Models of Data Extraction and Architecture in Relational Databases of Early Modern Private Political Archives

edited by
Dorit Raines

e-ISSN 2610-9093 ISSN 2610-9875

Studi di archivistica, bibliografia, paleografia 8

Edizioni
Ca'Foscari

Models of Data Extraction and Architecture in Relational Databases
of Early Modern Private Political Archives

Studi di archivistica, bibliografia e paleografia

Serie diretta da
Flavia De Rubeis
Dorit Raines

8

Edizioni
Ca'Foscari

Studi di archivistica, bibliografia, paleografia

Direzione e redazione
Dipartimento di Studi Umanistici
Palazzo Malcanton Marcorà
Dorsoduro 3484/D
30123 Venezia

Studi di archivistica, bibliografia, paleografia
e-ISSN 2610-9093
ISSN 2610-9875

URL https://edizionicafoscari.unive.it/it/edizioni/collane/studi-di-archivistica-bibliografia-paleografia/

Models of Data Extraction and Architecture in Relational Databases of Early Modern Private Political Archives

edited by Dorit Raines

Venezia
Edizioni Ca' Foscari - Venice University Press
2025

Models of Data Extraction and Architecture in Relational Databases of Early Modern Private Political Archives
edited by Dorit Raines

Certificazione scientifica delle Opere pubblicate da Edizioni Ca' Foscari: tutti i saggi qui raccolti hanno preliminarmente ottenuto il parere favorevole da parte di valutatori esperti della materia, attraverso un processo di revisione doppia anonima, sotto la responsabilità del Comitato scientifico della collana. La valutazione è stata condotta in aderenza ai criteri scientifici ed editoriali di Edizioni Ca' Foscari, ricorrendo all'utilizzo di apposita piattaforma.

Scientific certification of the works published by Edizioni Ca' Foscari: all essays published in this volume have received a favourable evaluation by subject-matter experts, through a double-blind peer review process under the responsibility of the Advisory board of the series. The evaluations were conducted in adherence to the scientific and editorial criteria established by Edizioni Ca' Foscari, using a dedicated platform.

Edizioni Ca' Foscari | Fondazione Università Ca' Foscari | Dorsoduro 3246 | 30123 Venezia
edizionicafoscari.unive.it | ecf@unive.it

1st edition May 2025
ISBN 978-88-6969-919-1 [ebook] | ISBN 978-88-6969-920-7 [print]

Cover design: Lorenzo Toso

The publication of this volume was made possible through the support of the Ministero dell'Università e della Ricerca, as part of the PRIN project 2017JMPYTA: *Papal Diplomacy and European Multi-denominational Societies Before the Thirty Years War.*

Models of Data Extraction and Architecture in Relational Databases of Early Modern Private Political Archives / edited by Dorit Raines — 1. ed. — Venezia: Edizioni Ca' Foscari, 2025 — viii + 296 pp.; 23 cm. — (Studi di archivistica, bibliografia, paleografia; 8). — ISBN 978-88-6969-920-7.

URL https://edizionicafoscari.unive.it/it/edizioni4/libri/978-88-6969-920-7/
DOI http://doi.org/10.30687/978-88-6969-919-1

Abstract

The essays included in this publication are penned by scholars renowned for their expertise in digital humanities and historical research, providing multidimensional insights into the evolving landscape of historiography. Through meticulous examination, they illustrate the transformative power of digital tools in reshaping the methodologies of historical inquiry, augmenting traditional practices with innovative approaches. By addressing these issues, scholars can better navigate the intricacies of historical narratives and contribute to a deeper understanding of the past.

Keywords Relational database. Data architecture. Historical research methodology. Digital tools. Historical narratives. Archival complexity. Metadata models.

Models of Data Extraction and Architecture in Relational Databases of Early Modern Private Political Archives
edited by Dorit Raines

Table of Contents

Introduction
Dorit Raines 3

PERSPECTIVES: HISTORICAL ARCHIVES AND DIGITAL HUMANITIES

The Digital Historiographic Turn and the Historian's Changing Toolkit: From 'Facts' and 'Events' to 'Datasets'
Dorit Raines 11

Is There a Reception of Algorithm-Based Research in Traditional Historical Scholarship?
Three Case Studies from Academic "Trading Zones"
Thomas Wallnig 47

The Representation of Historical Uncertainties as the Outcome of Competing and Incompatible Certainties
Fabio Vitali, Valentina Pasqual 61

Metapolis: Spatializing Histories Through Archival Sources
Lukas Klic 79

EXPERIENCES: HISTORICAL ARCHIVES, DATABASE AND ONLINE PUBLICATION

Including the Archival Context in the Historian's Materials: The Advantages of Archival Standard Databases in Historical Research
VINCULUM Project Database and Information System Guide
Maria de Lurdes Rosa 93

Cracking the Historical Code
From Unstructured Correspondence Corpora to Computational Analysis
Agata Bloch, Clodomir Santana,
Demival Vasques Filho, Michał Bojanowski 115

Methods and Tools of Quantification in Historical Research
Napoleonic Employment Applications as a Case Study
Valentina Dal Cin 143

Gendered Data in Medieval and Early Modern Sources
The *Gendered Networks* and *Digital Edgeworth Network* Projects
Máirín MacCarron 175

Extraction, Architecture and Recovery of Family Correspondence Data
The Platform "EpiCAT. Family Letters from Catalonia (Sixteenth-Nineteenth Centuries)"
Javier Antón Pelayo 193

CHALLENGES: GRAZIANI ARCHIVE AND OMEKA S

Historical Research and Archival Sciences in a Digital Perspective
Relational Database, Data Architecture and Data Extraction in Graziani Archives Portal
Dorit Raines 207

Reconciling Complex Historical Records with Omeka S Relational Database
The Case of the Graziani Archive
Gabriella Desideri 249

A Puzzle with Missing Pieces
Extracting, Deciphering, and Digitally Rearranging Data in Antonio Maria Graziani Private Archives
Carlo Baja Guarienti 269

How to Digitally Reconstruct the History of an Early Modern Private Library?
Antonio Maria Graziani (1537-1611) and the Vicissitudes of His Books
Luca Iori 281

Models of Data Extraction and Architecture in Relational Databases of Early Modern Private Political Archives

Introduction

Dorit Raines
Università Ca' Foscari Venezia, Italia

Over the past two centuries, the increasing adoption of scientific methodologies in historical studies has prompted debates among historians regarding the most suitable unit of measurement for analysing historical phenomena and structuring them coherently to uncover underlying processes. Central to this discourse is the distinction between 'facts' and 'events.' While a fact recorded in historical sources requires contextualization to attain historical significance, an event represents a network of interconnected facts that, despite being open to multiple interpretations, provides the necessary framework for historical analysis.

Recent technological advancements have shifted this debate toward data-driven methodologies. Historians now conceptualize historical data as discrete informational units categorized into typologies such as persons, places, objects, and institutions. Although these units, when stored in databases, often lose their original context, they can be aggregated and analysed statistically to reveal latent historical structures. Nevertheless, their transformation into historical facts remains contingent upon contextualization.

The emergence of relational databases – such as entity-relationship models and linked data – has further raised questions regarding the extent to which historical records can be accurately represented through structured data extraction. From a historiographical perspective, reconstructing historical contexts necessitates the development of complex data architectures capable of expressing the intricate relationships between data units. This approach reflects the evolving nature of historical methodology, integrating traditional historiographical concerns with contemporary digital humanities techniques.

Edizioni Ca'Foscari

Studi di archivistica, bibliografia, paleografia 8
e-ISSN 2610-9093 | ISSN 2610-9875
ISBN [ebook] 978-88-6969-919-1 | ISBN [print] 978-88-6969-920-7

Peer review | Open access
Submitted 2025-03-21 | Published 2025-05-22

DOI 10.30687/978-88-6969-919-1/000

Tim Hitchcock argued back in 2013 that academic historians had largely failed to respond effectively to challenges such as algorithm-driven discovery, misleading forms of search, poor OCR, and biased selection of sources. He suggested that while historians had preserved the form of scholarly good practice such as a critical use of evidence, they had ignored some important underlying principles that the 'keyword searching' culture imposed upon them (Hitchcock 2013, 9, 12, 14). The complexity of historical narration or the multidimensional reality of the past (Dedieu 2016, 1; Diaz-Ordoñez, Rodríguez Baena, Yun-Casalilla 2023, 1016) was challenged again and again in several projects striving to create 'structured historical narrations' by the limited binary capacity of data processors. It became clear that a database, even a relational one, had difficulties coping with a nonlinear narration expressed in prose and with sequential logic (Bodenhamer 2008, 224).

In 2020, the nationally funded project *Nuncio's Secret Archives: Papal Diplomacy and European Multidenominational Societies Before the Thirty Years' War* (2017JMPYTA) was launched with the aim of enhancing a stratified private political archive through the creation of an open-access research portal. The central objective was to develop a relational database representing the contents of the correspondence of two envoys of the Holy See to Eastern Europe in the second half of the sixteenth century, utilizing a 'keyword-searching' approach. However, as the team grappled with designing the database and defining the entities to be extracted, they realized that the well-established historical methodologies they had relied on clashed with the binary structure of the digital framework. This realization prompted a creative approach to overcoming the challenges encountered along the way.

In essence, the project raised the question of whether historians should critically reassess their scholarly practices and engage in broader methodological reflection – specifically, whether what Hitchcock refers to as 'keyword searching' can be challenged or at least refined (Hitchcock 2013, 14). The starting point was the recognition that traditional practices based on source criticism and classification must not only coexist with but also inform the evolving digital research methodologies. Accordingly, the team began by identifying the specific challenges posed by this integration and exploring potential solutions.

This intellectual inquiry culminated in an international conference held at Ca' Foscari University of Venice on 27-28 October 2022, aimed at examining the present and future relationships between historical research and relational databases. The conference specifically addressed the following topics:

- modelization of historical serial records for data extraction;
- models of historical data architecture in relational databases;

- different uses of historical structured data in relational databases;
- the future of relational databases for historical research.

The present publication is based on part of the papers presented at the conference and on other essays specifically requested to cover several topics emerging from the conference but not specifically addressed in it. The outcome provides multidimensional insights into the evolving landscape of historiography. Through meticulous examination, the essays illustrate the transformative power of digital tools in reshaping the methodologies of historical inquiry, augmenting traditional practices with innovative approaches.

The section *Perspectives* delves into the theoretical frameworks that underpin the integration of digital methodologies in historical research, offering a comprehensive analysis of the long-standing interplay between narrative contextualization and source authentication. It highlights the shift towards entity classification and the restructuring of data, ultimately facilitating the remodulation of historical narratives.

In *Experiences*, five historians share their firsthand encounters with digital technologies, detailing their initial expectations, the integration of sources within digital platforms, the challenges faced, and the solutions devised. This section underscores the practical implications of digital tools in historical research, revealing the dynamic relationship between historians and the digital realm.

Lastly, *Challenges* presents a case study unfolding the creation of the Graziani archives portal, and illustrating the complexities involved in constructing a digital repository. This section provides a step-by-step account from four distinct perspectives, encompassing data architecture, handling disambiguation and uncertainty, leveraging relational databases for enhanced historical evidence, and describing material culture within a relational database framework.

What emerges from the essays included in the present publication are several key problems and their corresponding solutions.

According to the essays included in the present publication the problems the historian encounters when using relational databases are:

- complexity of historical context – historical databases must maintain archival context and relational integrity, which is often neglected in favour of specific inquiries;
- metadata standards – existing metadata standards and relational databases do not fully capture the complexity of historical narratives;
- fragmented research process – numerous digital tools exist for spatio-temporal inquiries, but they often operate in isolation, leading to a fragmented research process.

- disambiguation of entities –the process of disambiguating entities can complicate documentation and standardization, impacting historical research. In fact, it is essential to understand the process that leads historians to identify or disambiguate an event, a person or a place;
- complexity of private archives – historians struggle with organizing unstructured data, particularly from fragmented private collections. These archives often have complexities such as lack of explicit attribution, missing information, and unknown dates or locations of composition;
- absence of long-term preservation infrastructures for relational databases portals – the lack of support and general services stems from the absence of shared common vision regarding the use and management of research data.

Several solutions were proposed by the scholars, based on their experience with relational databases:

- simulations for metadata extraction – employing simulations to extract metadata from archives results in quicker and more accurate identification compared to manual methods;
- structured data – the transition from unstructured to structured data through preprocessing techniques like content management, topic modelling, and social network analysis forms the basis for effective historical research;
- text mining techniques – integrating text mining with metadata about authors provides a robust method for investigating historical phenomena;
- critical approach to data – emphasizing the difference between 'data' (objective observations) and 'capta' (selected, interpreted information) helps maintain a critical approach to knowledge production.

In conclusion, while the challenges in historical data organization and analysis are manifold, the adoption of digital tools, structured approaches, and advanced methodologies offer promising solutions for more accurate and efficient historical research. Undoubtedly, non-digital historiography can benefit from algorithm-based research, focusing on structured data extraction and analysis. By addressing these issues, researchers can better navigate the intricacies of historical narratives and contribute to a deeper understanding of the past. Projects like the Graziani Archives facilitate digital rearrangement and analysis of data, addressing the inherent complexities and fostering easier identification and classification.

Bibliography

Bodenhamer, D.J. (2008). "History and Gis: Implications for the Discipline". Knowles, A.K. (ed.), *Placing history. How Maps, Spatial Data, and GIS Are Changing Historical Scholarship*. Bedlands (CA): ESRI Press, 219-33.

Dedieu, J.P. (2016). "Designing Databases for Historical Research with Special Reference to Fichoz". *Réseaux et Histoire*, 8 July.
https://reshist.hypotheses.org/files/2016/07/Dedieu_Historical_Databases.pdf

Diaz-Ordoñez, M.; Rodríguez Baena, D.S.; Yun-Casalilla, B. (2023). "A New Approach for the Construction of Historical Databases – NoSQL Document-Oriented Databases: The Example of *AtlantoCracies*". *Digital Scholarship in the Humanities*, 38(3), 1014-32.
https://doi.org/10.1093/llc/fqad033

Hitchcock, T. (2013). "Confronting the Digital or How Academic History Lost the Plot". *Cultural and Social History,* 10(1), 9-23
https://doi.org/10.2752/147800413X13515292098070

Perspectives: Historical Archives and Digital Humanities

Models of Data Extraction and Architecture in Relational Databases of Early Modern Private Political Archives
edited by Dorit Raines

The Digital Historiographic Turn and the Historian's Changing Toolkit: From 'Facts' and 'Events' to 'Datasets'

Dorit Raines
Università Ca' Foscari Venezia, Italia

Abstract This essay explores the impact of the digital revolution on historical research, traditionally grounded in 'facts' and 'events.' It first reviews shifts in historical sciences since the 1960s, influenced by social sciences' mid-twentieth-century adoption of mathematical tools and holistic methods for studying group behaviour. Historians developed their own analytical toolkit in response, blending classical concepts like 'fact' and 'event' with the social sciences' notion of 'actor.' Today's challenge is whether these tools suffice or need adaptation to leverage new technologies. Metadata of entities and properties offers a conceptual model for classifying occurrences into 'critical systems', where known variables frame analysis, and unforeseen reactions merit investigation. By linking these reactions to actors, historians could dynamically measure and interpret change within these systems, transforming their methodological approach.

Keywords Fact. Event. Structured data. Narratives of explanation. Historiographic turn.

Summary 1 The Historian's Toolkit. – 1.1 The 'Fact' and the 'Event'. – 1.2 Narratives of Explanation. – 1.3 The Archive – From 'Deposit of Authenticity' to 'Place of Knowledge'. – 2 The Impact of Social Sciences on the Historian's Toolkit. – 2.1 The 'Actor' and the 'Event'. – 2.2 Connecting the Actors in a Web of Relationships. – 2.3 Not Only Human – The Nature of the 'Actor'. – 3 Information Units and Datasets. – 3.1 Breaking the Event into Information Units. – 3.2 Modelizing Content. – 3.3 Object-Centric, Event-Centric and Complex Network Descriptions.

Edizioni Ca'Foscari

Studi di archivistica, bibliografia, paleografia 8
e-ISSN 2610-9093 | ISSN 2610-9875
ISBN [ebook] 978-88-6969-919-1 | ISBN [print] 978-88-6969-920-7

Peer review | Open access
Submitted 2024-09-05 | Accepted 2024-10-28 | Published 2025-05-22

DOI 10.30687/978-88-6969-919-1//001

1 The Historian's Toolkit

1.1 The 'Fact' and the 'Event'

Over the past two centuries, as historical studies have increasingly adopted a scientific approach,[1] historians have frequently debated the appropriate units of analysis for describing past episodes or phenomena and the most effective methods for organizing these elements into a coherent and logical framework to uncover the underlying processes that explain why events unfolded as they did (Little 2010, 5). The primary concepts proposed were the 'fact' and the 'event'.[2] In the case of the 'fact', it was not considered sufficient for a fact mentioned in historical records to be inherently of historical significance; a contextual framework was necessary to transform it into a historical fact.[3] Regarding the 'event', the French historian Paul Veyne had already observed that

> facts do not exist in isolation, in the sense that the fabric of history is what we shall call a plot, a very human and not very 'scientific' mixture of material causes, aims, and chances [...]. That plot is not necessarily arranged in chronological order; like an interior drama, it can unfold from one plane to another. (Veyne 1971, 32)

In essence, the event was understood as a complex of facts open to multiple interpretations, which ensured that each fact was situated within its appropriate context, thereby conferring historical significance:

1 Leopold von Ranke was perhaps the first historian to discuss the objectivity of facts, as evidenced by his famous phrase "wie es eigentlich gewesen" (simply to show how it really was), quoted by E.H. Carr from Ranke's preface to his *Histories of the Latin and Germanic Nations from 1494-1514* (*Geschichten der romanischen und germanischen Völker von 1494 bis 1514*), first published in 1824 (Carr 1961, 5). See Stern's (1956, 54-8) translation of the preface.

2 Carr was influenced by Talcott Parsons and his Theory of Action, particularly by Parsons' discussion of the meaning of 'fact'. Parsons cited Henderson's definition of a fact as an "empirically verifiable statement about phenomena in terms of a conceptual scheme" (Henderson 1932) and argued that not all facts are alike (Parsons 1937, 41-2). This observation was fully embraced by Carr, who argued that "the historian is necessarily selective" (1961, 9-10).

3 "The same fact is or is not historical according to the manner in which it is known. It is only the mode of acquiring knowledge that is historical" (Langlois, Seignobos 1904, 64). "The facts are available to the historian in documents, inscriptions, and so on, like fish on the fishmonger's slab. The historian collects them, takes them home, and cooks and serves them in whatever style appeals to him" (Carr 1961, 6). Perhaps Carr was influenced by Bloch's observation (although the former talks about facts and the latter speaks of documents): "the historian collects them, reads them, attempts to weigh their authenticity and truthfulness. Then, and only then, he makes use of them". However, it seems that Carr disregarded Bloch's conclusion: "no historian has ever worked in such a way" (1953, 64).

"Then what are the facts worthy of rousing the interest of the historian? – pursued Veyne – All depends on the plot chosen; in itself, a fact is not interesting or uninteresting" (Veyne 1971, 33; Roth 2020, 3). According to Veyne, the same principle applied to the 'event': "the historian never draws the map of the eventworthy – at the very most he can multiply the routes that cross it" (Veyne 1971, 34).

Veyne's concept of the "eventworthy" has been regarded as significant within the field of the philosophy of history. The historian's primary materials – facts and dates, or, as Mary Poovey distinguishes them, ancient and modern facts – do not permit extensive flexibility in interpretation, whether one is dealing with an interpretive account or a numerical table (Poovey 1998, 1-28). While the historian may hypothesize, surmise, speculate, suppose, and conjecture, these practices cannot form the foundation of the profession, as relying on them would be akin to building sandcastles. In essence, the historian requires solid evidence to differentiate between facts that are merely an unintentional 'twitch', as Gilbert Ryle described it, and a 'wink', a deliberate signal that initiates a sequence of actions – an eventworthy occurrence (Ryle 2009, 494; Geertz 1973, 6). The following example may illustrate this point **[tab. 1]**.

Table 1 A simulation of the meaning of 'fact' and 'event' according to the classical historians' toolkit

Statement	Nature of Statement	Historical Relevance
Emmanuel met Vladimir	A fact	Two individuals meet, no historical relevance.
Emmanuel Macron met Vladimir Putin	Historical fact?	May be considered historical fact only if we know beforehand the identity of these two individuals.
The French President Emmanuel Macron met Russian President Vladimir Putin	Historical fact	Lack of context; it is a historical fact but not yet an event.
The French President Emmanuel Macron met Russian President Vladimir Putin in Moscow on 7 February 2022	Historical fact	Lack of context; even though the date of the meeting is provided, a sign of a historical fact hinting to an event.
The French President Emmanuel Macron met Russian President Vladimir Putin in Moscow on 7 February 2022, leading Europe's diplomatic efforts to defuse Ukraine crisis. To Moscow, Macron has presented himself as a 'quality interlocutor", as Putin described him. The two men who set at the opposite sides of a long white table spent more than five hours locked in head-to-head talks.	Historical event	The context for the meeting is provided, thereby establishing a connection to the broader ongoing event – the Ukraine War.

