella Cecchini: investigation, resources (diplomatic transcription, photographic digitisation), validation, writing – original draft. The authors used generative AI to assist drafting, to prototype controlled vocabularies and extraction heuristics, and to develop the *Risposte* platform (Anthropic Claude Opus 4.8), with a final consistency check of the workflow (Anthropic Claude Fable 5); all outputs were reviewed and verified by the authors.

References

- Borgherini-Scarabellin, M. (1925). *Il Magistrato dei Cinque Savi alla Mercanzia dalla istituzione alla caduta della Repubblica. Studio storico su documenti d'archivio*. Venezia: R. Deputazione.
- Colavizza, G. (2017). A View on Venetian Apprenticeship through the Garzoni Database. In A. Bellavitis, M. Frank & V. Sapienza (eds.), *Garzoni. Apprendistato e formazione tra Venezia e l'Europa in età moderna*, 235-260. Mantova: Universitas Studiorum.
- de Vivo, F. (2007). *Information and Communication in Venice: Rethinking Early Modern Politics*. Oxford: Oxford University Press. DOI: 10.1093/acprof:oso/9780199227068.001.0001

- di Lenardo, I., Viaccoz, C., Guhenec, P., Musso, C. & Kaplan, F. (2025). The 1808 Napoleonic Land Registers of Venice: Spatial and Textual Dataset. *Journal of Open Humanities Data* 11: 67. DOI: 10.5334/johd.371
- Doerr, M. (2003). The CIDOC Conceptual Reference Module: An Ontological Approach to Semantic Interoperability of Metadata. *AI Magazine* 24(3): 75-92.
- Flanders, J. & Jannidis, F. (eds.) (2019). *The Shape of Data in Digital Humanities: Modeling Texts and Text-based Resources*. Abingdon, Oxon-New York: Routledge.
- Fusaro, M. (2015). *Political Economies of Empire in the Early Modern Mediterranean: The Decline of Venice and the Rise of England, 1450-1700*. Cambridge: Cambridge University Press.
- Karch, T., Saydaliev, J., di Lenardo, I. & Kaplan, F. (2025). LLM Agents for Interactive Exploration of Historical Cadastre Data: Framework and Application to Venice. *Computational Humanities Research* 1: e11. DOI:10.1017/chr.2025.10014
- Kiessling, B., Tissot, R., Stokes, P. & Stökl Ben Ezra, D. (2019). eScriptorium: An Open Source Platform for Historical Document Analysis. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, vol. 2, 19. IEEE. DOI: 10.1109/ICDARW.2019.10032
- Mühlberger, G. et al. (2019). Transforming Scholarship in the Archives through Handwritten Text Recognition: Transkribus as a Case Study. *Journal of Documentation* 75(5): 954-976. DOI: 10.1108/JD-07-2018-0114.
- Pezzolo, L. (2013). The Venetian Economy. In E. Dursteler (ed.), *A Companion to Venetian History, 1400-1797* (Brill's Companions to European History 4), 255-289. Leiden-Boston: Brill. DOI: 10.1163/9789004252523_007
- Sella, D. (1968). Crisis and Transformation in Venetian Trade. In B. Pullan (ed.), *Crisis and Change in the Venetian Economy in the Sixteenth and Seventeenth Centuries*, 88-105. London: Methuen.
- Sella, D. (1994). L'economia. In G. Cozzi & P. Prodi (eds.), *Storia di Venezia, VI: Dal Rinascimento al Barocco*, 651-711. Roma: Istituto della Enciclopedia Italiana.
- Tomasin, L. (2001). *Il volgare e la legge. Storia linguistica del diritto veneziano (secoli XIII-XVIII)*. Padova: Esedra.
- Tomasin, L. (2010). *Storia linguistica di Venezia*. Roma: Carocci.

Programme Committee:

Anna Maria Marras, PC Chair, University of Turin

Marina Buzzoni, Ca' Foscari University of Venice

Elisa Cugliana, University of Cologne

Swantje Dogunke, University of Jena

Rafał Górski, Jagiellonian University

Eetu Mäkela, University of Helsinki

Christopher Nunn, Heidelberg University

Petr Plecháč, Czech Academy of Sciences

Vassilis Routsis, University College London

Joëlle Weiss, Trier University

Local Organizing Committee:

Joanna Byszuk-Podsadniuk, Institute of Polish Language of the Polish Academy of Sciences

Maciej Eder, Institute of Polish Language of the Polish Academy of Sciences

Agata Kwaśnicka-Janowicz, Jagiellonian University

Wojciech Łukasik, Jagiellonian University

Adam Pawłowski, The University of Wrocław

Jan Rybicki, Jagiellonian University

Aleksandra Rykowska, Jagiellonian University

Michał Woźniak, Institute of Polish Language of the Polish Academy of Sciences