The challenge of selecting facts and events to construct a plausible narrative, which so troubled Ranke and Bloch (Stern 1956, 54-8; Bloch 1953 [1949], 51), was addressed in 1962 by the American philosopher of aesthetics, Arthur C. Danto, with the introduction of the concept of "narrative sentences" (1962). Danto continued his research on this topic a few years later, formulating the idea that "narrative is a *form* of explanation" (Danto 1965, 233, 237). He grappled with the problem of how to connect two events through an explanation, noting that

> this connection is not a causal connection: rather, the events in question are connected as end-points of a temporally extended change [...] and it is the change thus indicated for which a cause is sought. (235)

Although Danto may not have been the first to consider the dynamic nature of history – Carr had addressed it a year earlier[4] – he undoubtedly drew attention to the issue of events accumulating in the "History container". Some of these events, he argued, are already "utterly impervious to modification", while others are still in the making as "time-extended entities" (Danto 1962, 150-1). In essence, certain past events only become significant due to future developments that confer meaning upon them. What is particularly intriguing about Danto's explanation is that the narrative is no longer composed of a succession of events, each representing a complex of facts assembled meaningfully, but rather of events that are connected as endpoints. This marks a preliminary step toward a new approach – still unformulated at the time – hinting at the concept of a network.

1.2 Narratives of Explanation

In the 1970s, the first signs of what would later be termed the Linguistic Turn began to emerge. Hayden White had already anticipated this shift in *Metahistory*, published in 1973. Setting aside the debate on ideology as a driving force behind the selection of facts and events in constructing historical narratives (White 1973, 1-29), White was focused on exploring the methods of writing history. Addressing Danto's concept of a "conceptual narrative", which refers to "a 'true' historical account of what actually happened in history [that] would be one which remained on the levels of synchronic classification of

4 "When, therefore, I spoke of history in an earlier lecture as a dialogue between past and present, I should rather have called it a dialogue between the events of the past and progressively emerging future ends" (Carr 1961, 164).

the data on the one hand and of diachronic representation of them on the other" (275), White proposed the nineteenth-century idea of "narrative representations" of the historical process. A key aspect of White's work is his analysis of the "linguistic protocol" employed by him to "prefigure the historical field prior to bringing to bear upon it the various 'explanatory' strategies used to fashion a 'story' out of the 'chronicle' of events contained in the historical record" (White 1973, 426). In doing so, White expanded upon Danto's notion that "narrative is a *form* of explanation". According to White, the selection of facts and events is closely tied to the narrative of explanation that the historian presents: there is "an elective affinity between the act of prefiguration of the historical field and the explanatory strategies used by the historian in a given work" (427). In 1987, White further elaborated on the two components of the historiographic process in *The Content of the Form*, where he argued that history becomes a form of communication, conceived as "a 'message' about a 'referent' (the past, historical events, and so on) the content of which is both 'information' (the 'facts') and an 'explanation' (the 'narrative' account)" (White 1987, 40).

White, in fact, set the stage for distinguishing between the 'materials' ('facts' and 'events') employed by the historian and the 'explanation' that is already shaped by the selection of these materials. As noted, White advanced this idea further in 1987 by addressing a previously overlooked aspect: the different methods of communicating the selection of these 'materials' – whether through chronological order, diachronic arrangement, or narrative explanation. The concept of linking the selection of materials to the construction of their meaning gained traction in the historiographic debate inspired by the Linguistic Turn. As early as 1981, historian William J. Bouwsma argued that "intellectual history requires more radical treatment than may be appropriate for other dimensions of historiography" (Bouwsma 1981, 279). Bouwsma identified the "construction of meaning" as a trend with the potential to reshape intellectual history, gradually transforming it into cultural history. From the historian's perspective, the "construction of meaning", as Bouwsma suggested, involved "an increasing concern with the location, the description, and perhaps the explanation of what passes for meaning in a variety of historical situations" (284). He recognized Geertz's concept of 'thick description' and developments in linguistics as emerging trends that could bring about what he viewed as a much-needed shift in historiography. However, Bouwsma lamented that historians had yet to seriously explore the connections between language and the perception of reality (289-90).

Six years later, concurrently with the publication of White's *The Content of the Form*, the Canadian historian John Toews responded to Bouwsma's observations, presenting – whether accurately or

not – Bouwsma's article as a call for a Linguistic Turn in intellectual history, a term Bouwsma himself had not used. The critical issue, according to Toews and others, was determining "which among a variety of linguistic theories of meaning a historian should choose" (Toews 1987, 881). Toews extended this argument by asserting, in response to Bouwsma, that "language not only shapes experienced reality but constitutes it" (882).

Toews may have pinpointed the core issue that historians have consistently faced with the Linguistic Turn. Even today, there remains considerable disagreement regarding its definition, purpose, and impact, particularly in the realm of historical research (Eley 1996, 193-4, 225; Loriga, Revel 2022, 15-19; Martinat 2023, 231-41). Jacques Revel and Sabina Loriga, in their analysis of the implications of this turn for the historical sciences, argue that it introduced turbulence into a field traditionally characterized by its stability (Loriga, Revel 2022, 10). They contend that, while the Linguistic Turn did not significantly affect European history, it had a more pronounced impact on American history (358). This assertion is partially true, as the Linguistic Turn's emphasis on categories such as 'culture,' 'women,' 'experience,' 'other,' and 'agency' as foundational to narrative construction did influence European history from the 1990s onward, particularly in the field of social history (Vernon 1994, 8-96; Mullaney 1996, 161).

Reviewing today the contemporary debate on the impact of discursive constructs across various disciplines, one gets the impression that this "mutually constitutive and open-ended dialectic, the relationship between literary and other cultural discourses, the discursive and the social" (Mullaney 1996, 162) facilitated greater permeability between disciplines and promoted interdisciplinarity. At the same time, it challenged the most deeply held beliefs within each discipline.

The Linguistic Turn posed two major questions for historians:

1. Whether the narrative categories proposed by the Linguistic Turn influenced the historian's selection of meaningful facts and events to accurately describe the historical process (Tucker 2004, 92; Loriga, Revel 2022, 15);
2. Whether historians were prepared to deprioritize the significance of documents (and archives) in establishing authenticity as a prerequisite for the historical narrative in favour of a more effective literary representation (Spiegel 2005, 2-7; Martinat 2023, 235).

In both cases, it appears that historians were not being asked to abandon the 'fact' or the 'event' as fundamental units in their toolkit, but rather to broaden their scope to include other types of investigative materials and to untangle the Gordian knot that linked documents and archives to authenticity.

Regarding the question of whether narrative categories influence the historian's selection of meaningful facts and events to accurately describe the historical process, historians were already turning to other disciplines, such as anthropology, to explore the role of the 'fact' as a fundamental unit of analysis in constructing explanatory narratives. As early as 1973, Clifford Geertz, with his concept of 'thick description', emphasized the centrality of facts in anthropological work:

> The aim is to draw large conclusions from small, but very densely textured facts; to support broad assertions about the role of culture in the construction of collective life by engaging them exactly with complex specifics. (Geertz 1973, 28)

In 1995, Geertz directly addressed the issue of 'fact' and 'event', though it remains unclear whether his ideas influenced historians like Koselleck or sociologists such as Latour. In *After the Fact*, Geertz delved into a central concern of the social sciences: how to construct a meaningful narrative by linking facts.[5] He raised a critical point, noting that the events being narrated refer to a situation that no longer exists by the time all the facts are gathered:

> To form my accounts of change, in my towns, my profession, my world, and myself, calls thus not for plotted narrative, measurement, reminiscence, or structural progression, and certainly not for graphs; though these have their uses (as do models and theorizings) in setting frames and defining issues. It calls for showing how particular events and unique occasions, an encounter here, a development there, can be woven together with a variety of facts and a battery of interpretations to produce a sense of how things go, have been going, and are likely to go. (Geertz 1995, 3)

Like Danto, Geertz viewed narratives not as a mechanical cause-and-effect chain linking one event to another but as dynamic occurrences capable of influencing other events.

Nine years after Geertz's *After the Fact*, Reinhart Koselleck, who curiously never references either Danto or Geertz, revisited the question of the nature of historical events and how to construct narratives based on them. In *Futures Past: On the Semantics of Historical*

5 "It is not history one is faced with, nor biography, but a confusion of histories, a swarm of biographies. There is order in it all of some sort, but it is the order of a squall or a street market: nothing metrical […] What we can construct, if we keep notes and survive, are hindsight accounts of the connectedness of things that seem to have happened: pieced-together patternings, after the fact" (Geertz 1995, 2).

Time, Koselleck examined the role of historical time as a means of ordering events, questioning its adequacy:

> The ordering and narration of events only has need of an exact chronology. But precise dating is only a prerequisite and does not determine the content of what may be called 'historical time'. (2004, 1)

In analysing the transformation of the concept of 'History' in the German language, Koselleck identified a shift that began in the 1770s and was further intensified by the French Revolution, which had a substantial impact on nineteenth-century historiography. This shift was from *Historie*, meaning an account of what had occurred, to *Geschichte*, which "principally signified an event, that is, the outcome of actions either undertaken or suffered" (Koselleck 2004, 32, 35; Carr 1961, 5; Roth 2020, 3).

In essence, philosophers and historians began to realize that history was far more complex than a simple chronological narrative based on historical time. Beyond 'facts' and 'events', it was necessary to consider ongoing processes or "processual contexts" – from the past to the future (Koselleck 2004, 95). It was precisely these two levels – events that can only be narrated and structures that can only be described – that allowed Koselleck to treat events as measurable units, discussing their irreversibility, their repeatability, and, most importantly, "a differential classification of historical sequences" that can be measured against one another (95, 105). The mere chronology of events was no longer sufficient; to accurately describe the complexity of events, their diachronic structure was essential (106, 108).

Koselleck, in his effort to make historical analysis more complex and precise in determining the role of each historical element, laid the groundwork for a new methodological approach. The debate over 'fact' and 'event' became integral to historians' development of a toolkit. By assuming that the historical narrative reveals the elements that drive change, the first step in constructing a set of analytical tools was to choose a stable unit to measure ongoing developments: the fact and its more complex form, the event. This approach led historians to deconstruct the narrative into a series of smaller, basic units, which were preferably measurable or at least classifiable, enabling them to be reassembled according to the logic of classification.

Regarding White's second question – whether historians were prepared to deprioritize the importance of documents (and archives) in establishing authenticity as a prerequisite for the historical narrative in favour of a more effective literary representation – most historians in the 1990s and 2000s continued to hold tightly to records as the guarantor of authenticity. However, Postmodernism has prompted a rethinking of archives, leading to what is now referred to as the Archival Turn.

1.3 The Archive – From 'Deposit of Authenticity' to 'Place of Knowledge'

With the advent of Source Criticism and its criteria of validity, reliability, and relevance to the subject under investigation, reliability became closely associated with the world of archives.[6] Following the principle of *Respect des fonds*, archives ensured that all documents contained within them maintained their context and authenticity (Langlois, Seignobos 1904, 36-7; Duchein 1983, 64-82). However, archival science has evolved, considering as reliable not only texts but also any kind of artifact or remnant from the past. These new inclusions necessitated a different set of criteria to establish their relevance, authenticity, and validity (Howell, Prevenier 2001, 17-20). Postmodernism posed a significant challenge to the established principles of historians. As early as 1969, Michel Foucault viewed the archive as the ordering principle guiding the production of knowledge. According to him,

> the archive is […] the system that governs the appearance of statements as unique events. But the archive is also […] all […] things said […] grouped together in distinct figures, composed together in accordance with multiple relations, maintained or blurred in accordance with specific regularities. (Foucault 2004, 101)

Foucault, it seems, was interested not in the records themselves but in the statements, which he viewed as events in their own right. This perspective marked a significant departure from the notion of the archive as a historicized repository of knowledge about the human past: the archive, according to Foucault, "does not have the weight of tradition" (102).

The long-standing dominance of Leopold von Ranke's archival practices, which emphasized the framing of supposedly neutral records, began to give way to the belief that the mechanisms of collecting and preserving records actively shaped the historical process, and consequently, its historiographic analysis and narrative. Archives, once considered merely as places of enquiry, became subjects of enquiry themselves (Carr 1961, 6, 15-16; Rolnik 2012, 215). In *Birth of the Archive*, Markus Friedrich refers to early modern archives as "places of knowledge", asserting that "archives […] were multivalent elements of a multifaceted culture and society" (2018, 4).

6 Source Criticism enumerates the following criteria to establish the degree of faithfulness of the source to past events: the document's 'genealogy', its genesis, originality, interpretation, authority, and finally, the competence and trustworthiness of the observer (Howell, Prevenier 2001, 60-8).

This suggests that narrative forms themselves played a role in shaping how societies processed their pasts. This concept was foundational for cultural studies, which in the 1990s introduced the notion of the 'Archival Turn'. This idea was further expanded upon by fields such as anthropology, colonial studies, and art and image history, particularly in relation to archives.[7]

At this juncture, historians could no longer treat archives solely as places of provenance and authentication; they were now required to evaluate the formation and development of records within the archive and to integrate them into the historiographic narrative, assigning them a role comparable to that of other historical actors.

The Archival Turn, which invites us to place archives at the centre of scholarly inquiry, also reveals a deeper connection to historiographic research. The development of archival practices, which gradually evolved into what we now call archival science, shares a common approach with historiography in terms of the relationship between fact and event: a document, by itself, does not constitute meaningful testimony to the past; it requires context to be considered reliable evidence of historical facts and events. Ultimately, the positivist perspective that has dominated our understanding of the physical world established context within a hierarchical system (fact-event; record-archival series) and assigned the creator the limited role of provenance. In contrast, historians and archivists were permitted – albeit within a rigid framework of rules – to re-arrange and evaluate the material. This process was believed to guarantee "impartiality, authenticity, immutability, reliability, evidentiality, integrity, truth, authority, accuracy, order, uniqueness, and trustworthiness" (Lane, Hill 2010, 4).

Postmodernism challenged this worldview by rejecting the notion of a unified society progressing towards a single, all-encompassing goal. Instead, it embraced Jean-François Lyotard's concept of "incredulity towards metanarratives", viewing these grand narratives as mere mechanisms of legitimation, with no bearing on what is true or just (Lyotard 1984, xxiv-xxv). In the archival context, Postmodernism influenced the development of a new perspective: archives, once seen as places of fixed and stable meaning, have become contested spaces where meanings are hidden, subverted, altered, or absent (Lane, Hill 2010, 7). It is not surprising, then, that Bruno Latour, already recognizing the potential dangers of this shift, called for a renewal of empiricism in 2004: "The question was never to get *away* from facts but *closer* to them, not fighting empiricism but, on the contrary, renewing empiricism" (Latour 2004, 231).

7 Simon 2002, 101-2; Stoler 2009, 44-51; Hölling 2015, 73-81; Callahan 2024, 74-88.

This situation, however, has led to a "new divide" between history and archives, which threatens to impact future historical research (Blouin, Rosenberg 2011, 13-93). Records were no longer preserved as evidence of specific transactions but were instead archived in clusters that reflected complex social and political processes. These records encompassed materials that were evaluative, descriptive, prescriptive, or advisory, depending on the intent and function of their creators. The archival logic used to organize items such as correspondence, orders, decrees, individual diaries, architectural plans, and similar documents following the understanding of their context within organizational processes: the agency, function, or, in some cases, the particular significance of the individual who generated them. The need for archivists to address the growth of multiple archival constituencies, particularly in response to changes in the private archival environment, led to the idea that standardization was the key to shared archival access. This standardization was based on the creation of measurable units or entities. The construction of these data archives was guided by the way historians interpreted the data, which in turn formed authoritative coding practices and created the categories under which the data could be archived and retrieved. In short, the Archival Turn, inevitably altered by the disruptive nature of digital record production, also began to re-examine historical material archives, applying this new logic to them as well.

2 The Impact of Social Sciences on the Historian's Toolkit

2.1 The 'Actor' and the 'Event'

As early as the 1930s, inspired by the Chicago School's Systems theory, Harvard professor Talcott Parsons developed the Theory of Action, which immediately attracted the interest of historians. Indeed, historians were also grappling with the challenge of defining the appropriate unit of measure to detect change and describe the ongoing relationships among all the elements or actors in an unfolding narrative. Parsons' theory perceived human action by moving beyond the traditional categories of social groups, hierarchy, power and control, social organization, and role distribution. Instead, Parsons selected what he called the "unit act" as the fundamental unit, characterized by four features:

1. the agent or actor;
2. the end, objective, or goal of the action;
3. the conditions or means of action – specifically, the situations in which the actor does or does not have control over his action;
4. the range of choices available to the actor regarding his action (Parsons 1937, 43-4).

The following table illustrates the hypothetical benefits historians might derive from the Theory of Action **[tab. 2]**.

Table 2 A simulation of an historian's reading into an event according to the Theory of Action

Statement	Nature of Statement	Historical Relevance
Emmanuel met Vladimir	Insufficient information	Two individuals meet.
Emmanuel Macron met Vladimir Putin	Insufficient information	No information is provided regarding the nature of the exchange nor of the role of the actors.
The French President Emmanuel Macron met Russian President Vladimir Putin	Insufficient information	Even if we know in advance who these two actors are, we lack information regarding the nature of the exchange or their respective roles.
The French President Emmanuel Macron met Russian President Vladimir Putin in Moscow on 7 February 2022	Insufficient information	no context is provided even though we have the date of the meeting.
The French President Emmanuel Macron met Russian President Vladimir Putin in Moscow on 7 February 2022, leading Europe's diplomatic efforts to defuse Ukraine crisis	Insufficient information	The context for the meeting is provided, but we lack information regarding the nature of the exchange, despite its connection to a broader event – the Ukraine War.
The French President Emmanuel Macron met Russian President Vladimir Putin in Moscow on 7 February 2022, leading Europe's diplomatic efforts to defuse Ukraine crisis. To Moscow, Macron has presented himself as a "quality interlocutor", as Putin described him. The two men who set at the opposite sides of a long white table spent more than five hours locked in head-to-head talks.	Exchange between the two leaders (actors) as part of a historical event	The Theory of Action focuses on the exchange between the two interlocutors, which imparts meaning to an event. The statement describes only a portion of the exchange between the two actors. It is evident that they decided to meet to discuss solutions for the Ukraine crisis.

The Theory of Action seeks to trace the individual across all of his relationships, analysing the added value and contribution of these relationships to social dynamics. It views the individual as a social actor embedded within a web of various relationships, making decisions about the degree, intensity, and type of attachment based on the 'social capital' that each relationship may offer. Whether we agree or disagree with Parsons' analysis, his theory had a profound impact on the social sciences during the 1950s and 1960s. Although it was eventually set aside in favour of parallel theories – such as Herbert Hyman's *The Psychology of Status* (1980) and Muzafer Sherif's work on reference and membership groups (1968) – and by more complex ideas that aimed to reinvent functionalism, the Theory of Action made a comeback. By the late 1970s, it reemerged in a different form, and in the 1980s, it renewed interest in the actor and his 'social capital,' as exemplified by Pierre Bourdieu in *Distinction* (1984) and in interaction systems.

One of the fundamental concepts underlying the Theory of Action is the concept of 'exchange' (Parsons 1937, 98-101, 460-70), which was previously explored by Durkheim in his studies on exchange and property [1895] (1982) and later reintroduced by George C. Homans in 1951 in his *The Human Group* with his social exchange theory (268-72). While economists primarily focused on economic exchange, the market, and the economy, sociologists incorporated additional influences into their analysis of economic action. Homans observed that if an action had previously resulted in an advantage, it is more likely to be repeated, as its potential social capital has been demonstrated. Furthermore, interaction between two or more actors can either facilitate, deflect, or constrain individuals' actions in the market. For example, a long-standing friendship between a buyer and a seller might prevent the buyer from abandoning the seller simply because a similar item is available at a lower price elsewhere in the market. Economists, recognizing the value of social studies, began applying behavioural patterns to the economic field (Smelser, Swedberg 1994, 3-26).

This interdisciplinary approach was further advanced through historical sociology, as analysed by Dennis Smith in his 1991 work *The Rise of Historical Sociology*, where he argues that historical sociology seeks to uncover "the mechanisms through which societies change or reproduce themselves" (Smith 1991, 1). Likewise, economic sociology, whose origins lie in the works of Max Weber (particularly his concept of 'social action' in *Economy and Society*) (Weber 1978, 1375-80) and Durkheim (with his concept of 'organic solidarity' in *The Division of Labor in Society*) (Durkheim 1947, 68-87; 149-75), views markets – whether of money, labour, or goods – as social structures shaped over time by networks of social actors, organizations, and national cultures (Smelser, Swedberg 1994, 3-26; Beckert 2009, 246).

The potential of this approach was quickly recognized. The economic domain, which sought to apply the frames of reference, variables, and explanatory models of sociology to the complex activities involved in the production, distribution, exchange, and consumption of scarce goods and services, integrated the Theory of Action's emphasis on the actor and his social capital, ultimately leading to the development of Network Theory.

2.2 Connecting the Actors in a Web of Relationships

The concept of a network had already been explored in the 1930s by a group of German psychologists led by Jacob Moreno, who was interested in Gestalt psychology and was the first to draw a 'sociogram'. The idea gained further momentum in 1953 when psychologist Dorwin Cartwright and mathematician Frank Harary connected network theory with graph theory, ultimately leading to the development of network models (Moreno 1934, 32-3; Cartwright, Harary 1956, 277-93; Scott 2012, 8-11). A significant advancement occurred in 1976 with the publication of an article in the *American Journal of Sociology* titled "Social Structure from Multiple Networks: Part I. Blockmodels of Roles and Positions", followed shortly by "Part II. Role Structures" (White, Boorman, Breiger 1976, 730-80; Boorman, White 1976, 1384-446). This work was authored by Harrison White and two of his students. White, who had been associated with the Chicago School and, since 1963, served as an associate professor of sociology in the Harvard Department of Social Relations – the same department as Talcott Parsons – emerged as the leader of the 'Harvard Revolution' in social network analysis with this publication.

Several years earlier, White had conducted an analysis on a model of social mobility, which he published in 1970 in the book *Chains of Opportunity*. In this work, he demonstrated that the interdependence between patterns of relationships and an individual's position could be described through a quantitative analysis of social roles without relying on statistical aggregates.

The assumption was that to study the structure of social action, it would be necessary to describe the stable exchange modes in Durkheim's sense (or, as White termed them, "stable interaction patterns"; White, Boorman, Breiger 1976, 756) among actors – essentially the foundation of the network approach. The notion that all structures, whether historical or social in nature, are composed of basic components (facts or actors) that interact on various levels, and even across time and space (763), was a precursor to – or a reflection of – a philosophical approach that was gradually evolving into what we now refer to as Hyperconnectivity and/or Complexity. This

approach, in simplified terms, concerns the behaviour of a system or model whose components interact across multiple levels of relations.

Network theory entered historical studies primarily through the field of historical demography. As early as 1987-88, during the 'tournant critique' of the Annales school – a period of re-evaluation of the theories of Ernest Labrousse and the *Nouvelle histoire* (Burguière 2009, 128-31) – the French academic world, particularly in the social sciences, experienced what Gérard Noiriel termed a 'crise de l'histoire' (Noiriel 1996). This period saw a renewed interest in the Theory of Action, as well as in micro-history and world history, although networks and their social roles were studied by only a few scholars. Historical demography was among the first to focus on historical networks, particularly addressing the issue of 'family networks' ("réseaux de parenté"), as seen in the 1995 issue of the *Annales de démographie historique*, which was dedicated to this theme (Nassiet 1995, 105-23).

Historical studies only began to recognize the necessity of theoretically discussing the role of historical networks about 25 years ago. In the 1998 publication *Réseaux, familles et pouvoirs dans le monde ibérique à la fin de l'Ancien Régime*, Jean-Pierre Dedieu and Zacarias Moutoukias emphasized in the preface the issue of the "content and volume of messages", noting that these factors "depend on the nature of the links between senders and recipients, which are, in turn, influenced by all other existing interpersonal relationships within the network".[8] However, while acknowledging that network theory was "under the current circumstances, the only available tool" to "formalize, […] find a framework that allows for grouping and making sense of the multiplicity of fragmented observations",[9] the two scholars cautioned against the 'metaphoric use' ("utilisation métaphorique") of networks (Dedieu, Moutoukias 1996, 7-30).

The internet, with its network structure, has quickly produced significant repercussions across various aspects of social life (e.g., family, relationships, work, information dissemination), extending the network model to virtually all possible domains.[10] Meanwhile, French historical demography continued its research in this area. Its leading journal, *Annales de démographie historique*, dedicated a special

8 "Contenu et volume des messages" which "dépendent de la nature des liens entre expéditeurs et destinateurs, et que ceux-ci sont à leur tour affectés par l'ensemble des autres relations interindividuelles existantes dans le réseau".

9 "Dans les circonstances actuelles, le seul outil disponible" to "formaliser, […] trouver un cadre qui permette de regrouper et de donner sens à la multiplicité des observations fragmentaires".

10 Lee Rainie and Barry Wellman (2012, 21-57) speak of the "social network revolution" which refers not to a technological shift, but to a relational shift, in which networks – rather than groups – become the systems of support.

issue to "Histoire de la famille et analyse de réseaux" in 2005, and another to "Les réseaux de parenté, refonder l'analyse" in 2008. It was only in this latter issue that network analysis was truly adopted, yielding some interesting results. However, this represented only a small step forward; Network theory required a more practical approach. In February 2010, Michel Bertrand, Claire Lemercier, and Sandro Guzzi published what could be considered a manifesto, in which they observed:

> Historians' discussions of networks refer to a wide range of authors and, in particular, to various theories depending on the period or sub-discipline in question – such as the history of family and marriage, business and boards of directors, sciences or literature, and correspondence or citations [...] – as well as other social sciences they engage with. There is therefore no 'school' of network analysis in history, nor is there often a dialogue between those who have found a certain fertility in this concept.[11]

The questions raised by the three scholars were as follows:

1. the existence and nature of a unified approach to both systematic and qualitative analysis;
2. the presence of particular challenges associated with historical sources that might complicate or condition the study of networks;
3. the scale of network analysis and its adaptation to historical research.[12]

From the outset, the challenge appeared to be not only how to calculate intensity, density, length, or other metrics in Network theory concerning long-gone societies, but also how to schematically and visually represent the web of relationships between historical social actors. Sociology and historical demography have already addressed this problem through the use of sociograms and connected graphs. However, the task of the historian is different: historical research seeks to enumerate and then analyse the circumstances that led to, and resulted in, the creation of one or more networks, with the aim

11 "Les discussions historiennes des réseaux se réfèrent à des auteurs, et en particulier à des théories, très variés selon la période ou la sous-discipline concernée – histoire de la famille et de l'alliance, de l'entreprise et des conseils d'administration, des sciences ou des lettres et des correspondances ou citations... – et les autres sciences sociales fréquentées. Il n'existe donc pas d''école' d'analyse de réseaux en histoire, ni même le plus souvent de dialogue entre ceux et celles qui ont trouvé une certaine fécondité à ce concept".

12 A question probably inspired by the micro-history approach and its "principe de la variation d'échelle" (Revel 1996).

of dynamically describing their development, influence on social actors, and evolution over time. Historians do not aim to create models or schematically simplify the description of any given network. On the contrary, they seek to explore its complexity, ramifications, and interactions with other realities or factors that may be deemed relevant, in order to situate this phenomenon within a broader context. This approach does not imply viewing social actors as mere expressions of structures, as was common in classical social history, nor does it attempt to align with 'micro' or 'macro' perspectives. The problem with the concept of 'network' lies in treating it as a "category of analysis", which is misleading since it is an unstable, fluctuating, and ephemeral system. Moreover, its scope may shift depending on the social actor's perspective. José Maria Imízcoz Beunza, one of the foremost historians studying Network analysis, has already cautioned against the danger of inadvertently using "categories of analysis" that are, in fact, social actors themselves (Imízcoz Beunza 2011, 23; Imízcoz Beunza, Arroyo Ruiz 2011, 133). Even when historians attempt to trace the action and determine its trajectory, the task is challenging, particularly when considering the difficulty of defining the boundaries of historical networks.

2.3 Not Only Human – The Nature of the 'Actor'

Independent of the ongoing debate on Network analysis, the well-known French sociologist Bruno Latour introduced his Actor-Network Theory (ANT). As early as 1999, Latour published an article titled "Factures/Fractures: From the Concept of Network to the Concept of Attachment", in which he argued that while

> networks are extremely efficacious in redistributing force, they are not at all effective in renewing a theory of action specific to each of the nodes. […] The move towards the network of attachments should permit us to keep the distributive effects of the network, while at the same time enabling us to entirely reconceptualize the nature and source of action. (Latour 1999, 31)

By the concept of 'attachment', he referred to "the formidable proliferation of objects, properties, beings, fears, techniques that make us do things unto others" (24). Latour further developed these ideas in 2005 with *Reassembling the Social: An Introduction to Actor-Network-Theory*. ANT posits the 'principle of generalized symmetry', which asserts that both human and non-human entities (e.g., artifacts, organizational structures) should be incorporated into the same conceptual framework and granted equal agency. According to Latour, this approach allows for a detailed description of the concrete

mechanisms that hold the network together, while ensuring an impartial treatment of the actors involved (Latour 2005, 109-15). Within the flat network of associations that ANT scholars trace, every actor is considered a mediator, translating one type of information into another; thus, there is only one scale but multiple levels of relations (128-33). By using the term 'network', Latour did not intend to adopt the Network Analysis approach, which he found inadequate for describing the complexity of exchanges and their nature: "[Network] is a tool to help describe something, not what is being described" (131). Here is an example of the novelty introduced to historians by Actor-Network Theory **[tab. 3]**.

Table 3 A simulation of an historian's reading into an event according to the Actor-Network Theory

Statement	Nature of Statement	Historical Relevance
Emmanuel met Vladimir.	Insufficient information	Two individuals meet.
Emmanuel Macron met Vladimir Putin.	Insufficient information	No information is provided regarding the nature of the exchange or the role of the actors.
The French President Emmanuel Macron met Russian President Vladimir Putin.	Insufficient information	Even if we know beforehand who these two individuals are, we have no information on the nature of the exchange nor of the role of the actors
The French President Emmanuel Macron met Russian President Vladimir Putin in Moscow on 7 February 2022.	Insufficient information	No context is provided, even though the date of the meeting is known.
The French President Emmanuel Macron met Russian President Vladimir Putin in Moscow on 7 February 2022, leading Europe's diplomatic efforts to defuse Ukraine crisis.	Insufficient information	The context for the meeting is provided, but we lack information about the nature of the exchange, despite its connection to the broader event of the Ukraine War.
The French President Emmanuel Macron met Russian President Vladimir Putin in Moscow on 7 February 2022, leading Europe's diplomatic efforts to defuse Ukraine crisis. To Moscow, Macron has presented himself as a "quality interlocutor", as Putin described him.	Exchange between human actors as part of a historical event	Actor-Network Theory seeks to understand the nature of the exchange between the actors, which provides partial meaning to an event.

The French President Emmanuel Macron met Russian President Vladimir Putin in Moscow on February 7, 2022, leading Europe's diplomatic efforts to defuse Ukraine crisis. To Moscow, Macron has presented himself as a "quality interlocutor", as Putin described him. The two men who set at the opposite sides of a long white table spent more than five hours locked in head-to-head talks.	Exchange between all actors (humans and objects) as part of a historical event	Actor-Network Theory seeks to understand the nature of the exchange among all actors, which collectively gives meaning to an event. The long table possesses agency equal to that of the two actors – its presence on the scene starkly contrasts with Putin's characterization of Macron as a "quality interlocutor".
	Historical image, which in itself, plays a part in the event	The image possesses its own agency, emphasizing the exaggerated distance between the actors created by the long white table – a detail scarcely perceived in the textual description.

In short, "An actor-network is traced whenever, in the course of a study, the decision is made to replace actors of whatever size by local and connected sites instead of ranking them into micro and macro" (Latour 2005, 179). Whether one agrees with Latour or not, the intriguing aspect here is the potential connection to questions of networks in historical studies. The challenge historians face with network theory is, to put it bluntly, the attempt to classify complexity itself – its degree, recurrence, and intensity – tasks that historians typically leave to mathematicians (Ahnert et al. 2020, 21). Table 3, which attempts to view ANT from a historian's perspective, might be seen by network scholars as a misunderstanding of the theory's essence. However, for historians, moving beyond merely dealing with 'facts' and 'events' to closely examining instances of exchange and mediation, while attempting to understand the agency of all actors, holds the potential for more accurately describing and analysing situations and structures previously unexplored. For instance, the white table played a crucial role in the nature of the exchange between Putin and Macron, a role that cannot be ignored by historians. Moreover, the image of the summit, deliberately released by the Russians, had a significant impact on Western leaders' decision-making process regarding the Ukraine war.

Historians were now confronted with a new set of tools. Their traditional reliance on 'facts' and 'events' had been well-suited to the 'narratives of explanation' proposed by Danto, White, and others (Passmore 1962, 105-23). However, establishing relationships among objects, properties, and beings to create networks for describing historical complexity, and further explaining the 'why' and the 'how',

required a new set of guidelines. Historians began to recognize the need for basic units of measure – but what were these units exactly, and where could they be found? For years, historians had built their analyses on facts and events as documented in archival records (Little 2010, 6). Yet, the complexity paradigm, Network theory, and, most significantly, the Digital Turn, now suggested a different approach to the relationship between historical research and archives.

3 Information Units and Datasets

3.1 Breaking the Event into Information Units

> Numbers have the ability to capture certain attributes that cannot be gleaned simply by reading text or viewing images. Statistics can make an argument that cannot be expressed by words alone. Despite this, the quantitative is perceived as at odds with the normal practices and tropes of cultural commentary. (Ahnert et al. 2020, 74)

As some historians ventured into the realm of numbers using digital tools, their efforts were criticized by colleagues who viewed these approaches as incoherent with the complexities of historical narrative: "The tendency is to see these differences in terms of binary oppositions: the quantitative as opposed to the qualitative, numbers as opposed to words, graphs as opposed to text" (74).

Historians influenced by the Digital Turn were not necessarily seeking a purely quantitative approach. They temporarily set aside the reconstruction of narratives based on 'facts' – particularly 'historical facts' – and began to think in terms of data. Historical data comprise all those basic units of information, categorized into value-oriented typologies or entities: persons, places, objects, institutions, etc. When placed in a simple database, these units serve primarily as access points to records, facilitating researchers' searches for specific information. However, as demonstrated in Table 4 and Figure 1, these units lose their original context when grouped into datasets **[tab. 4; fig. 1]**.

Table 4 Statements of facts and events – classical historical narrative

Statement	Nature of Statement
Józef met Joachim.	Fact
Józef Lipski met with Joachim von Ribbentrop.	Fact
Józef Lipski met with Joachim von Ribbentrop on 4 October 1938.	Historical fact?
Polish ambassador Józef Lipski met with German Foreign Minister Joachim von Ribbentrop on 4 October 1938.	Historical fact
Polish ambassador Józef Lipski met with German Foreign Minister Joachim von Ribbentrop on 4 October 1938, in Berchtesgaden to discuss the German annexation of Danzig.	A cluster of facts – historical event

If Józef and Joachim are identified as Józef Lipski and Joachim von Ribbentrop, their meeting undoubtedly holds historical significance, provided we understand who these individuals are, as well as when and why they met. For instance: "Polish Ambassador Józef Lipski met with German Foreign Minister Joachim von Ribbentrop on 4 October 1938, to discuss the German annexation of Danzig". While this statement presents a brief narrative, it does not fully convey the complexities of the locations or the individuals involved. To deepen our understanding of this event, we might employ a data model description akin to the Resource Description Framework (RDF) vocabulary. However, doing so risks losing the broader context **[fig. 1]**.

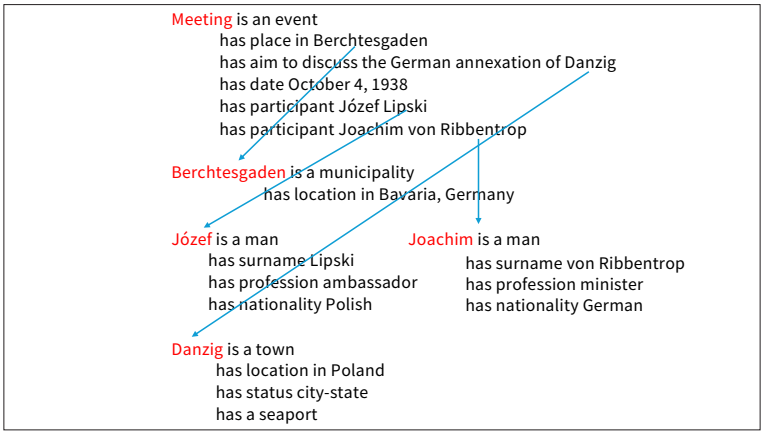


Figure 1 Data extracted from a statement and arranged according to the RDF model

The advantage of this approach lies in its ability to confront these information units with other similar data variables, thereby constructing robust knowledge for studies aimed at uncovering hidden structures that may reveal ongoing processes or patterns of human behaviour. For example, within this context, one might ask the following questions to detect a pattern:

1. Before the outbreak of any war, how many times did representatives from the two opposing sides meet to reach an understanding or negotiate a peace treaty?
2. Who were the representatives, and what roles did they play?
3. Is there an increasing importance in the type of representation as the outbreak of war approaches?
4. Can we identify recurring models in the meeting rituals?
5. Are there differences between historical and contemporary processes of this nature? (Ahnert et al. 2020, 25)

These questions aim to model a type of event to foresee its evolution. Until today historians have not had sufficient data to investigate such inquiries.

Paradoxically, the Archival Turn has, in a sense, precipitated the Digital Turn in historical research. As observed, just as many historians began to engage with the implications of archival evidence in the context of social and cultural history, gender, race, ethnicity, discourse, colonial studies, and other emerging analytical approaches, archivists were starting to pose a very different set of questions. Driven by technological advancements, they quickly produced a substantial body of new literature on archival theory and practice in the electronic age, reevaluating and redefining their profession.

This raises the question: What implications does this have for the future of historiographical research?

First, the essential identifiers – such as names, places, agencies, and dates of creation – simplify the complexities inherent in many documents by condensing them into a rigid set of subject constructions, thereby reducing the range of descriptive tags, such as those found in Dublin Core metadata. Although reductionist, these identifiers seem to embody the qualities of the record: enduring, applicable to historical records, and both uniform and acceptable across various types of institutions. Contemporary historians are increasingly interested in this new form of documental description, contemplating whether they might extend historiographic boundaries further by attempting to model contents and create essential descriptive categories. Such categories could facilitate cross-searching within and across records, even those originating from different institutions and creators.

Second, the advent of relational databases – such as the entity-relationship model, linked data, and graph databases – that organize

data into tables which can be linked based on common data points, has brought to the forefront the question of whether the contents of records can be faithfully synthesized using structural data extracted from them. From the historian's perspective, this suggests that to accurately recreate historical contexts and link data units into coherent facts, a further step may involve classifying archival records in types and then treating each type in a serial manner. This would require the creation of a complex data architecture capable of expressing the intricate relationships between data units, while also considering a future where each data unit is created with automated, tagged provenance.

Lastly, the traditional historiographic approach that interprets the world in terms of ascertained facts and events has become a limiting framework for a society eager to explore new interpretations and to view historical processes differently. Archivists, when discussing 'context', now openly refer to "a realm of contextualities" (a term borrowed from quantum mechanics) (Lane, Hill 2010, 12; Nesmith 2005, 260). This suggests that while the old structure based on the authoritative power of provenance may still apply to individual records, the intersection of datasets can create situations where the narrative exceeds the sum of its parts and takes unforeseen directions. Historians should consider this point in the near future, as the future of archival science, shaped by the Digital Turn, will increasingly influence historical research. This is especially pertinent at a time when recent technological advancements, such as Intelligent Data Extraction (Natural Language Processing, Large Language Models, and Machine Learning), may render the debate over using facts and events as measurable units in historical research somewhat outdated (Hodel, Prada Ziegler, Schneider 2023; Mahadevkar et al. 2024). Nevertheless, the critical question remains: are structured datasets the future of historical research?

3.2 Modelizing Content

As previously noted, recent technological developments may have rendered the debate over 'facts' or 'events' as the fundamental units of historical research somewhat outdated. Historians today think in terms of basic units of information, grouped into value-oriented typologies. With the development of ontologies, these units are now referred to as 'entities', where each object is clearly distinguishable. Each entity is associated with a set of attributes or 'properties'. Each property has a name and one or more values, and it can have values of more than one type. Additionally, two entities can have different types of values for the same property. For example, all human beings possess thousands of different properties, two of which are 'height'

and 'hair colour'. For any particular entity, each property has a 'value': hair colour might be black, fair, grey, etc. If a historian deemed it important to describe the properties of Lipski and von Ribbentrop, they would assign the properties of 'black hair' and 'medium height' to the entity named 'Józef Lipski'. and 'fair hair' and 'medium height' to the entity named 'Joachim von Ribbentrop'.

When placed in a simple database, entities are primarily considered as access points to records, aiding researchers in their search for specific information. However, although these information units may lose their original context, when grouped into datasets, they can be analysed alongside similar data variables or serve as the basis for statistical studies aimed at uncovering hidden structures that might reveal an emerging process. These entities can only be considered 'facts' when they are related back to the historical context that recorded them.

To return to the question posed earlier: can the contents of records be faithfully synthesized using structured data extracted from them? The serial nature of archival records of the same type, created by the same entity, often results in texts that are morphologically similar yet different in content. These module-like texts facilitate data extraction, as the records typically contain recurrent entities that are easily typified, with properties that are readily classified. Consequently, the proposed solution would involve treating archival records in a serial manner, by creating an intricate data architecture capable of expressing the complexity of the relationships between data units, along with the source of each.

Attempting to structure the contents of archival records is not a simple task. Computers operate on binary logic, which does not easily accommodate complexity, while texts are often complex, incorporating descriptive, chronological, comparative, or cause-and-effect elements. Furthermore, historical texts contain layers of datasets that store facts and events interlinked in various ways. It may seem as though the historian is using a Lego-like system to build a narrative, linking facts and events as he/she sees fit. However, this is not the case. A historian cannot link two facts together arbitrarily or simply because the connection appears logical (as illustrated by the "The Plan" in Umberto Eco's *Foucault's Pendulum*). The historical process imposes a set of rules that establish a working framework.

Let us return to the example previously discussed in Tables 4 and 5 regarding Józef Lipski and Joachim von Ribbentrop. When we examine the relationships between the common entities from the perspective of 'facts' and 'events', we encounter two different types of relationships:

1. A is related to B (the type of relationship could be: son, husband, father, is of age, has someone as a friend, employer, or in this case: is an interlocutor). This formula can be expressed

in a graph statement using Resource Description Framework (RDF), composed of a triple statement: subject-predicate-object. For example, Giovanni is a husband; Africa is a continent; Maria has an email address xyz; Józef has the title of ambassador. Typically, this ontology is straightforward because it uses concepts that are understandable or can be verified and measured. The framework also anticipates the use of fuzzy logic, where the truth value of variables can be any number between 0 and 1, functioning like Boolean operators (and/or/not) or using if-then rules. However, these two options are not applicable in any historical context that requires truth statements. In our case, we are dealing with entities related to others, an expression that constitutes a verifiable fact: e.g., Africa is a continent.

2. A acts on B (or vice versa) – this is not merely a description of the subject using the predicate and the object. Here, we are dealing with an event that serves as a cause leading to an effect: A orders/tells/meets B. What is missing here is the information: to do what? As previously mentioned, an event is a cluster of facts and their relationships, arranged in a chronological and/or logical succession of cause and effect. In our case, discussing the annexation of Danzig may be considered an explanatory narrative for the meeting. However, RDF falls short of providing the tools needed to express the complexity of the event and its relationship to a succession of events – whether successive or collateral: why was an insignificant location (so it seems) such as Berchtesgaden chosen for the meeting? In this case we have to link the whole meeting to a new statement (in itself highly complex): "Berchtesgaden was Hitler's favourite vacation location which served as an outpost to the German Chancellary".

When historians decide to extract data from archival records, following Klaus Krippendorff's approach in *Content Analysis: An Introduction to Its Methodology* (2004, 413), they must make several key decisions: what data to analyse, how to define the data, the relevant context, the boundaries of the analysis, and what is to be measured. Regarding the scope or boundaries of the analysis, the question arises whether historians are primarily interested in inserting access points into a database to facilitate research by leveraging the recurrent entities we mentioned – such as persons (name, surname), institutions, objects, places, and their properties. Alternatively, do they aim to describe the complexity of the texts by detailing the relationships between entities and creating a formal content template for each item? In the first scenario, the focus would be less on reconstructing the item itself through the relationships between the

entities expressed in it; in the second scenario, they would proceed to architect the data. In this context, the data architecture should consider the entities necessary to determine the item's type. For example, if dealing with items like letters or dispatches, one would extract the names of the sender and receiver, the date of composition, and any mentioned persons, places, institutions, or objects. In the case of wills, one would extract the date and place of composition, the testator, witnesses, notary, and the estates or sums of money to be distributed to heirs.

However, what we actually extract are the metadata that describe the type of document. This form of data extraction is insufficient for fully representing the contents of the records. Consider, for example, a will. When a testator disposes of his estate, he/she may do so either by listing his heirs and detailing the properties bequeathed to each, or by listing each property and specifying the heir to whom it is bequeathed, along with any conditions attached.

In the first scenario, the model would have the following schema:

- A bequeaths properties 1, 2, and 3 of his estate to B.
- A bequeaths properties 4 and 5 to C, and so on.

In the second scenario, the schema would be:

- Property 1 is bequeathed by A to B.
- Property 2 is bequeathed by A to C with conditions (e.g., a prohibition on selling, a percentage to be given to another party after a certain number of years, etc.).
- Property 3 is bequeathed by A to D.
- Property 4 is bequeathed by A to B.

Each item listed in these two schemas represents an event in itself: a transfer of property. How, then, does the historian convey this cause-effect link (a word that better describes the nature of the bond rather than relationship), that is, how does he/she qualify the link between A, B, and the bequeathed property? One approach is to create a hierarchical structure in which the will serves as the source, linked to multiple items (corresponding to the number of properties in the estate), each representing a property with its testator, its heir, and, where applicable, the conditions attached to the property. This straightforward schema is effective for a will, but how does it apply to a more complex event, i.e., a meta-event, that encompasses a large number of other complex events, each of which is composed of minor ones?

Consider, for example, an event referred to in history as World War II. This is, in a sense, a constructed event, first named by American President Franklin Delano Roosevelt in 1941. Can a historian adequately express the intricate clusters of events that comprise this meta-event, each made up of smaller events, in a manner that places all

under the umbrella of World War II? In theory, it is possible to use a genealogical tree-like structure; however, as we know, some events led to multiple simultaneous effects, generating a series of additional events. This complexity complicates the task of establishing clear cause-and-effect links. In this case, the only model that approximates the description of such a meta-event is CIDOC-CRM, the Conceptual Reference Model, which offers a formal structure for describing concepts and relationships through 80 classes and 130 relationships.[13]

3.3 Object-Centric, Event-Centric and Complex Network Descriptions

In their 2006 paper, "Documenting Events as Metadata", Doerr and Kritsotaki present a conceptual framework embedded in CIDOC-CRM that addresses the challenge of describing events as relationships between entities. The principles of CIDOC-CRM, which are based on the concepts of people, ideas, and objects – in short, persistent items that can intersect in space and time – are, to quote the authors:

> Modelling of events can be used for the representation of metadata and content relationships […], such as participation in an event, part-whole relation, reference information, and classification, which are the most fundamental relationships that connect things, concepts, people, time, and place. (Doerr, Kritsotaki 2006, 56)

The authors aim to provide an accurate representation of the life history of a cultural object and propose a model of event-centric documentation, rather than object-centric, to structure historical data. However, even this conceptual model has its limitations. The first issue concerns the description of cause-and-effect. The authors assume that every event is a form of interaction between A and B. While this assumption may be accepted, it remains unclear how a historian can effectively describe the effects that generate subsequent events – a task traditionally handled by the 'narrative of explanation'.

A significant advancement has been achieved with the recent publication of the *Records in Context Conceptual Model* (RiC-CM). The ontology component, RiC-O, employs Linked Open Data (LOD) techniques. Within this ontology, an 'event' (ID RiC-E14) is classified under a higher-level entity termed 'Thing' (ID RiC-E01). A 'Thing' is defined as "Any idea, material thing, or event within the realm of human experience" (ICA 2023, 19), while an 'event' is described as

13 https://www.cidoc-crm.org/.

> something that happens in time and space [...] An *event* may be discrete, happening at a specific moment in time, or may occur over an extended period of time. *Events* may have *events* as parts, and *events* may precede or follow one another. Multiple *agents* may participate in the same *event*, and in different roles. (32)

RiC-CM thus acknowledges that an 'event' can be more complex than a mere occurrence, potentially encompassing a network of interrelated events that are components of the overarching event.

Even more intriguing in the RiC-CM is the concept of 'activity' (ID RiC-E15), which is defined as "a kind of *event*". It is an *agent* designed and performed *event* that has an intended purpose or purposes. The RiC-CM addresses a dual interpretation of 'activity': "purpose and process are complementary understandings of *activity*. Together, the two perspectives address why the *activity* is performed, the expected ends or outcomes, and how the *activity* fulfils the purpose" (ICA 2023, 32-3). The RiC-CM proposes the following attributes for 'event': General Description, Identifier, Name, Event Type, and History, while for 'activity', it adds the attribute of 'Activity Type' (70). The 'event' or 'activity' may also be associated with the following entities: 'date', 'place', 'thing', or 'subevent'.

Consider the following historical statement: "World War II is generally considered to have begun on 1 September 1939, when Nazi Germany, under Adolf Hitler, invaded Poland". In Figure 2, we will attempt to establish relationships using the RiC conceptual model between the entities to fully convey the meaning of this statement **[fig. 2]**.

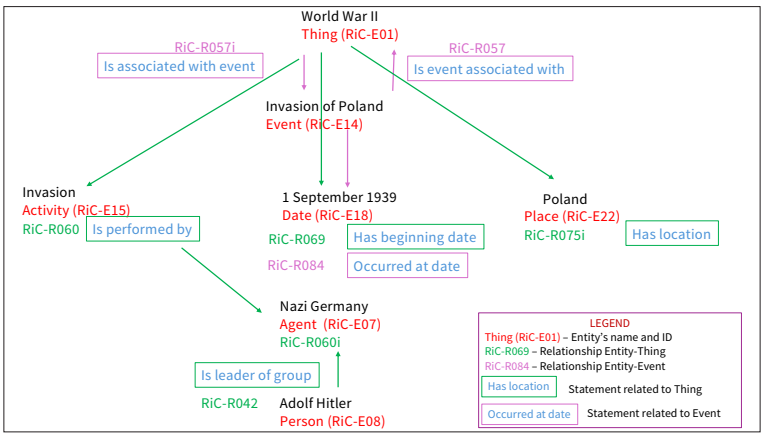


Figure 2 Simulation of historical statement described according to RiC Conceptual Model

While the RiC-CM is undoubtedly a powerful tool that aims to adapt archival ontology to new conceptual models such as Linked Open Data, network diagrams, and the Entity-Relationship Model, it is primarily designed to address the activity of describing records rather than historical events. Its flexibility is impressive, but it still falls short in enabling the description of complex meta-events like World War II, with its intricate web of cause-and-effect events and the agency of both humans and non-humans.

One issue encountered when designing the statement's data architecture using RiC-CM involves the 'event' entity (second level in the entity hierarchy) and its relationship to 'activity' (third level in the entity hierarchy) (ICA 2023, 17). As discussed above, RiC-CM considers 'activity' to be a kind of event, yet it does not establish a formal relationship (using a specific ID) between the two, while it does create a formal relationship between 'thing' (first level in the entity hierarchy) and 'activity'. Moreover, even if we attempt to replace 'event' with 'activity', we find that 'activity' lacks a formal relationship with 'date' or 'place', whereas 'event' does have a formal relationship with 'date'. This situation necessitates the use of the first-level 'thing', which is formally related to 'activity', 'date', and 'place', bypassing the 'event' entity altogether.

It is evident, then, that a more complex statement such as: the causes of World War II, which included unresolved tensions in the aftermath of World War I and the rise of fascism in Europe and militarism in Japan, and was preceded by events including the Japanese invasion of Manchuria, the Spanish Civil War, the outbreak of the Second Sino-Japanese War, and the German annexation of Austria and the Sudetenland[14] would be nearly impossible to describe using RiC-CM.

Another issue, illustrated by the example of World War II, involves multiple assumptions based on differing interpretations of facts. World War II is an event with a debated start date. Is it the September 1939 invasion of Poland by Hitler, as Shirer (1960, 535) suggests? Or is it the March 1938 *Anschluss* of Austria into the German Reich, or the 1938 Sudetenland crisis (Davies 2006, 56-7)? The end date is equally problematic: is it 8 May 1945, with the German surrender, or 2 September of the same year, with the Japanese surrender (11-20)?

How can these differing interpretations, based on different understandings of the nature of the meta-event itself, be meaningfully represented in a relational database? This issue is not about standardization or disambiguation; rather, it concerns how to express data extracted from various archival sources, each presenting its own

14 Both statements are taken from 'World War II' at `https://en.wikipedia.org/wiki/World_War_II`.

logic. A solution for expressing disputed attributions and dating (*capta*) is introduced by Vitali and Pasqual (*infra*), who, after evaluating existing models of RDF reification, n-ary relations and CIDOC CRM ontology, Named Graphs, and Wikidata statements and rankings, propose representing *capta* using a conjectural graph based on RDF triplets, incorporating a newly created predicate specifically designed for this model. However, applying the Conjectures model to the description of a meta-event may hinder our understanding of its complexities, as the expressed uncertainty could potentially distort the cause-and-effect links between events.

It becomes evident that the mere metadata of entities falls short of meeting the historical research requirement for 'narratives of explanation', which prioritizes the selection of facts and 'eventworthy' occurrences to be included in a meaningful narrative that elucidates how events unfolded. The lack of a solution for describing complex events through the construction of datasets hinders the integration of the historical discipline into the Digital Humanities domain. Historians may need to reconsider their methodological tools or broaden their definitions of 'facts' and 'events' to encompass perspectives such as that of Charles Tupper.

Tupper, when discussing data organization practices in the business world, defines an 'event' as "an occurrence that sets in motion an activity that changes the state of an entity within the model".

The key idea here is that an event relates to entities, not facts, and therefore the focus of data architecture is not the 'construction of meaning', but rather the impact of the occurrence on the entity or entities: "events represent the triggering of processes and procedures" because they add, delete, or update the entities' properties (Tupper 2011, 180). If we accept this perspective, then we might also accept the creation of links between entities that participate in events (as does RiC-CM), rather than connecting the events themselves. This approach can be applied using a network analysis method proposed by Kenna and MacCarron – a complex networks analysis of narratives (specifically, epic narratives in the authors' case).

In a sense, following Koselleck's concept of "a differential classification of historical sequences", the idea underlying Kenna and MacCarron's approach is the "notion of universality of critical phenomena". This notion suggests that, despite differences at a microscopic level, the macroscopic properties of systems undergoing phase transitions (so-called 'critical systems') depend only on a few parameters, such as symmetries and dimensions. Critical systems can, therefore, be categorized according to their universality class, which is in turn defined by a small set of numbers known as "critical exponents" (Kenna, MacCarron 2022, 23). Ultimately, the two scholars argue that the nature of relationships between entities can be reduced to a small set of critical exponents.

If future historians are willing to accept that, instead of focusing on the 'event', the unit of analysis should be the nature of the relationship between entities (of any kind, not necessarily human) that influences an occurrence, and if they can agree on the parameters that define 'critical systems', then we may develop, alongside the 'narratives of explanation', an analysis of World War II from a radically different perspective – one that is no less significant and potentially even more illuminating: we would be able to measure the intensity and density of these relationships.

By establishing a set of 'critical systems' with predefined exponents (for example, if A is hostile to B, B may or may not react; if A is aggressive towards B, B is much more likely to respond in kind), historians could focus on investigating only the unforeseen reactions. Furthermore, with the advent of technological innovations or shifts in cultural and ethical approaches, historians would be able to dynamically detect and measure the extent of change occurring within these 'critical systems'.

Today, we stand at a crossroads. The central question is whether historians should continue to adhere to a conceptual framework based on entities, properties, and their relationships, or whether they should adopt a new framework that better suits their needs – one that, instead of relying on facts and events as units of analysis, centres on the nature of actors' relationships as the driving force behind a chain of occurrences that ultimately become 'eventworthy'.

Only time will tell.

Bibliography

Ahnert, R.; Ahnert, S.E.; Coleman, C.N.; Weingart, S.B. (2020). *The Network Turn. Changing Perspectives in the Humanities*. Cambridge: Cambridge University Press.

Annales de démographie historique (2005). "Histoire de la famille et analyse de réseaux", 109.
https://shs.cairn.info/revue-annales-de-demographie-historique-2005-1?lang=fr

Annales de démographie historique (2008). "Les réseaux de parenté, refonder l'analyse", 116.
https://shs.cairn.info/revue-annales-de-demographie-historique-2008-2?lang=fr

Beckert, J. (2009). "The Social Order of Markets". *Theory and Society*, 38(3), 245-69.

Bertrand, M.; Lemercier, C.; Guzzi, S. (2010). "Analyse de réseaux et histoire". *Calenda – O calendário de letras e de ciências sociais e humanas*.
https://doi.org/10.58079/fs6

Bloch, M. (1953). *The Historian's Craft*. New York: Vintage Books. Transl. of *Apologie pour l'histoire ou métier d'historien*. Paris: Armand Colin, 1949.

Blouin, F.X. Jr.; Rosenberg, W.G. (2011). *Processing the past: Contesting Authorities in History and the Archives*. Oxford: Oxford University Press.

Boorman, S.A.; White, H.C. (1976). "Social Structure from Multiple Networks. II. Role Structures". *The American Journal of Sociology*, 81(6), 1384-446.

Bourdieu, P. (1984). *Distinction. A Social Critique of the Judgement of Taste*. Cambridge (MA): Harvard University Press. Transl. of: *La distinction. Critique sociale du jugement*. Paris: Les Éditions de Minuit, 1979.

Bouwsma, W.J. (1981). "From History of Ideas to History of Meaning". *The Journal of Interdisciplinary History*, no. monogr., *The New History: The 1980s and Beyond*, 279-91.
https://www.jstor.org/stable/203030?seq=1

Burguière, A. (2009). *The Annales School. An Intellectual History*. Ithaca; London: Cornell University Press. Transl. of *L'École des Annales. Une histoire intellectuelle*. Paris: Odile Jacob, 2006.

Callahan, S. (2024). "When the Dust Has Settled: What Was the Archival Turn, and Is It Still Turning?". *Art Journal*, 83(1), 74-88.

Carr, E.H. (1961). *What is History?*. New York: A Vintage Book.

Cartwright, D.; Harary, F. (1956). "Structural Balance: A Generalization of Heider's Theory". *Psychological Review* 63(5), 277-93.

Danto, A.C. (1962). "Narrative Sentences". *History and Theory*, 2, 146-79.

Danto, A.C. (1965). *Analytical Philosophy of History*. Cambridge: Cambridge University Press.

Davies, N. (2006). *Europe at War 1939-1945. No Simple Victory*. London: Macmillan.

Dedieu, J.-P.; Moutoukias, Z. (1998). "Approche de la théorie des réseaux sociaux". Castellano, J.L.; Dedieu, J.-P. (éds), *Réseaux, familles et pouvoirs dans le monde ibérique à la fin de l'Ancien Régime*. Paris: CNRS Éditions, 7-30.

Doerr, M.; Kritsotaki, A. (2006). "Documenting Events in Metadata". Ioannides, M.; Arnold, D.; Niccolucci, F.; Mania, K. (eds), *The 7th International Symposium on Virtual Reality, Archaeology and Cultural Heritage VAST (2006).* Budapest: ARCHAEOLINGUA, 56-60.

Duchein, M. (1983). "Theoretical Principles and Practical Problems of Respect des Fonds in Archival Science". *Archivaria*, 16, 64-82.

Durkheim, É. (1947). *On the Division of Labor in Society*. New York: The Free Press. Transl. of *De la division du travail social. Étude sur l'organisation des sociétés supérieures*. Paris: Félix Alcan, 1893.

Durkheim, É. (1982). *The Rules of Sociological Method*. New York: The Free Press. Transl. of *Les Règles de la méthode sociologique*. Paris: Félix Alcan, 1895.

Eley, G. (1996). "Is All the World a Text? From Social History to the History of Society Two Decades Later". McDonald, T.J. (ed.), *The Historic Turn in the Human Sciences*. Ann Arbor: University of Michigan Press, 183-243.

Foucault, M. (2004). *The Archeology of Knowledge*. London: Routledge. Transl. of: *L'Archéologie du savoir*. Paris: Gallimard, 1969.

Friedrich, M. (2018). *The Birth of the Archive. A History of Knowledge*. Ann Arbor: University of Michigan Press.

Geertz, C. (1973). *The Interpretation of Cultures. Selected Essays*. New York: Basic Books.

Geertz, C. (1995). *After the Fact. Two Countries, Four Decades, One Anthropologist*. Cambridge (Mass.): Harvard University Press.

Henderson, L.J. (1932). "An Approximate Definition of Fact". *University of California Publications in Philosophy* 14(7), 179-200.

Hodel, T.; Prada Ziegler, I.; Schneider, Ch. (2023). "Pre-Modern Data: Applying Language Modeling and Named Entity Recognition on Criminal Records in the City of Bern". *Digital Humanities Conference 2023*. https://doi.org/10.5281/zenodo.8141708

Hölling, H.B. (2015). "Archival Turn: Toward New Ways of Conceptualising Changeable Artworks". Smits, R.; Manovich, L. (eds), *Data Drift: Archiving Media and Data Art in the 21st Century*. Riga: RIXC and Liepaja's University, Art Research Lab, 73-81.

Homans, G.C. (1951). *The Human Group*. London: Routledge & Kegan Paul.

Howell, M.; Prevenier, W. (2001). *From Reliable Sources. An Introduction to Historical Methods*. Ithaca; London: Cornell University Press.

Hyman, H.H. [1942] (1980). *The Psychology of status*. New York: Arno Press.

Imízcoz Beunza, J.M. (2011). "Actores y redes sociales en Historia". Carvajal de la Vega, D.; Añíbarro Rodríguez, J.; Vítores Casado, I. (coords.), *Redes sociales y económicas en el mundo bajomedieval*. Valladolid: Castilla Ediciones, 17-29.

Imízcoz Beunza, J.M.; Arroyo Ruiz, L. (2011). "Redes sociales y correspondencia epistolar. Del análisis cualitativo de las relaciones personales a la reconstrucción de redes egocentradas". *REDES – Revista hispana para el análisis de redes sociales* 21, 98-138.

International Council on Archives Expert Group on Archival Description (ICA) (2023). *Records in Context Conceptual Model, version 1.0*. https://www.ica.org/resource/records-in-contexts-conceptual-model/

Kenna, R.; MacCarron, P. (2017). "A Networks Approach to Mythological Epics". Kenna, R.; MacCarron, M.; MacCarron, P. (eds), *Maths Meets Myths: Quantitative Approaches to Ancient Narratives*. Cham: Springer, 21-43.

Koselleck, R. (2004). *Futures Past. On the semantics of historical time*. New-York; Chichester: Columbia University Press. Transl. of *Vergangene Zukunft. Zur Semantik geschichtlicher Zeiten*. Frankfurt am Main: Suhrkamp Verlag, 1979.

Krippendorff, K. (2019). *Content Analysis: An Introduction to Its Methodology*. London: sage.

Lane, V.; Hill, J. (2010). "Where Do We Come From? What are We? Where are We Going? Situating the Archive and Archivists". Hill, J. (ed.), *The Future of Archives and Recordkeeping: A Reader*. London: Facet Publishing, 7-26.

Langlois, Ch.-V.; Seignobos, Ch. (1904). *Introduction to the Study of History*. New York: Henry Holt. Transl. of *Introduction aux études historiques*. Paris: Hachette, 1898.

Latour, B. (1999). "Factures/Fractures: From the Concept of Network to the Concept of Attachment". *Res: Anthropology and Aesthetics*, 36, 20-31.

Latour, B. (2004). "Why Has Critique Run out of Steam? From Matters of Fact to Matters of Concern". *Critical Enquiry*, 30(2), 225-48.

Latour, B. (2005). *Reassembling the Social – An Introduction to Actor-Network-Theory*. Oxford: Oxford University Press.

Little, D. (2010). *New Contributions to the Philosophy of History*. Dordrecht; Heidelberg; London; New York: Springer.

Loriga, S.; Revel, J. (2022). *Une histoire inquiète: Les historiens et le tournant linguistique*. Paris: EHESS; Gallimard; Seuil.

Lyotard, J.-F. (1984). *The Postmodern Condition: A Report on Knowledge*. Manchester: Manchester University Press. Transl. of *La Condition postmoderne: rapport sur le savoir*. Paris: Les Éditions de Minuit, 1979.

Mahadevkar, S.V.; Patil, S.; Kotecha, K.; Soong, L.W.; Choudhury, T. (2024). "Exploring AI-Driven Approaches for Unstructured Document Analysis and Future Horizons". *Journal of Big Data*, 11, 92.
https://doi.org/10.1186/s40537-024-00948-z

Martinat, M. (2023). "New Paradigms in French Historiography, or the Same Old Ones?" *Literature* 3, 231-41.

Moreno, J.L. (1934). *Who Shall Survive? A New Approach to the Problem of Human Interrelations.* New York: Beacon Press.

Mullaney, S. (1996). "Discursive Forums, Cultural Practices: History and Anthropology in Literary Studies". McDonald, T.J. (ed.), *The Historic Turn in the Human Sciences*. Ann Arbor: University of Michigan Press, 161-89.

Nassiet, M. (1995). "Réseaux de parenté et types d'alliance dans la noblesse à l'époque moderne". *Annales de démographie historique* 1, 105-23.

Nesmith, T. (2005). "Reopening Archives: Bringing New Contextualities into Archival Theory and Practice". *Archivaria*, 60, 259-74.

Noiriel, G. (1996). *Sur la "crise de l'histoire".* Paris: Belin.

Parsons, T. (1937). *The Structure of Social Action*. New York: The Free Press of Glencoe.

Passmore, J. (1962). "Explanation in Everyday Life, in Science, and in History". *History and Theory* 2(2), 105-23.

Poovey, M. (1998). *A History of the Modern Fact. Problems of Knowledge in the Sciences of Wealth and Society.* Chicago: The University of Chicago Press.

Rainie, L.; Wellman, B. (2012). *Networked. The New Social Operating System.* Cambridge (MA): The MIT Press.

Revel, J. (éd.) (1996). *Jeux d'échelles: la micro-analyse à l'expérience*. Paris: Gallimard.

Rolnik, S. (2012). *Archive Mania: 100 Notes, 100 Thoughts*. Berlin: Hatje Cantz

Roth, P.A. (2020). *The Philosophical Structure of Historical Explanation*. Evanston (IL): Northwestern University Press.

Ryle, G. [1971] (2009). *The Thinking of Thoughts. What is 'le Penseur' Doing?* Ryle, G. (ed.), *Collected Essays 1929-1968*. London; New York: Routledge, 494-510.

Scott, J. (2012). *What is Social Network Analysis?.* London: Bloomsbury Academic.

Sherif, M. (1968). "The Concept of Reference Groups in Human Relations". Hyman, H.; Singer, E. (eds), *Readings in Reference Group Theory and Research.* New York: The Free Press, 84-94.

Shirer, W.L. (1960). *The Rise and Fall of the Third Reich. A History of Nazi Germany*. New York: Simon and Schuster.

Simon, Ch. (2002). "Introduction: Following the Archival Turn". *Visual Resources*, 18, 101-7.

Smelser, N.J.; Swedberg, R. (1994). "The Sociological Perspectives on the Economy". Smelser, N.J.; Swedberg, R. (eds), *The Handbook of Economic Sociology*. Princeton: Princeton University Press, 3-26.

Smith, D. (1991). *The Rise of Historical Sociology*. Philadelphia: Temple University Press.

Spiegel, G.M. (2005). "Introduction". Spiegel, G.M. (ed.), *Practicing History. New Directions in Historical Writing After the Linguistic Turn*. New York; London: Routledge, 1-31.

Stern, F. (1956). *The Varieties of History. From Voltaire to the Present*. Cleveland; New York: Meridian Books.

Stoler, A.L. (2009). *Along the Archival Grain. Epistemic Anxieties and Colonial Commonsense*. Princeton; Oxford: Princeton University Press.

Toews, J.E. (1987). "Intellectual History after the Linguistic Turn: The Autonomy of Meaning and the Irreducibility of Experience". *The American Historical Review*, 92(4), 879-907.

Tucker, A. (2004). *Our Knowledge of the Past. A Philosophy of Historiography.* Cambridge: Cambridge University Press.

Tupper, Ch.D. (2011). *Data Architecture. From Zen to Reality*. Burlington (MA): Morgan Kaufmann; Elsevier.

Vernon, J. (1994). "Who's Afraid of the 'Linguistic Turn'? The Politics of Social History and Its Discontents". *Social History*, 19(1), 81-97.

Veyne, P. (1971). *Writing History. Essay on Epistemology*. Middletown (CT): Wesleyan University Press. Transl. of *Comment on écrit l'histoire: essai d'épistémologie*. Paris: Seuil, 1970.

Weber, M. (1978). *Economy and Society: An Outline of Interpretive Sociology.* Berkeley: University of California Press. Transl. of *Wirtschaft und Gesellschaft*. Tübingen: Verlag von J.C.B. Mohr (Paul Siebeck).

White, H.C. (1970). *Chains of Opportunity: System Models of Mobility in Organizations*. Cambridge (MA): Harvard University Press.

White, H.C.; Boorman, S.A.; Breiger, R.L. (1976). "Social Structure from Multiple Networks. I. Blockmodels of Roles and Positions". *American Journal of Sociology*, 81(4), 730-80.

White, H. (1973). *Metahistory. The Historical Imagination in Nineteenth-Century Europe*. Baltimore; London: The Johns Hopkins University Press.

White, H. (1987). *The Content of the Form. Narrative Discourse and Historical Representation*. Baltimore; London: The Johns Hopkins University Press.

Models of Data Extraction and Architecture in Relational Databases of Early Modern Private Political Archives
edited by Dorit Raines

Is There a Reception of Algorithm-Based Research in Traditional Historical Scholarship?
Three Case Studies from Academic "Trading Zones"

Thomas Wallnig
Universität Wien, Österreich

Abstract A decade or more into the 'digital transformation', Digital Humanities papers are notably absent in traditional research bibliographies. This study examines three use cases where non-digital historiography could benefit from algorithm-based research focused on the same subject matter: the medieval treatise *Imitatio Christi*, the rise of the Medici in fifteenth-century Florence, and the French *Encyclopédie*. The conclusion is that without any convergence on content-related matters, Digital Humanities might emerge as a separate historical discipline.

Keywords Digital Humanities convergence. Trading zones. Imitatio Christi. Medici. Encyclopédie.

Summary 1 Introduction. – 2 *Imitatio Christi*. –3 *Robust action and the rise of the Medici*. – 4 *Encyclopédie*: Can Algorithmic Research Contribute to Intellectual History?. – 5 Outlook.

1 Introduction

It seems that the limitlessness of digital resources has taken its toll on the conceptual coherence of humanities scholarship. While scholars with digital expertise try to formulate the traits of the ongoing digital transformation in regular iterations and inscribe it into the

Edizioni Ca'Foscari

Studi di archivistica, bibliografia, paleografia 8
e-ISSN 2610-9093 | ISSN 2610-9875
ISBN [ebook] 978-88-6969-919-1 | ISBN [print] 978-88-6969-920-7

Peer review | Open access
Submitted 2023-10-03 | Accepted 2024-01-18 | Published 2025-05-22

DOI 10.30687/978-88-6969-919-1//002

longer history of epistemological change in the humanities,[1] generations of students – and later PhD candidates – are still being academically supervised in keeping with a mindset in which digital references merit little more than a sheepish footnote.

If – as a central argument of the sceptical faction goes – digital scholarship fails to prove its added value, or at least one that justifies the significant efforts its practice entails, then it may be appropriate to ask in which ways traditional scholarship eventually should engage with those outcomes of digital research that can be considered secure and established knowledge. For this broader set of questions, Max Kemman (2021) has aptly introduced the Galsionian concept of "trading zones" to describe the interaction between history and digital tools and methods.[2] For this paper, I have selected three case studies providing 'digital' results which, as I will show, align with some of the dominant traditional research questions; this selection deliberately leaves aside more tentative and exploratory research where this connection is less obvious.

The choice of these three case studies is arbitrary and reflects my own background as a German-academia historian of premodern Central Europe.[3] It is a field in which – unlike archaeology, for example – there is traditionally no eye-level interaction between data analysis and interpretative scholarship. This is because (in general terms and with all due exceptions) its scholarly efforts consist of close reading and interpretative rendering of written historical source material that a scholar has chosen, ordered and prepared for argumentation

1 As one example among many, cf. the excellent and epistemologically ambitious series "Studies in Digital History and Hermeneutics": https://www.degruyter.com/serial/sdhh-b/html.

2 Kemman has dedicated a book to the academic successes and hindrances of that dialogue: "Yet as part of the detachment of historical research from the tool and technology, supervisors recommended that PhD candidates in history ultimately needed to produce a historical thesis […] and the disciplinary value of working not just with digital data, but also with 'paper archives'" (2021, 127). Cf. Kemman, Kleppe, Scagliola 2014; Cady 1990, 374-86.

3 However, the broader range of possible alternative case studies illustrate that the findings of this paper may by and large be considered representative: in early-modern newspaper and media studies as well, the algorithmic operationalisation of research questions is in its infancy (https://books18c.hypotheses.org/) despite existing historically and algorithmically informed research. Kittelmann, Purschwitz 2019; historical outlines of "Actor network theory" (like Füssel, Neu 2021) do not come with digital dimensions although reflected digital operationalisations of the concept exist, as in the *Campus Medius* project which deals with digital cartography in cultural and media studies: https://campusmedius.net/overview?lang=de#p:7; and the scholarly findings generated within the scope of COST Action "Reassembling the Republic of Letters" (Hotson, Wallnig 2019, especially section IV) have only very gradually found their way into traditional scholarship on the matter.

and selective presentation.[4] The fact that the methodological challenges brought about by the social sciences since the 1970s have perhaps never been fully digested makes scepticism of the digital seem like the echo of a repressed conflict over qualitative and quantitative methods.

What is more, there seems to be divergence even with regard to basic terminology. 'Methods' in the humanities denote epistemic frameworks and terminological toolkits rather than algorithms, which sometimes causes misunderstandings because of differing or even largely incompatible assumptions about the nature and value of respective research assertions: a non-structured description of an archival fonds, even if it contains all the necessary information, is of little use to a researcher using computer-based methods, while the comparison of network metrics in a given prosopography may remain strange even to scholars closely familiar with the data.

Rather than complaining or theorizing about or even trying to change this state of affairs, however, this paper attempts to shed light on what is actually happening in the bibliographies and footnotes of the respective books and papers. In all three case studies, an outline of the research problem will be provided before the 'digital' research in the field is described and its consideration in ongoing research is discussed.

2 *Imitatio Christi*

The spiritual treatise on the Imitation of Christ was not only "among the most successful texts of the fifteenth and sixteenth centuries" (Harrap 2017, 3), but also one of the texts most contested among ecclesiastical scholars of the early-modern period. While Jesuits, Augustinian Canons and others argued (rightly, as we know today) in favour of its creation by Canon Regular Thomas a Kempis (d. 1471), Benedictines of the seventeenth and eighteenth centuries in particular tried to claim authorship for their (possibly fictional) confrere Giovanni Gersen, abbot of Vercelli in the early thirteenth century – a view that still informs the 1982 edition of the text.[5]

While the early-modern controversy was mainly historical-critical (Benz 2003, 539-49), the debate already began to acquire a dimension of quantitative linguistics before 1900. Pierre Puyol, historian and bibliographer, tried to approach the question (and sustain the

4 Projects addressing premodern topics (like manuscript studies: `https://cima.or.at/links/`) from the perspective of heritage science remain an exception. With a stunningly externalist view (and often silence) on matters digital: Spoerhase, Martus 2022, 92.

5 *Imitatio Christi et contemptus omnium vanitatum mundi*. `https://geschichtsquellen.de/werk/3014`.

Gersen hypothesis) by grouping parts of the large surviving manuscript evidence into a stemma. While this grouping remains accepted, Puyol's conclusions were challenged from around 1940 by statistician George Udny Yule, who had turned to the matter of statistical analysis of texts towards the end of his career and with no particular historical ambition (Yates, Yule 1952, 312). By comparing first sentence length and then other linguistic features of the *Imitatio Christi* with the same features in the known works of the two potential authors, Yule made a robust argument for the authorship of Thomas a Kempis; while this can be considered confirmed in the meantime, today's historical and philological scholarship asks other questions than those related to original authorship, turning to the practices of textual transmission in the context of the *Devotio Moderna*.

As a result, recent scholarship will generally briefly mention the statistical approaches (Habsburg 2017, 15) and place them in the context of previous research (Harrap 2017, 15). However, it will not discuss any of the details or assume any validity of Yule's findings beyond the resolved question of authorship attribution. Conversely, one of the classic overview works of literary statistics reports Yule's criticism and further development of Puyol's work among its case studies on the statistical analysis of writing style (Oakes 1998, 203-5):[6] with the help of a computer as early as 1939, Yule began by comparing the sentence length of text samples before introducing additional criteria for stylistic analysis in a second step – among them the 'K characteristic', a measure of vocabulary richness. This method makes it possible to determine whether a text contains words that are rare and thus distinctive, and Yule – basing his statistics exclusively on nouns – combined this characteristic with the total size of the vocabulary, the frequency distribution of the different words used, the mean frequency of the words in a given sample and the number of nouns occurring only in that sample. Simply put, he developed basic stylometry based on the *Imitatio Christi*.

It should be added that none of the underlying data – a machine-actionable version of the text – have survived, and that to this day there is no digital version of the *Imitatio*. Together with the decreasing (if not entirely ceasing) interest concerning the question of authorship and the general shift in philological perspectives in the sense of the "New Philology" (Roelli 2020, chs. 1-2), this may explain why no further use has been made of the first-level statistical research of earlier decades.

But why should anyone interested in the *Imitatio* today make use of it? Recent studies deal with Catholic and Protestant translations of the work (Habsburg 2017) or more generally with its early-modern reception and adaptation (Aurnhammer, Steiger 2019), while others

6 For Yule's merely statistical refutation of Puyol, see Yule 1946, 44-52.

address the use of scriptural sources in the text (Becker 2002). For the former, it would not be inconceivable given the huge quantity of material at stake to reproduce Yule's statistically tested analysis on appropriate samples, which would allow for stylistic clustering alongside theological observations. For the latter, Yule's analyses cannot be repurposed – but the carefully distinguished types of use of scriptural material are a typical test case for text reuse software, where the adjustment of fuzziness parameters can become a hermeneutical counterpart to researchers' analytical efforts.[7]

In conclusion, we can affirm that research on the *Imitatio* accepts the episode of statistical analysis as a (minor) part of its bibliographic genealogy with regard to the authorship question but does not consider making use of its results or algorithmic methods for a formalized approach to its current questions.

3 *Robust Action and the Rise of the Medici*

Like the previous one, the second use case also points to a period in which statistical analysis was not yet framed as data science[8] and the internet was in its early days: in 1993, American political scientists John F. Padgett and Christopher K. Ansell published their article "Robust Action and the Rise of the Medici 1400-1434" in the *American Journal of Sociology* (Padgett, Ansell 1993, 1259-319). The text, which results from the authors' interest in organisational theory and power structures more than from genuine historical curiosity, has become one of the classical works of historical network research (e.g., Petz 2022).

The question addressed by Padgett and Ansell is how it was possible for the Republic of Florence, one of the flourishing states of the Italian Renaissance, to be brought under the autocratic regime of the Medici family, whose main exponent in the corresponding period – Cosimo il Vecchio – did not hold any public office. The sources on which the study is based reflect the administrative ambitions of a late medieval (city) state: surviving marriage and business registers along with accounts of the political organs of the republic and maps of the individual families' habitations. These sources had already been

7 Research of the past decade into the writing practices underlying the *Encyclopédie* has led to results of this type. While the team was initially interested in detecting undeclared text reuse – especially Rousseau – they went on to use word co-occurrence metrics to tentatively describe the similarity of thought; see for example the works of Clovis Gladstone (`https://rll.uchicago.edu/clovis-gladstone-0`), in particular Gladstone 2020; Gladstone, Cooney 2020 and also below, § 4.

8 For the problem of data science merely being a new label for statistics, see O'Neil, Schutt 2013, 1-16.

edited at the time of the two researchers' efforts, so that they could focus their energy on making them mathematically processable.

Based on this material, Ansell and Padgett constructed a network of weighted ties resulting in a block model showing – in very simple terms – that the Medici maintained strong connections with two rivalling factions that would otherwise not have been in contact. This gave them (and in particular, Cosimo) factual power in the social field that meant the undermining of the republican institutions was only a matter of time.

"Robust Action" thus not only represents a key paper within the history of historical network research, but also one in the history of Renaissance Florence and that of premodern state building in general. But does recent scholarship in the latter two fields take account of the work?

Given the iconic character of Renaissance Florence for any narrative of the 'West', there is an insurmountable quantity of books on the matter. However, we can identify works published after 1993 which gravitate around the topic in the broader and narrower sense.[9] John Najemy's overview *History of Florence* (2008) treats the crucial phase in the early fifteenth century without any reference to Padgett and Ansell, and the same is true for specialized works focusing on the Medici affirmation period from a perspective of state building[10] and social order,[11] which could arguably profit from a discussion of the results of this network research. Conversely, much of the historical literature including "Robust Action" is cited and discussed in a paper about "Patronage, Citizenship, and the Stalled Emergence of the Modern State in Renaissance Florence" by sociologist Paul D. Rutgers (2005). In other words, the bibliographical divide here seems to be a unilateral one.

But this bibliographical divide does not end at Florentine or Tuscan history. There is likewise no consideration of historical-sociological accounts or organization theory in recent historical scholarship on the formation of the modern state, including new administrative

9 Most of the following bibliography is taken from Brege 2021, 333 fn. 21.

10 E.g., Cohn 1999, who in his appendix I works with statistics as well (Regression Models: Wealth, Migration, and Taxes), and who in an introductory chapter named "Networks of Culture and the Mountains" (13-54) could have made explicit that the term, in a methodologically more rigorous sense, had been applied to the same matter only a few years earlier.

11 An engagement with "Robust Action" would have been obvious, to name but one example, in the context of the following passage in Taddei 2004, 39-62, section 26: "Cette forme de gouvernement non officiel mis en place par les Médicis nécessitait donc sans cesse la construction subtile d'un réseau de rapports personnels, familiaux et diplomatiques qui devait leur assurer l'appui politique des familles les plus influentes de l'oligarchie".

history.[12] This is surprising, since sources similar to those extant for fifteenth-century Florence would also be available for other early-modern polities;[13] also, because network-based studies on early-modern court culture[14] suggest that the existing research questions could be addressed fruitfully by taking the existing findings in the field seriously (Enderlin-Mahr 2022, 35-48).

An additional example seems to confirm the trend of the diverging bibliographies: the work of Katalin Prajda (2018), a historian of Renaissance Florence who collaborates closely with John Padgett, is referenced with regard to network research in a book on management studies (Tavanti 2019, 230), whereas she appears in traditional scholarship as a historian of the relations between Florence and the Kingdom of Hungary (Baker, Maxson 2020, 12 fn. 5). Similar observations can be made with regard to other medieval historians as well (e.g. Gramsch-Stehfest 2013).

Does this mean we will not only need 'institutional' "trading zones" but also long-term 'bibliographical' ones?

4 *Encyclopédie*: Can Algorithmic Research Contribute to Intellectual History?

The *Encyclopédie* was one of the most famous works of the French Enlightenment. Edited by Denis Diderot, Jean-Baptiste le Rond d'Alembert, Louis de Jaucourt and others, this highly controversial dictionary reframing existing knowledge from the perspective of enlightened thought was published in 17 volumes (and 11 volumes of plates) between 1751 and 1772. Large yet circumscribed works like this lend themselves to digital processing, and today, two major projects offer searchable full-text databases of the material: the French *Édition Numérique Collaborative et Critique de l'Encyclopédie* (ENCCRE), with scans of the original copy from the Bibliothèque Mazarine, and the Chicago-based ARTFL project,[15] which in turn comes with highly advanced search options.[16]

12 E.g., Reinhard 2007; Blockmans, Genet 1996, in particular part D: "Power Elites and State Building". See also the research design of the journal of administrative history, *Administory*: https://www.bar.admin.ch/bar/de/home/service-publikationen/verwaltungsgeschichte/administory--online-zeitschrift.html.

13 Taking the Habsburg Monarchy as an example, see the prosopographical material recorded in Hochedlinger, Mat'a, Winkelbauer 2019; as well as in the database https://viecpro-project.oeaw.ac.at/.

14 See Ahnert et al. 2020, 93-5, on a network-based analysis of the Tudor State Papers.

15 http://enccre.academie-sciences.fr/encyclopedie/.

16 https://encyclopedie.uchicago.edu/.

From the ARTFL context (and through the availability of machine-readable versions of 74,000 articles), research has emerged that attempts to bring together the analytical potential of the resource with the traditional *Encyclopédie*-related questions in intellectual history. For example, the article "Discourses and Disciplines" (Roe, Gladstone, Morrissey 2015) discusses the application of supervised and unsupervised machine-learning techniques to the corpus, and in doing so addresses at least three fields relevant to intellectual historians.

First, a tentative but convincing connection is made between corpus analyses and the early phases of French discourse analysis, specifically Pêcheux's "automatic discourse analysis" (Roe, Gladstone, Morrissey 2015, 5). These approaches, as the authors admit, were too structuralist in their perspective and too narrow in their scope, but could – following the broadening of the concept of discourse by Foucault – be re-evaluated with a view to techniques like topic modelling:

> It is not unreasonable, for instance, to posit that Foucault's concept of archeology, in fact, justifies the 'bag of words' analytical model used by topic modelling and other machine-learning algorithms; a model that has often come under scrutiny (for good reason) by humanists. By locating words within a set of discursive practices rather than linguistic rules, Foucault's concept of discourse frees us from exclusive interest in language structure, and what that structure conveys, and orients us more toward the association of the various words, concepts, or 'topics' that form a discourse.

However, this line of argumentation is not developed further in current introductions to discourse analysis,[17] and engagement with algorithmic approaches is generally rare in discourse studies and not recognized in its encompassing dimension.[18]

Second, the topics generated by using Latent Dirichlet allocation (LDA) on the *Encyclopédie* articles highlight a circumstance already known from earlier studies:[19] the "subversive classification practices" used by the *Encyclopédistes* to hide potentially controversial content and articles by either not classifying them or classifying them

17 E.g., Landwehr 2018, connecting discourse analysis with cultural history.

18 E.g., Bubenhofer 2018, 208-41. This chapter aptly introduces corpus building and analysis, but it also has to broadly introduce basics (like the TEI) and remains an isolated case, although the author states that there still is a need to "critically reflect the theoretical and methodical implications of corpus-analytical approaches against the background of the digital age" (209; Author's transl.). Although synthetic research is gradually also emerging (Vásquez 2022), a similar picture is also encountered in English-language handbooks, like Handford, Gee 2023, which strongly focuses the digital aspect on the analysis of digital communication.

19 Roe, Gladstone, Morrissey 2015, section "Topic Modelling the French *Encyclopédie*".

under different headings (for instance, Newton's biography can be found in the 'new geography' article on his birthplace, Wolstrope). However, the knowledge about the discursive makeup of individual articles acquired in this way is subsequently not applied, for example in a recent study on the relative chronology of the final ten volumes published all at once in 1765 (Boussuge 2020-21, 287-302).[20]

Finally, "Discourses and Disciplines" also identifies topics that "attest to the function of the *Encyclopédie* as a reference work, concerned, in the most general sense, with making comparisons and distinctions [...], providing the meaning of words [...], and appealing to the authority of the ancients".[21] Apart from the latter assertion, which leads directly to the debate on the actual making of the *Querelle des anciens et des modernes* (Edelstein 2011; Peper, Wallnig 2022), the other findings could certainly provide significant contributions to the understanding of compilatory works of the early-modern period. But while literary studies rarely engage with this type of lexical 'meta-discourses', historical studies dealing with compilation frequently use textual evidence like quotations, references and the like as material for praxeological analysis – rather than for discursive analysis of a text in its entirety.[22]

What we see in this third case study is a situation in which Digital Humanities scholarship with sound methodology[23] and questions directly related to the current state of the scholarly debate is simply not considered in traditional scholarship – not even with the intent of questioning its validity or foregrounding opposing methodological arguments. This corresponds to the – justifiable and reasonable – practice of not considering the bibliography of other disciplines commonly encountered in the social sciences and the various philologies, even when they treat the same historical topic. Where there is no trade, there can be no "trading zone".

20 The approach used in Roe, Gladstone, Morrissey 2015, could, for example, shed additional light on the chronological table on page 295.

21 Roe, Gladstone, Morrissey 2015, section "Topic Modeling the French *Encyclopédie*".

22 Zweifel 2021, see especially the introduction (1-52), in which recent scholarship on encyclopaedias is presented without any reference to digital approaches.

23 Cf. the detailed methodological description in Roe, Gladstone, Morrissey 2015, section "Supervised Machine Learning and the Encyclopedic Disciplines".

5 Outlook

As we have seen in the three described cases, the 'digital turn' seems in many ways to be stuck between the rhetoric of emergence and the reluctant and often clueless stances of traditional historical scholarship; "trading zones", if one seeks them, are rare. The arbitrary collection of material presented in this paper, however, should also have made clear that there is not one individual instance responsible for this situation, but rather that institutional and epistemic constrictions make fruitful interaction rare and often difficult. I would like to highlight two specific points that emerge from the examples provided.

First, the use of the same data by different disciplines can easily lead to methodological misunderstandings. 'Discourse' and 'style' mean different things to the linguist and the historian; a 'network' is something else in historical studies than in sociology or organisational theory. However, how much of the professional discussion on these matters – which in every mentioned case is extensive and detailed – can reasonably be transferred to other disciplines has not yet been fully determined: history has not reasonably digested the explosion of methodological diversity during the past three decades, and thus has good reason to not open the door to yet another set of epistemologies whose centres admittedly lie far outside the humanities in mathematics, statistics and programming.

Second, the few recent cases of gradual integration of algorithm-based scholarship into the core of historical research show that as is often the case, the 'paradigm shift' is likely to happen by way of a few key figures drawing their breakthrough success from the existence of a critical mass whose concerns they will be able to place in relation to the potential of the emerging new approach. However, if such an integration fails to occur, it is also conceivable that the Digital Humanities, isolated among the traditional humanities, could become a separate discipline beside (and thus in epistemic competition with) history – like art history, sociology or linguistics.

I will conclude by admitting that I criticize a scholarly practice in this paper that I myself have engaged in, namely that of drawing tentative conclusions from messy data on vague methodological ground (Wallnig 2020, 441-62; 2021, 207-19; Wallnig, Gasteiner, Bekesi 2023, 79-92). Nonetheless, I (and information theory; Weber, Hasenauer, Mayande 2018) do believe that nescience can best be addressed if made explicit in the first place – and not simply ignored, as is often the case, which also limits the "trading zones" approach to those fields where actual exchange is happening. In the second place, given the undisputable overload of potential research to consider, informed and reflexive collaboration remains the best way to master disciplinary plurilingualism before it results in academic language politics and power struggles (Wyatt et al. 2017, 102):

> one way of reconciling the epistemological contradictions between quantitative and qualitative methods [...] is to recognize that all types of constructs require interpretation in contexts of use, not only by other researchers but also by social actors, including respondents, policy makers, and other audiences.

If traditional scholarship and Digital Humanities should stipulate a 'marriage of convenience',[24] it will be a matter of academic politics – just like any "trading zone" ultimately is.

Bibliography

Ahnert, R. et al. (2020). *The Network Turn*. Cambridge: Cambridge University Press.

Ash, M.G. (2012). "Wissenschaftsgeschichte und Wissenschaftsphilosophie. Einführende Bemerkungen". *Berichte zur Wissenschaftsgeschichte*, 35, 87-98.

Aurnhammer, A.; Steiger, J.A. (Hrsgg) (2019). *Christus als Held und seine heroische Nachfolge. Zur imitatio Christi in der Frühen Neuzeit*. Berlin; Boston: De Gruyter.

Baker, N.S.; Maxson, B.J. (2020). "Where in the World is Renaissance Florence? Challenges for the History of the City after the Global Turn". Baker, N.S.; Maxson, B.J. (eds), *Florence in the Early Modern World. New Perspectives*. London; New York: Routledge, 1-19.

Becker, K.M. (2002). *From the Treasure House of Scripture. An Analysis of Scriptural Sources in De imitatione Christi*. Turnhout: Brepols.

Benz, S. (2003). *Zwischen Tradition und Kritik. Katholische Geschichtsschreibung im barocken Heiligen Römischen Reich*. Husum: Matthiesen.

Boussuge, E. (2020-21). "La chronologie de l'Encyclopédie interdite. Les dix derniers volumes: tomes VIII à XVII (1762-1765)". *Dix-huitième siècle*, 52, 287-302.

Brege, B. (2021). *Tuscany in the Age of Empire*. Cambridge (MA): HUP.

Bubenhofer, N. (2018). "Diskurslinguistik und Korpora". Warnke, I.H. (Hrsg.), *Handbuch Diskurs*. Berlin; Boston: De Gruyter, 208-41.

Cady, S.A. (1990). "The Electronic Revolution in Libraries: Microfilm Déjà Vu?". *College & Research Libraries*, 51(4), 374-86.

Cohn, S.K. Jr. (1999). *Creating the Florentine State. Peasants and Rebellion, 1348-1434*. Cambridge: CUP.

Edelstein, D. (2011). *The Enlightenment. A Genealogy*. Chicago: CUP.

Enderlin-Mahr, A. (2022). "Akteure und Netzwerke im Umfeld der k. u. k. Kabinettskanzlei". Ableidinger, C. et al. (Hrsgg), *Im Büro des Herrschers*. Göttingen: Vandenhoeck & Ruprecht, 35-48.
https://theemperorsdesk.univie.ac.at/

Füssel, M.; Neu, T. (2021). "Reassembling the Past?!". Füssel, M.; Neu, T. (Hrsgg), *Akteur-Netzwerk-Theorie und Geschichtswissenschaft*. Paderborn: Schöningh, 1-25.

Gladstone, C. (2020). *Rousseau et le matérialisme*. Liverpool: Liverpool University Press. Oxford University Studies in the Enlightenment.

24 The term is used by Mitchell Ash with regard to the relation between the history and philosophy of science, their different epistemologies and the fact that they are often merged because of institutional pressure (2012, 88).

Gladstone, C.; Cooney, C.M. (2020). "Opening New Paths for Scholarship: Algorithms to Track Text Reuse in ECCO". Burrows, S.; Roe, G. (eds), *Digitizing Enlightenment. Digital Humanities and the Transformation of Eighteenth-Century Studies*. Oxford: Voltaire Foundation, 353-74. Oxford University Studies in the Enlightenment.

Gramsch-Stehfest, R. (2013). *Das Reich als Netzwerk der Fürsten. Politische Strukturen unter dem Doppelkönigtum Friedrichs II. und Heinrichs (VII.) 1225-1235*. Ostfildern: Thorbecke.

Habsburg, M. (2017). *The Devotional Life. Catholic and Protestant Translations of Thomas a Kempis' Imitatio Christi, c. 1420-c. 1620* [PhD Dissertation]. St. Andrews: University of St. Andrews.

Handford, M.; Gee, J.P. (eds) (2023). *The Routledge Handbook of Discourse Analysis*. London: Routledge.

Harrap, D.A. (2017). *The Phenomena of Prayer: The Reception of the Imitatio Christi in England (1438–c.1600)* [PhD Dissertation]. London: Queen Mary University London.

Hochedlinger, M.; Mat'a, P.; Winkelbauer, T. (Hrsgg) (2019). *Verwaltungsgeschichte der Habsburgermonarchie in der Frühen Neuzeit*. Bd. 1, *Hof und Dynastie, Kaiser und Reich, Zentralverwaltungen, Kriegswesen und landesfürstliches Finanzwesen*. Vienna: Böhlau.

Hotson, H.; Wallnig, T. (eds) (2019). *Reassembling the Republic of Letters in the Digital Age. Standards, Systems, Scholarship*. Göttingen: GUP.
https://univerlag.uni-goettingen.de/handle/3/isbn-978-3-86395-403-1

Kemman, M. (2021). *Trading Zones of Digital History*. Berlin; Boston: De Gruyter.

Kemman, M.; Kleppe, M.; Scagliola, S. (2014). "Just Google It". Mills, C.; Pidd, M.; Ward, E. (eds), *Proceedings of the Digital Humanities Congress 2012*. Sheffield: The Digital Humanities Institute.
https://www.dhi.ac.uk/books/dhc2012/just-google-it/

Kittelmann, J.; Purschwitz, A. (2019). *Aufklärungsforschung digital: Konzepte, Methoden, Perspektiven*. Halle (Saale): Mitteldeutscher Verlag. Kleine Schriften, IZEA 10.

Landwehr, A. (2018). *Historische Diksursanalyse*. Frankfurt; New York: Campus.

Najemy, J.M. (2008). *A History of Florence, 1200-1575*. Malden (MA): Blackwell.

Oakes, M. (1998). *Statistics for Corpus Linguistics*. Edinburgh: EUP.

O'Neil, C.; Schutt, R. (2013). *Doing Data Science. Straight Talk from the Front Line*. Sebastopol (CA): O'Reilly.

Padgett, J.F.; Ansell, Ch.K. (1993). "Robust Action and the Rise of the Medici, 1400-1434". *American Journal of Sociology*, 98, 1259-319.
https://home.uchicago.edu/~jpadgett/papers/published/robust.pdf

Peper, I.; Wallnig, T. (eds) (2022). *Central European Pasts. Old and New in the Intellectual Culture of Habsburg Europe*. Berlin; Boston: De Gruyter.

Petz, C. (2022). "A (Not So) Short History of Historical Network Research. Part 3". *Digital Humanities Lab*, 9 September.
https://dhlab.hypotheses.org/3194

Prajda, K. (2018). *Network and Migration in Early Renaissance Florence, 1378-1433. Friends of Friends in the Kingdom of Hungary*. Amsterdam: AUP.

Reinhard, W. (ed.) (1996). *Power Elites and State Building*. New York: Clarendon Press. The Origins of the Modern State in Europe, 13th to 18th Centuries 2.

Reinhard, W. (2007). *Geschichte des modernen Staates. Von den Anfängen bis zur Gegenwart*. Munich: C.H. Beck.

Roe, G.; Gladstone, C.; Morrissey, R. (2015). "Discourses and Disciplines in the Enlightenment: Topic Modeling the French *Encyclopédie*". *Frontiers in Digital Humanities*, 2.
https://doi.org/10.3389/fdigh.2015.00008

Roelli, Ph. (ed.) (2020). *Handbook of Stemmatology. History, Methodology, Digital Approaches*. Berlin; Boston: De Gruyter.
https://doi.org/10.1515/9783110684384

Rutgers, P.D. (2005). "Patronage, Citizenship, and the Stalled Emergence of the Modern State in Renaissance Florence". *Comparative Studies in Society and History*, 47(3), 638-64.
https://doi.org/10.1017/S0010417505000289

Spoerhase, C.; Martus, S. (2022). *Geistesarbeit. Eine Praxeologie der Geisteswissenschaften*. Berlin: Suhrkamp.

Taddei, I. (2004). "Le système politique florentin au XVe siècle". Boutier, J.; Landi, S.; Rouchon, O. (éds), *Florence et la Toscane, XIVe–XIXe siècles. Les dynamiques d'un État italien*. Rennes: Presses universitaires de Rennes, 39-62.
https://books.openedition.org/pur/15775

Tavanti, M.; Wilp, E.A. (2019). "Humanistic Renaissance for Good. Leadership Lessons from Florence to Silicon Valley". Stachowicz-Stanusch, A. et al. (eds), *Humanistic Values from Academic Community Perspective*. Charlotte (NC): Information Age Publishing, 221-40.

Vásquez, C. (ed.) (2022). *Research Methods for Digital Discourse Analysis*. London: Bloomsbury.

Wallnig, T. (2020). "Distant Reading Austria. Ein Essay über die Habsburgermonarchie des langen 18. Jahrhunderts und die digitale Transformation der Geschichtswissenschaften". Seitschek, F.-S; Hertel, S. (Hrsgg), *Herrschaft und Repräsentation in der Habsburgermonarchie (1700-1740). Die kaiserliche Familie, die habsburgischen Länder und das Reich*. Berlin; Boston: De Gruyter, 441-62.

Wallnig, T. (2021). "Wissen in Wien, 1780: digitale Annäherungen". Fischer, N.; Mader-Kratky, A. (Hrsgg), *Schöne Wissenschaften. Sammeln, Ordnen und Präsentieren im josephinischen Wien*. Vienna: Verlag der ÖAW, 207-19.

Wallnig, T.; in collaboration with Bekesi, J.; Gasteiner, M. (2023). "Digital Humanities and Wissenschaftsgeschichte: A New Unity (Called Txt)?". Hunger, H. (Hrsg.), *Einheit oder Vielheit. Über Methode und Gegenstand in der Geschichte und Philosophie der Wissenschaften*. Vienna: Verlag der ÖAW, 79-92. Forschung und Gesellschaft 23.

Wyatt, S.; Milojević, S.; Park, H.W.; Leydersdorff, L. (2017). "Intellectual and Practical Contributions of Scientometrics to STS". Felt, U. et al (eds), *The Handbook of Science and Technology Studies, Fourth Edition*. Cambridge (MA): MIT Press, 87-112.

Weber, Ch.M.; Hasenauer, R.P.; Mayande, N.V. (2018). "Toward a Pragmatic Theory for Managing Nescience". *International Journal of Innovation and Technology Management*, 15(5).
https://doi.org/10.1142/S0219877018500451

Yates, F. (1952). "George Udny Yule (1871-1951)". *Obituary Notices of the Fellows of the Royal Society*, 8(21), 309-23.

Yule, G.U. (1946). "Cumulative Sampling. A Speculation as to What Happens in Copying Manuscripts". *Journal of the Royal Statistical Society*, 109(1), 44-52.

Zweifel, S. (2021). *Aus Büchern Bücher machen. Zur Produktion und Multiplikation von Wissen in frühneuzeitlichen Kompilationen*. Munich; Vienna: De Gruyter.

Models of Data Extraction and Architecture in Relational Databases of Early Modern Private Political Archives
edited by Dorit Raines

The Representation of Historical Uncertainties as the Outcome of Competing and Incompatible Certainties

Fabio Vitali
Alma Mater Studiorum - Università di Bologna, Italy

Valentina Pasqual
Alma Mater Studiorum - Università di Bologna, Italy

Abstract This study delves into the ontological implications associated with 'data' and 'capta', highlighting the profound importance of contextual factors and interpretation within the realm of critical discourse. More specifically, 'data', encapsulating the notion of objective observations of facts, differ from 'capta', which encompass selections, opinions, controversies, and debates. These capta elements contribute interest and added value to scholarly knowledge. The formalised separation of data and capta proves to be pivotal in shaping research practices conducive to a critical approach to knowledge production. Upon an evaluation of the efficacy of existing technologies in addressing this issue, the Conjectures model is introduced as a potential solution. This model aims to explicitly represent capta while concurrently facilitating the evolution of critical discourse within a one Knowledge Graph designed to capture complex information within a particular domain. To underscore our point, an illustrative example is drawn from the field of historiography.

Keywords Conjectures. GLAM. Provenance. Uncertainty. RDF.

Summary 1 Introduction. – 2 Related Works. – 3 Expressing Capta with Semantic Web Technologies. – 3.1 RDF Reification. – 3.2 N-ary Relations and CIDOC CRM. – 3.3 Wikidata Statements and Ranking. – 3.4 Named Graphs. – 3.5 RDF-star. – 4 Expressing Capta with Conjectures. – 5 Conclusions.

Edizioni Ca'Foscari

Studi di archivistica, bibliografia, paleografia 8
e-ISSN 2610-9093 | ISSN 2610-9875
ISBN [ebook] 978-88-6969-919-1 | ISBN [print] 978-88-6969-920-7

Peer review | Open access
Submitted 2023-10-03 | Accepted 2024-01-18 | Published 2025-05-22

DOI 10.30687/978-88-6969-919-1//003

1 Introduction

The motivation that propels the scholar is not solely confined to the pursuit of distinguishing true from false. Rather, it extends to the discernment, within the realm of veracity, of aspects that elicit interest, intrigue, the capacity to narrate a compelling storyline, and the potential to yield significant implications for scholarly work. This implies that our best scholarly stories are not simply the final results of our studies, but also, and most importantly, the narration of how we have reached them. Furthermore, in order to narrate these stories, we must represent the scholarly activity in its complexity, including the points of view and opinions we have worked on, discussed, objected to, found lacking or false, or, possibly, rescued from obscurity and discredit.

In this paper our discourse revolves around methodologies and models, all with the aim of not solely encapsulating the final result of a scholarly investigation, but potentially encompassing the building blocks that substantiate it. This ultimate aim is intended to render the result accepted and appreciated in the full strength of its scholarly complexity. We therefore deem it highly important to correctly represent the two activities of collecting and interpreting data that scholars regularly perform. To this end, we turn to Checkland and Holwell (2005), who distinguish between two distinct approaches to data collection, named 'data' and 'capta'.

'Data' refers to the mass of facts that are given and observed. The term comes from the Latin word *dare*, meaning 'to give'. Data is often collected through observation and measurement of external reality (e.g., the material, the dimensions, the current location of a painting), and while individual data points may not be particularly interesting on their own, it is their collection and analysis that may become important. This approach to data collection is often associated with a realist perspective, where data is seen as an objective representation of the world.

On the other hand, 'capta' refers to the small fraction of the available data that we actively take, filter, select, and interpret. The term comes from the Latin word *capere*, meaning 'to take'. Capta is often associated with a constructivist perspective, where knowledge is viewed as constructed and subjective (e.g., the attribution of the same painting to a specific artist or artistic school). Capta involves searching for relevant information, filtering out irrelevant data, selecting the most important information, and interpreting it based on our situated, partial, and constitutive knowledge. This process of knowledge construction is often seen as a more humanistic approach to research, as it acknowledges the role of the researcher in shaping the knowledge that is produced.

Johanna Drucker (2011) further expands on this distinction, arguing that data and capta have different ontological implications.

Data is seen as representing pre-existing facts, while capta is seen as representing situated, partial, and constitutive knowledge of a constructed nature involving selections, interpretations, and expressions of opinions and points of view. This distinction has important implications for knowledge production, as it challenges the notion of data as an objective representation of the world and highlights the importance of context and interpretation in shaping our understanding of the world.

Despite these differences, to this date both data and capta are inserted and described in the same digital collections without differentiation. This approach neglects the correct representation of complexities in knowledge construction and limits the ability to engage in more nuanced and critical analysis. We argue that it is possible to develop a better critical approach to knowledge production by recognising the differences between data and capta and by incorporating this understanding into research practices. This approach can involve paying attention to one's own situated knowledge and perspectives, recognising the role of interpretation and context in shaping our understanding of the world, and being open to alternative perspectives and interpretations. Ultimately, this approach can lead to a computable, and a more robust model of the world and of the issues that we seek to understand.

Consider for instance the history of the evolving attributions of the (now known as) Hadrian's Wall as our guiding example. Flavius Eutropius, a historian who lived in Rome and Constantinople between 363 and 387 AD, wrote about the building of a wall in the northern part of Britain, under the orders of Septimius Severus, the Roman emperor who reigned between 193 and 211 AD. However, Eutropius had never been to Britain, and had no direct knowledge of the wall: he simply attributed the wall to Severus because he knew that he had been to Britain and died there. The venerable Bede, the English monk and historian who lived between 672 and 735 AD, tried to make sense of Eutropius' claim of a wall built by Severus, and since he knew the one built much further north by Antoninus Pius (142 AD), he inferred that the wall by Severus mentioned by Eutropius was the one situated further south. Given Bede's reputation and weight, his words and attribution were accepted for centuries, even though here and there a few historians raised some doubts.

The turn of the tide arrives in 1840, when the priest and historian John Hodgson, while writing his *magnus opus, History of Northumberland*, challenged the long-standing attribution of the wall to Severus. Given that his book held only peripheral relevance to the topic of the wall, he allocated a solitary footnote in the third volume to express his doubts. Being a thorough scholar, this footnote is a remarkable 173 pages long ground-breaking evidence supporting the attribution of the wall to Hadrian rather than Severus. Hodgson's

ideas were initially misread or ignored, but eventually became accepted and adopted as the correct attribution of the wall, until today.

The data in this story are few, and possibly uninteresting: the Hadrian's Wall is 117 km long, about 2.4 m high, and runs from Wallsend to Bowness-on-Solway. But here the capta is the real story. It concerns ancient superficial reporting, imprecise deduction from erroneous information, deference to historical (holy, even) authority preventing the obvious truth to emerge, the stubbornness of one man against centuries of accepted facts, the heterogony of ends in scholarly works, and much more. Reporting just the facts unearths a fraction of the story, and the uninteresting bits of it anyway.

At the same time, uncertainty is a common challenge that scholars face in dealing with capta. Uncertainty can arise from various causes, such as temporary ignorance or evolving information. Scholars must also contend with disagreements and controversies, carry out carefully thought experiments where likely and unlikely scenarios are examined, and many other scholarly challenges that introduce uncertainties in their scientific hypothesis. In such situations, avoiding being reticent is crucial (Barabucci et al. 2021).

Competing datings, attributions, and interpretations are examples of the uncertainties that scholars often face. Competing datings refer to instances where multiple timelines or chronological sequences are proposed for a particular event or object. The implications of such doubts can be significant, as they can impact the understanding and implications of related facts and captas. Similarly, competing attributions refer to instances where multiple authors or creators are proposed for a particular work of art or literature under examination. Their impact and ramifications can extend beyond the question of originality, even impacting the monetary value of the item. Finally, competing interpretations refer to instances where multiple meanings or implications are proposed for a particular piece of information or capta. Their implications and ramifications can be significant, as they can impact the understanding of the nature and relevance of the item.

In this paper we intend to introduce a formalisation of capta utilising Semantic Web technologies. In section 2, several existing approaches have been assessed. In section 3, these approaches have been analysed and their advantages and disadvantages highlighted. In section 3, our own formalisation, called Conjectures, is proposed. In section 4, our findings are outlined, and our conclusions are drawn with some insights for further developments.

2 Related Works

Uncertainty ontologies focus on addressing uncertainties that arise from competing and incompatible certainties in datasets. They operate under the assumption that they are dealing with single data items, whose factuality may not be fully known, with the purpose of measuring and reasoning about quantified uncertainty. Conversely, when facing capta, our underlying premise is that we are dealing with a plurality of competing assertions, each accompanied by its distinct degree of factuality, rendering their simultaneous acceptance untenable. The uncertainty inherent in capta arises from the acknowledgment that we are obliged to contend with multiple points of view, some of which may not be factual or true.

The issue of uncertainty is a unique situation that has recently received significant attention. The nature of uncertainty has been expressed by the W3C Incubator Group on Uncertainty Reasoning (Laskey et al. 2008) in terms of epistemic versus aleatory, objective versus subjective, and contingent versus generic. This uncertainty emerges due to ambiguous, inconsistent, vague, incomplete, or empiric information.

To address this challenge, various ontologies have been developed, such as URREF (Blash et al. 2019), which provide a means to express and reason upon uncertainty. These ontologies offer a way to represent uncertainty in a structured and organised manner, allowing for more effective analysis and decision-making. By utilising these tools, researchers can better navigate the complexities of uncertain data and draw meaningful insights from them. The ontologies have been developed based on the assumption that datasets accurately identify the level, nature, and cause of uncertainty. Consequently, they target data and are designed to measure and reason about quantified uncertainty in single data items, even if their factuality is not known.

However, this approach falls short when dealing with capta, as capta represent multiple competing assertions whose factuality cannot all be accepted. As such, our target is not data but rather capta, and our assumption is that we must deal with the complexity of these multiple competing assertions. Our purpose is to represent this complexity and to acknowledge that uncertainty arises when dealing with multiple points of view that cannot all be accepted as factual and true.

Through the years, several models have been developed to express competing claims in the same Knowledge Base as Reification (Hayes 2004), N-ary relations (Noy, Rector 2006) used for example by CIDOC CRM (Doerr et al. 2007), Singleton Properties (Nguyen et al. 2015), and more recently RDF-star (Hartig 2017), but no common strategy has been defined by the community yet. The way of best expressing capta with either of these models heavily depends on the

complexity of the information to be rendered, and the kind of queries on large datasets of uncertain assertions needed to be performed.

In the next section we shall examine how five different syntaxes can express the Hadrian Wall's controversy, and the relative advantages and disadvantages of each.

3 Expressing Capta with Semantic Web Technologies

In this section, the concurring claims about Hadrian's Wall concurring attributions are formalised on behalf of several models, namely reification, n-ary relations, Wikidata statements, named graphs and RDF-star quoted triples.

These statements all express two (competing) sets of statements:

1. According to the Venerable Bede, in his *Historia Ecclesiastica* (702 AD), the wall was built by emperor Septimius Severus in 189 AD.
2. According to John Hodgson, in his *History of Northumbria* (1840), the wall was built by emperor Hadrian in 122 AD.

3.1 RDF Reification

```
:S1 a rdf:Statement.                          :S2 a rdf:Statement.
:S1 rdf:Subject :Wall.                        :S2 rdf:Subject :Wall.
:S1 rdf:Predicate p:creator.                  :S2 rdf:Predicate dc:creator.
:S1 rdf:Object :Hadrian.                      :S2 rdf:Object :Severus.
:S1 prov:wasAttributedTo :Hodgson.            :S2 prov:wasAttributedTo :Bede.
:S1 prov:wasDerivedFrom :HistOfNorthumbria.   :S2 prov:wasDerivedFrom :HistEcclesiastica.
:S1 prov:atTime"1840"^^xsd:Year.              :S2 prov:atTime"702"^^xsd:Year.

:S3 a rdf:Statement.                          :S4 a rdf:Statement.
:S3 rdf:Subject :Wall.                        :S4 rdf:Subject :Wall.
:S3 rdf:Predicate p:inception.                :S4 rdf:Predicate p:inception.
:S3 rdf:Object "122"^^xsd:Year.               :S4 rdf:Object "189"^^xsd:Year..
:S3 prov:wasAttributedTo :Hodgson.            :S4 prov:wasAttributedTo :Bede.
:S3 prov:wasDerivedFrom :HistOfNorthumbria.   :S4 prov:wasDerivedFrom :HistEcclesiastica.
:S3 prov:atTime"1840"^^xsd:Year.              :S4 prov:atTime"702"^^xsd:Year.
```

Figure 1 Disputed attributions and dating of the wall expressed with reification

The concurring claims are represented in figure 1 through reification (Hayes 2004) **[fig. 1]**, where S1 is a statement attributed to John Hodgson, whose subject is the 'Wall', whose predicate is being 'created

by' (p:creator), whose object is 'Hadrian'. S2 is a statement attributed to Bede, whose subject is the 'Wall', whose predicate is 'being created by', whose object is 'Severus'. The same patterns are followed to encode the concurring dating of the wall (S3 and S4), showing that a high number of triples is required to express such concurring claims; in particular each claim is represented by at least 4 triples each, for a grand total of 28 statements. Additionally, reified triples require the introduction of a 'fictitious' class (rdf:Statement) to express the claims, requiring the repetition of contextual information for each claim (e.g. who made the claim, when the claim was made and in which edition it has been published).

3.2 N-ary Relations and CIDOC CRM

In figure 2, the same concurring claims can be represented with n-ary relations (Noy, Rector, 2006) and the CIDOC CRM ontology (Doerr et al. 2007) **[fig. 2]**:

```
:Wall a crm:E24_Physical_Human-Made_Thing   :P1  a crm:E12_Production;
                                              crm:P108_has_produced :Wall.

:A1 a crm:E13_Attribute_Assignment;         :A2 a crm:E13_Attribute_Assignment;
   crm:P177_assigned_property_of_type          crm:P177_assigned_property_of_type
crm:P14_carried_out_by;                     crm:P14_carried_out_by;
   crm:P140_assigned_attribute_to :P1;         crm:P140_assigned_attribute_to :P1;
   crm:P141_assigned :Hadrian                        crm:P141_assigned :Severus;
   crm:P14_carried_out_by :Hodgson.                  crm:P14_carried_out_by :Bede.

:A3 a crm:E13_Attribute_Assignment;         :A4 a crm:E13_Attribute_Assignment;
   crm:P177_assigned_property_of_type          crm:P177_assigned_property_of_type
crm:P4_has_time_span;                       crm:P4_has_time_span;
   crm:P140_assigned_attribute_to :P1;         crm:P140_assigned_attribute_to :P1;
   crm:P141_assigned "122"^^xsd:Year;          crm:P141_assigned "189"^^xsd:Year;
   crm:P14_carried_out_by :Hodgson.            crm:P14_carried_out_by :Bede.
```

Figure 2 Disputed attributions and dating expressed with CIDOC CRM (using n-ary relations)

According to CIDOC CRM, as shown in figure 2, the wall is a man-made thing (crm:E24_Physical_Human-Made_Thing). A production event P1 that produced the wall exists. An attribution A1 exists and it regards the identity of the person who realised the event P1, which is 'Hadrian' according to John Hodgson. An attribution A2 exists, referring to the person who realised the event P1, which is 'Severus', according to Bede. An attribution A3 exists, recording the time when the event

P1 occurred, which is the year '112', according to Hodgson. An attribution A4 exists, recording the time when the event P1 occurred, which is the year '189', according to Bede. Similarly to reification, n-ary relations require the inclusion of a 'fictitious' class to express each claim (crm:E13_Attribute_Assignment), and a high number of triples (23). Furthermore, their provenance information needs to be repeated to each claim (e.g., A1 and A3 are both attributed to Hodgson while A2 and A4 are attributed to Bede).

3.3 Wikidata Statements and Ranking

Similarly to the previous cases (reification and n-ary), by using Wikidata syntax and model, as shown in figure 3, we represent each claim as a wikibase:Statement (Erxleben et al. 2014) **[fig. 3]**. For the sake of clarity, we replaced the numerical ids used by Wikidata with meaningful labels. Wikidata has a rich way of representing provenance, including individuals, sources and dates. Additionally, it has interesting ways of reporting both claims asserted as preferred and claims proposed as deprecated through the use of ranking over statements and other syntactic methods. Unfortunately, deprecated and preferred statements are less than 2% out of all statements in Wikidata. Additionally, many preferences simply represent corrections of typos, changes in location or of inventory numbers. Furthermore, the 'nature of statement' predicates use a rich vocabulary to characterise statements. Of the more than 260 different types of natures of statements present in Wikidata, we counted about 60 dealing with uncertainty, such as approximation, controversy, 'disputed', 'dubious', estimate, guess, hypothesis, 'presumably', 'possibly', speculation, 'unconfirmed', 'unknown', 'unofficial', etc. Yet, 'nature of statement' predicates are very rare; they were added to fewer than 0.1% of the collected sample. Their scarcity, the vastly overlapping semantics and labelling of their values does not help in increasing their presence (Di Pasquale et al. 2024).

The first claim (S1) assigns a creator to the wall using the property wdt:creator and the value :Hadrian. Additionally, it provides a reference to support this claim using the Wikidata property wikibase:Reference. The reference is assigned the identifier ref:REF1 and includes the following information: it was stated in the publication *History Of Northumbria*, authored by :Hodgson, and published in the year '1840'. The same pattern is implied with statements S2, S3 and S4. As with reification and n-ary relations, each statement is independent and each must be repeated independently and independently attributed to the corresponding source (respectively, ref:REF1 and ref:REF2).

```
ref:REF1 a wikibase:Reference ;
  pr:statedIn :HistOfNorthumbria;
   pr:author :Hodgson;
   pr:publicationDate "1840"^^xsd:Year .

:Wall wdt:creator :Hadrian.
:Wall p:creator s:S1.
s:S1 a wikibase:Statement;
  wikibase:rank wikibase:PreferredRank;
  ps:creator :Hadrian;
  pq:natureOfStatement :attribution;
  prov:wasDerivedFrom ref:REF1.


:Wall wdt:inception "122"^^xsd:Year.
:Wall p:inception s:S3.
s:S3 a wikibase:Statement;
  wikibase:rank wikibase:PreferredRank;
  ps:inception "122"^^xsd:Year;
  pq:natureOfStatement :attribution;
  prov:wasDerivedFrom ref:REF1.
```

```
ref:REF2 a wikibase:Reference ;
   pr:statedIn :HistEcclesiastica;
   pr:author :Bede
   pr:publicationDate "702"^^xsd:Year .

# no assertion since this is deprecated
:Wall p:creator s:S2.
s:S2 a wikibase:Statement;
  wikibase:rank wikibase:DeprecatedRank;
  ps:creator :Severus;

  prov:wasDerivedFrom ref:REF2.


# no assertion since this is deprecated
:Wall p:inception s:S4.
s:S4 a wikibase:Statement;
  wikibase:rank wikibase:DeprecatedRank;
  ps:inception "189"^^xsd:Year;

  prov:wasDerivedFrom ref:REF2.
```

Figure 3 Disputed attributions and datings of the wall expressed with Wikidata statements

In contrast to previous cases, Wikidata distinguishes between accepted claims (S1 and S3, by Hodgson) and discarded ones (S2 and S4, by Bede). Both claims regarding the Hodgson attribution are ranked as preferred (wikibase:PreferredRank), marking the claims as 'accepted', and therefore stating their logical status validity. Additionally, in order to assert S1 and S3 a new triple is added to both (respectively, :Wall wdt:creator:Hadrian and :Wall wdt:inception "122"^^xsd:Year), meaning that this information can be retrieved by using simple queries in Wikidata SPARQL endpoint. On the other hand, deprecated claims (S2 and S3, by Bede) are marked as wikibase:DeprecatedRank, denoting the claims as 'not accepted', and therefore 'not true anymore'. Additionally, no triple is added to S2 and S4 claims, making them non-asserted claims, meaning that this information cannot be retrieved by using simple queries in Wikidata SPARQL endpoint. In general, rank is not only used for accepted/non-accepted assertions, but also for, e.g., temporal statements (a painting used to be in museum A, with DeprecatedRank, but it is now in museum B (with PreferredRank).

3.4 Named Graphs

"Named Graphs" (Carroll et al. 2005) provides a succinct model compatible with RDF 1.1 (2014) to express claims and their provenance information **[fig. 4]**. Differently from previous cases, the two attributions can be grouped in just one graph so that each graph shows its provenance just once, thereby reducing the number of triples needed to express the claims. However, the claims in figure 4 can be seen as both equally asserted (read it as 'the wall was created by Hadrian in 122 according to Hodgson and at the same time the wall was created by Severus in 189 according to Bede'), and the provenance is just some kind of additional information. Named graphs have a great potential of expressivity, but the lack of control over whether the content of the graph is asserted or not somehow prevents a full exploitation of their potentialities.

```
GRAPH :S1 {                                  GRAPH :S2 {
   :Wall :creator :Hadrian;                     :Wall :creator :Severus;
      :inception "122"^^xsd:Year.                  :inception "189"^^xsd:Year.
}                                            }
:S1 prov:wasAttributedTo :Hodgson;           :S2 prov:wasAttributedTo :Bede;
   prov:wasDerivedFrom :HistOfNorthumbria.      prov:wasDerivedFrom :HistEcclesiastica.
```

Figure 4 Disputed attributions and datings of the wall expressed with named graphs

3.5 RDF-star

RDF-star (Hartig 2017) extends RDF 1.1 (2014) syntax in order to express statements without asserting them, representing claims in angle brackets. The statement << The wall was created by Hadrian >> is attributed to Hodgson. The statement << The wall was created in 122 >> is attributed to Hodgson. The statement << The wall was created by Severus >> is attributed to Bede. The statement << The wall was created in 189 >> is attributed to Bede. Neither quoted statement is asserted. In order to assert one of them, the claim is repeated outside of the quote (:Wall p:creator:Hadrian and :Wall p:inception "122"^^xsd:Year). Moreover, every single quoted assertion must be associated separately to each provenance statement implying that there is no simple way to avoid repetitions as with reification **[fig. 3]**, n-ary relations **[fig. 4]** and Wikidata **[fig. 5]**.

In summary, all the above methods for representing capta have interesting characteristics but also some limitations. One issue is the necessity to incorporate numerous additional fictitious entities such as statements, activities, and events, which can lead to an

unnecessarily large number of statements (e.g., see rdf:Statement, crm:E13_Attribute_Assignment, wikibase:Statement). It is somewhat challenging to distinguish between asserted statements (predicates presented as true facts) and those that are only expressed (predicates not associated with a truth value).

```
:Wall :creator :Hadrian.
<< :Wall :creator :Hadrian. >>                 << :Wall :creator :Severus. >>
   prov:wasAttributedTo :Hodgson;                 prov:wasAttributedTo :Bede;
   prov:wasDerivedFrom :HistOfNorthumbria.        prov:wasDerivedFrom :HistEcclesiastica.

:Wall :inception "122"^^xsd:Year.
<< :Wall :inception "122"^^xsd:Year. >>        << :Wall :inception "189"^^xsd:Year. >>
   prov:wasAttributedTo :Hodgson;                 prov:wasAttributedTo :Bede;
   prov:wasDerivedFrom :HistOfNorthumbria.        prov:wasDerivedFrom :HistEcclesiastica.
```

Figure 5 Disputed attributions and datings of the wall expressed with RDF-star

Moreover, ontologies tend to include specialised entities for even the most straightforward scenarios, resulting in more complex and indirect representations that require a greater number of statements whose veracity is difficult to determine.

Each Wikidata statement uses 3 triples to express a claim, but differently from reification and n-ary relations, rankings allow one to distinguish between currently accepted and deprecated claims. While the named graphs method uses just one quadruple to express each claim, its semantics do not distinguish explicitly between currently accepted and deprecated claims. RDF-star uses one quoted triple to express a claim and allows for distinction between currently accepted claims (asserted claims) and deprecated claims (non-asserted).

Finally, controversial and debated propositions are often expressed individually (e.g., one reification block per proposition, one n-ary relation per proposition, one Wikidata ranked statement per proposition, one RDF-star per proposition) and no relation is explicitly provided to connect multiple claims belonging to a single theory (e.g., the attribution of the creator of the wall goes hand in hand with the inception date and they cannot be chosen independently). The endeavour of identifying uncertainties, ambiguities, and complex situations is thus exceptionally demanding. These challenges underscore the necessity for a more comprehensive and flexible approach to representing capta, one that can account for the numerous nuances and complexities inherent in these phenomena.

4 Expressing Capta with Conjectures

Our proposal, introduced in 2021 (Daquino et al. 2022), aims to provide a more nuanced approach to representing capta by using a specialisation of named graphs (Carroll et al. 2005). Unlike traditional named graphs, Conjectures do not assert the truth of their content by construction. Instead, we divide statements into three categories based on the level of agreement or disagreement surrounding them.

1. 'Undisputed statements' are those that have not been doubted by anyone so far, although this does not necessarily mean that the claim is true. Such statements are represented using plain RDF 1.1 named graphs.
2. 'Disputed statements' are those that have at least one known source casting doubts on the claim or providing incompatible and competing claims. To represent such statements, we use Conjectural Graphs, which allow for multiple competing claims to coexist within the same graph.
3. 'Settled statements' are those in which the author of the dataset has opted for a single claim over competing alternatives, even while acknowledging the existence of disagreement. These statements are represented using Collapsed Conjectural Graphs.

A complete formal model of Conjectures has been developed. It demonstrates the correctness of our approach both as an extension of RDF 1.1 in its strong form **[fig. 6]** and within plain RDF 1.1 in its weak form **[fig. 7]**.

In the strong form, undisputed statements are stored into plain named graphs. Each named graph is introduced by the keyword GRAPH and the content of each of them is asserted. Consider for example the claims in C1 (Wall:length "117km" and:Wall:height "2.4m") in figure 8, which was never questioned.

Disputed claims are stored into special named graphs introduced by the keyword CONJ, meaning that their content is not asserted. For example, the claim :C2 in figure 8 stores :Wall p:creator:Severus and p:inception "189"^^xsd:Year. In this case, both claims are being part of the now discarded attribution of the wall and therefore recorded as non-asserted. Additionally, with the use of named graphs, it is possible to group those claims which belong to the same theory (e.g., in this case, the wall attribution) and record the provenance referring to both the dating and the attribution of the wall (e.g., :C2 prov:wasAttributedTo:Bede).

Settled claims are stored into special named graphs introduced by the keyword SETTLED CONJECTURE, meaning that their content is asserted while being disputed in the past.

```
GRAPH :C1 {
   :Wall:length "117km".
   :Wall:height "2.4m".
}
CONJ :C2 {
   :Wall:creator:Severus.
   :Wall:inception "189"^^xsd:gYear.
}
:C2 prov:wasAttributedTo:Bede;
   prov:wasDerivedFrom:HistEcclesiastica.
SETTLED :C3 {
   :Wall:creator:Hadrian;
   :Wall:inception "122"^^xsd:gYear.
}
:C3 prov:wasAttributedTo:Hodgson;
   prov:wasDerivedFrom:HistOfNorthumbria.
```

Figure 6 Undisputed, disputed and settled attributions and datings of the wall expressed with Conjectures (strong form)

Figure 6 contains the representation of the claims about the wall dimensions (undisputed claim), the discarded attribution of the wall to Severus by Bede (disputed claim, definitely non-asserted) and the accepted attribution of the wall to Hadrian by Hodgson (disputed but settled claim, part of the debate but definitely asserted) with Conjectures in the strong form.

The weak form of Conjectures presents identical information as the strong form but adheres to the syntax of RDF 1.1.[1] As in the strong form, undisputed statements are stored into plain named graphs as shown in figure 7. Disputed claims on the other hand are formalised in the weak form according to the following definition (Daquino et al. 2022):

> A conjectural graph is a named graph where all triples (*s*, *p*, *o*) are represented with two triples, (*s*, *cp*, *o*) and (*cp*, *conj*:*isAConjecturalFormOf*, *p*), where *cp* is a unique newly minted predicate created specifically for the triple to conjecture.

Conjectures adopt newly-minted predicates used only once, which are mapped to their original predicate via the property *conj:isAConjecturalFormOf*. Similarly to :singletonPropertyOf (Nguyen et al. 2015), the property allows to easily retrieve original predicates. For example, as shown in figure 7, the attribution (:Wall:creator:Severus and :Wall:inception "189"^^xsd:Year) claimed by Bede (graph :C2) is now represented by four triples and each original predicate (:creator and :inception) is mapped to the newly minted predicates (C2:creator and C2:inception) by conj:isAConjecturalFormOf.

Settled claims are recorded according to the following definition (Daquino et al. 2022):

1 A parser of Conjectures in their weak and strong form is available at `http://conjectures.altervista.org/`.

A collapse graph *c*1 consists of two graphs: the first is the conjecture *c*1 and the second is a new graph *cc*1 including all the triples in *c*1 but with their original predicates, excluding *conj*: *isAConjectualFormOf*, and adding the triple (*cc*1, *conj*: *collapses*, *c*1).

As shown in figure 7, the claim by Hodgson is represented with two graphs **[fig. 7]**: :C3 represents the claim in its conjectural form, while :settlementOfC3 asserts the claim. The addition of the triple (:settlementOfC3 conj:settles:C3) explicitly links the two graphs.

```
GRAPH:C3 {
   :Wall C3:creator:Hadrian;
   :Wall C3:inception “122”^^xsd:gYear.
   C3:creator conj:isAConjecturalFormOf:creator.
   C3:inception conj:isAConjecturalFormOf:inception.
}
:C3 prov:wasAttributedTo:Hodgson;
   prov:wasDerivedFrom:HistOfNorthumbria.
GRAPH:settlementOfC3 {
   :Wall:creator:Hadrian;
   :Wall:inception “122”^^xsd:gYear.
   :settlementOfC3 conj:settles :C3.
}
GRAPH:C1 {
   :Wall:length “117km”.
   :Wall:height “2.4m”.
}
GRAPH:C2 {
   :Wall C2:creator:Severus.
   :Wall C2:inception “189”^^xsd:gYear.
   C2:creator conj:isAConjecturalFormOf:creator.
   C2:inception conj:isAConjecturalFormOf :inception.
}
:C2 prov:wasAttributedTo:Bede;
   prov:wasDerivedFrom:HistEcclesiastica.
```

Figure 7 Undisputed, disputed and settled attributions and datings of the wall expressed with Conjectures (weak form)

Since Conjectures in their weak form are compliant with RDF 1.1, they can be easily retrieved with plain and traditional SPARQL queries **[fig. 8]**.

Strong and weak forms aim to provide two (non-alternative) solutions to express critical discourse and theories, preventing

technological barriers in their adoption. The full semantics of Conjectures is separately documented (Rolfini 2021), along with a longer dissertation on the structure of Conjectures.

```
SELECT DISTINCT ?conj
WHERE {
 GRAPH ?conj {
   ?item ?conjpredicate ?author .
   ?conjPredicate conj:isAConjecturalFormOf :creator
 }
}
```

Figure 8 SPARQL query retrieving all claims (?conj) and items (?item) referring to discarded attributions

5 Conclusions

In conclusion, facts are not the only important aspect of scholarly knowledge that deserve attention. Capta, which include selections, opinions, controversies, and debates, are what make scholarly knowledge interesting and valuable. However, formally encapsulating these diverse and multifaceted aspects of knowledge in a structured dataset remains a formidable challenge.

It is essential for cultural heritage experts and professionals to call for adequate formalisms and tools to capture and preserve these capta. By increasing the quality and quantity of digital collections with non-objective facts, we can achieve an important objective for our future. All the reification methods presented in this work enable us to make statements about statements within RDF. This implies that in addition to the claimed content (e.g., :Wall C2:creator:Severus), supplementary triples can be included to provide context for the information about the claim itself (e.g., with Conjectures, :C2 prov:wasAttributedTo:Bede; prov:wasDerivedFrom:HistEcclesiastica). Consequently, several connections can be established through various claims, such as indicating agreement, or disagreement, specifying if a claim is based on another, referencing both primary and secondary sources, and declaring motivations and evidence on which the claim is based. Through such contextual annotation, the structure of critical discourse in the Humanities can be formalised and, therefore, retrieved.

This proposal provides a more flexible and nuanced way to represent capta proposing Conjectures in their weak and strong form, acknowledging the complexities and uncertainties inherent in these phenomena. Ultimately, it is through a comprehensive approach to scholarly knowledge that we can continue to expand our understanding of the world around us.

Bibliography

Barabucci, G.; Tomasi, F.; Vitali, F. (2021). "Supporting Complexity and Conjectures in Cultural Heritage Descriptions". *CEUR Workshop Proceedings*, 2810, 104-15. https://ntnuopen.ntnu.no/ntnu-xmlui/handle/11250/2736994

Blasch, E.P. et al. (2019). "Uncertainty Ontology for Veracity and Relevance". *2019 22th International Conference on Information Fusion (FUSION)*. Ottawa: Institute of Electrical and Electronics Engineers (IEEE), 1-8. https://doi.org/10.23919/FUSION43075.2019.9011402

Carroll, J.J. et al. (2005). "Named Graphs". *Journal of Web Semantics*, 3(4), 247-67. https://doi.org/10.1016/j.websem.2005.09.001

Checkland, P.; Holwell, S.E. (2005). "Data, Capta, Information and Knowledge". Hinton, M. (ed.), *Introducing Information Management: The Business Approach*. London; New York; Amsterdam: Elsevier, 47-55.

Cyganiak, R.; Wood, D.; Lanthaler, M. (2014). "RDF 1.1 Concepts and Abstract Syntax". *W3C Recommendation 25 February 2014*. https://www.w3.org/TR/rdf11-concepts/

Daquino, M. et al. (2022). "Expressing Without Asserting in the Arts". Di Nunzio, G.M. et al. (eds), *Proceedings of the 18th Italian Research Conference on Digital Libraries* (Padua, 24-25 February 2022). CEUR Workshop Proceedings 3160. http://ircdl2022.dei.unipd.it/downloads/papers/IRCDL_2022_paper_31.pdf.

Di Pasquale, A. et al. (2024). "On Assessing Weaker Logical Status Claims in Wikidata Cultural Heritage Records". *Semantic Web Journal*. https://www.semantic-web-journal.net/content/assessing-weaker-logical-status-claims-wikidata-cultural-heritage-records-1

Doerr, M.; Ore, Ch.-E.; Stead, S. (2007). "The CIDOC Conceptual Reference Model: A New Standard for Knowledge Sharing". *Proceedings of the Conference: Challenges in Conceptual Modelling. Tutorials, Posters, Panels and Industrial Contributions at the 26th International Conference on Conceptual Modeling – ER 2007* (Auckland, 5-9 November 2007). ACM: Association for Computing Machinery 83, 51-6. https://dl.acm.org/doi/pdf/10.5555/1386957.1386963.

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

Erxleben, F. et al. (2014). "Introducing Wikidata to the Linked Data Web". Mika, P. et al. (eds), *Proceedings of the Conference: The Semantic Web – ISWC 2014 = 13th International Semantic Web Conference* (Riva del Garda, 19-23 October 2014, part 1). Cham: Springer International Publishing, 50-65. Lecture Notes in Computer Science 8796. https://doi.org/10.1007/978-3-319-11964-9_4

Hartig, O. (2017). "Foundations of RDF* and SPARQL* (An Alternative Approach to Statement-Level Metadata in RDF)". Reutter, J.L.; Srivastava, D. (eds), *Proceedings of the 11th Alberto Mendelzon International Workshop on Foundations of Data Management and the Web 2017* (Montevideo, 7-9 June 2017), article ID 12. CEUR Workshop Proceedings 1912. https://ceur-ws.org/Vol-1912/paper12.pdf.

Hayes, P. (2004). "RDF Semantics". *W3C Recommendation 10 February 2004*. http://www.W3.Org/TR/Rdf-Mt/

Laskey, K.J. et al. (2008). "Uncertainty Reasoning for the World Wide Web". *W3C Incubator Group Report 31 March 2008*. http://www.w3.org/2005/Incubator/urw3/XGR-urw3/

Nguyen, V. et al. (2015). "On Reasoning with RDF Statements about Statements Using Singleton Property Triples". *arXiv*, 15 September. https://doi.org/10.48550/arXiv.1509.04513

Noy, N.; Rector, A. (2006). "Defining N-Ary Relations on the Semantic Web". *W3C Working Group Note 12 April 2006*.
https://www.w3.org/TR/swbp-n-aryRelations/.

Rolfini, A. (2021). "Semantics of Conjectures", *arXiv*, 17 October.
https://arxiv.org/abs/2110.08920

Models of Data Extraction and Architecture in Relational Databases of Early Modern Private Political Archives
edited by Dorit Raines

Metapolis: Spatializing Histories Through Archival Sources

Lukas Klic
I Tatti - The Harvard University Center for Italian Renaissance Studies, Italy

Abstract *Metapolis* endeavours to fill a gap within digital infrastructures tailored for the Humanities, particularly those designed to encompass the entire research continuum of geospatial historical research. While numerous digital tools exist to cater to the research requisites of scholars engaged in spatio-temporal inquiries at various levels, these tools often operate in isolation from one another, resulting in a fragmented research process. *Metapolis*, in contrast, amalgamates a suite of tools and provides essential support for humanities scholarship, thereby empowering researchers in their endeavour to reconstruct locations across temporal dimensions within a geospatial framework. Functioning as an interactive map-centric publication platform, it affords users the capacity to intermingle archival, bibliographic, and multimedia resources with interpretative research, thereby enabling their correlation and visual representation on a geographical map. Leveraging ResearchSpace, an open-source Semantic Web research environment, *Metapolis* facilitates the reuse and dissemination of Linked Open Data. Additionally, an extensive array of functionalities facilitates the enhancement of data through external knowledge repositories like VIAF, WikiData, Worldcat, and the Getty vocabularies. Conceived both as a research instrument and a publication medium, this software enables cohorts of scholars spanning a wide spectrum of humanistic disciplines to harmonize their analyses and bolster each other's discoveries through the overlaying of historical maps, interlinking them with sources to empower users in constructing a deeper understanding of the world and its historical evolution.

Keywords Metapolis. Linked Open Data. Research infrastructures. ResearchSpace. Geospatial data.

Summary 1 Introduction. – 2 Background. – 3 System Architecture. – 4 Spatializing Histories. –5 Data Sustainability. – 6 Conclusion.

Edizioni Ca'Foscari

Studi di archivistica, bibliografia, paleografia 8
e-ISSN 2610-9093 | ISSN 2610-9875
ISBN [ebook] 978-88-6969-919-1 | ISBN [print] 978-88-6969-920-7

Peer review | Open access
Submitted 2023-10-03 | Accepted 2024-01-18 | Published 2025-05-22

DOI 10.30687/978-88-6969-919-1//004

1 Introduction

The *Metapolis* project[1] is an initiative to develop a digital research infrastructure to support scholarship in the humanities that focuses on identifying and articulating the complex network of relationships that surround places throughout time. The architecture of the platform is designed both as a research and publication tool and allows groups of scholars from a wide range of humanistic disciplines to connect their research and interlink and augment each other's findings. While advancements brought about by the Open-Source Geospatial Foundation and other initiatives have been instrumental in advancing geospatial research, real-world applications and tooling for the humanities are still lacking. Furthermore, the various tools available to scholars that could support an end-to-end research workflow are fragmented across various pieces of software that do not communicate with one another (Waters 2023). In order to support novel and more innovative forms of inquiry into the history of urban spaces and culture, the field is in need of a map-based research infrastructure geared toward non-technical users, empowering them with the necessary set of digital tools to support innovative historical scholarship (Knowles 2008).

Metapolis builds on the advancements of the ResearchSpace initiative,[2] a collaborative research environment developed by the British Museum (Oldman, Tanase 2018; Oldman, Tanase, Santschi 2019). Various projects have demonstrated how the platform can be used as an interpretive research tool that powers the full life cycle of digital scholarly research and publication.[3] *Metapolis* can be used to map people, an idea, an art form, a philosophical movement, or any concept that can be georeferenced on a map. One could just as easily track the birth and proliferation of Jazz in New Orleans, the forcible relocation of the Cherokee Nation along the Trail of Tears or map the impact of COVID-19 on given communities. Historical census data, deeds, account books, wills, and inventories could be uploaded and georeferenced. Layers of digital media can then be interwoven to create a depth of information that is impossible to achieve with conventional tools. Most software that supports geospatial research

Metapolis is generously supported by a Level III Digital Humanities Advancement grant from the National Endowment for The Humanities. The project is a highly collaborative effort between Researchers in the Digital Humanities Lab at the Harvard Center for Italian Renaissance Studies: Ludovica Galeazzo, Remo Grillo, Lukas Klic, and Gianmarco Spinaci. This project would not be possible without this team of highly dedicated scholars.

1 https://dh.itatti.harvard.edu/#Metapolis.

2 http://researchspace.org/.

3 https://hokusai-great-picture-book-everything.researchspace.org/; https://florentinedrawings.itatti.harvard.edu/.

focuses on visualizing a limited set of data that is decoupled from these historical sources. *Metapolis* enables the publishing of these sources, together with scholarly narratives or articles alongside a map and offers innovative methods of interaction, search, and visualization. By georeferencing media such as photographs or oral histories, the historical record can be expanded to provide novel views of culture that could otherwise remain undiscovered.

ResearchSpace is an open-source collaborative Semantic Web environment designed to use and build knowledge about the world and its history. The platform is already rich with features: data extraction and integration, linking and enrichment, storage, querying, inferencing, search, visualization, and authoring, all using well-established open standards of the Semantic Web that enable reuse and repurposing in Linked Open Data environments. As an open-source system, the platform allows for enhancements from the community, which in turn can be contributed back to the source code, allowing for dissemination to a global community. The software architecture is data-centric, allowing user interfaces to be built and centered around any kind of research and collection data (Klic 2019, 31-2).

In order to accommodate complex historical reconstructions, a series of necessary customizations and upgrades to geospatial functionality have been implemented in *Metapolis*. Traditional Geographic Information Systems (GIS) such as ArcGIS[4] and QGIS[5] allow scholars to analyze and superimpose historical maps on modern Cartesian coordinates (Gregory, Geddes 2014). While these systems can be used to create visualizations of transforming urban spaces, they generally lack the ability to publish these findings in a web-based infrastructure that supports search and further interpretation by end-users (Huang, Harrie 2020). GIS systems typically employ functionality where map 'layers' serve to represent segments of time, a component that often places limitations on scholars' ability to express a more complex and dynamic urban landscape. Additionally, these systems limit scholars to representing an interpretive analysis that is the result of years of research without the ability to link back to original sources. As a result, archival documents, visual representations, and secondary literature that serve as the foundation of research become secondary and are often buried in footnotes. *Metapolis* seeks to overcome these limitations by extending a rich set of features in ResearchSpace, with functionality commonly found in GIS systems. It aims to create a platform where users can seamlessly navigate between historical sources and spatial data in ways that foster new opportunities for interpretation and research.

4 https://www.arcgis.com/index.html.

5 https://qgis.org/it/site/.

Institutions and scholars globally will have a research infrastructure capable of hosting their own geospatial projects. Interactive publications and narratives can be coupled with annotations and assertions on maps, historical documents, and iconographic sources. These semantically enriched and structured data would in turn contribute to a culture of open scholarship and collaboration among researchers, disrupting barriers posed by proprietary databases and publishing platforms where information is kept in silos. The platform would also serve as a pedagogical instrument that could be incorporated into the curriculum of courses worldwide, both in-person and in Massive Open Online Course (MOOC) environments. The nature of this machine-readable data lends itself well to playful and serendipitous discoveries, making it an attractive and engaging tool for both undergraduates and seasoned scholars alike. The project aims to advance the global paradigm shift in publishing models, away from an inward-looking, closed, and costly model towards an open and inclusive one based on collaboration and open access (Klic 2019, 24-5).

2 Background

The number of humanities projects that employ digital geospatial methodologies and tools is vast since geographic coordinates, together with dates, form part of a limited set of historical data that are computationally actionable at a larger scale. International consortiums such as the Open Source Geospatial Foundation (OSGeo)[6] have done exceptional work on this front, allowing for unprecedented interdisciplinary collaboration in the field. Groundbreaking open-source tools developed by this consortium, such as QGIS,[7] a desktop Geographical Information System, and OpenLayers,[8] a set of open-source libraries for visualizing and publishing geospatial data on the web have enabled numerous digital humanities projects to be brought to fruition. Other tools, like Neatline,[9] have been integrated into the open-source digital publishing software such as Omeka,[10] enabling for the visualization of humanities data to be published in web environments in productive ways. These have even resulted in digital publications such as *The Chinese Deathscape,*[11] a digital volume from Stanford University Press that explores grave relocation and burial reform in

6 https://www.osgeo.org/.
7 https://www.osgeo.org/projects/qgis/.
8 https://openlayers.org/.
9 http://www.neatline.org/about/.
10 https://omeka.org/.
11 https://chinesedeathscape.org/.

China. Projects such as *Digimap*[12] have allowed for the publication of stratified historical and contemporary maps to facilitate scholarly analysis. Research projects like *DECIMA*[13] (Terpstra, Rose 2016), *Visualizing Venice/Visualizing Cities*[14] (Huffman, Giordano, Bruzelius 2018) and the *Venice Time Machine*[15] (Abbott 2017) have sought to reconstruct the history of urban landscapes at scale.

All these research and infrastructure projects have used desktop GIS systems, either ArcGIS or QGIS, to perform the complex historical analysis of these spaces. By georeferencing historical maps over modern-day ones and cross-referencing historical documents, researchers were able to determine the history of political boundaries, the transformation of urban spaces, and the movement of objects, actors, or goods over space and time in novel ways that are often impossible in the analog form. Notwithstanding the gamut of digital tools at the disposal of scholars, the output of this often-painstaking research has generally been a print publication that does not provide access to the source data. Publications in the digital form have not allowed users to search for historical data and sources within a particular space and time or visualize the results on a map in meaningful ways that can facilitate further interpretation (Gaehtgens 2013). Additionally, these publications often employ flat data structures that do not do justice to the rich scholarship behind the complex network of relationships that compose the publication.

Historical reconstructions require the ability to visualize and search for historical features, generally in the form of polygons and lines, that can take the form of buildings, outdoor spaces or political boundaries. Although Omeka does offer tools for data input and visualization on a map, it is not able to author or visualize anything other than individual points or allow for geospatial search. Tools such as GeoNetwork[16] can manage complex geospatial features but are built on data management tools that do not adapt well to humanities research and are often used for managing sets of geospatial data rather than publishing them in an intuitive user interface. GeoBlacklight[17] allows for the publishing of geospatial datasets and has limited visualization options but is more geared towards providing raw data that the user can download and reuse in other contexts, rather than

12 https://digimap.edina.ac.uk/.

13 https://decima-map.net/.

14 https://www.beniculturali.unipd.it/www/ricerca/linee-di-ricerca/visualizing-venice-exploring-the-citys-past/.

15 https://www.epfl.ch/research/domains/venice-time-machine/.

16 https://geonetwork-opensource.org/.

17 https://geoblacklight.org/.

serving as a tool for scholars to author and interact with data. Carto[18] offers powerful visualization tools that can be pre-configured by the author to visualize a single idea but does not allow for subsequent reconfiguration and interpretation by end-users, nor has any data management functionality. The *Metapolis* project has evolved out of years of experimentation with these existing tools and hopes to fill these gaps by extending existing work in the field and building on top of it rather than in parallel.

3 System Architecture

The ability to represent artistic production, historical documentation and urban change over time is instrumental for research projects that tell the complex histories of the urban, architectural, cultural, and socio-economic dynamics of early modern cities, as well as of the actors and objects pertaining to these spaces. The *Metapolis* project is rooted in an iterative expansion of the ResearchSpace infrastructure in order to be able to represent and publish these histories.

After several years of successfully working within the ResearchSpace ecosystem, researchers in the Digital Humanities Lab at I Tatti, The Harvard University Center for Italian Renaissance Studies, began planning for a more ambitious set of tools to incorporate geospatial mapping functionality. This work has largely been informed by a collaboration with the ERC project *Venice's Nissology – VeNiss*, led by PI Ludovica Galeazzo (2022),[19] and her previous project *Archipelago* (2021).[20] This research aims to digitally reconstruct the history of a cluster of sixty islands scattered across the Venetian lagoon from the late fifteenth century onwards. Both these projects originally employed a relational database, FileMaker Pro, to collect and store all historical data and sources, alongside a geographic information system, ArcGIS or QGIS, for mapping and geocoding urban data. While geodatabases store spatial data, including digital reconstructions of the ancient landscapes represented through points, lines, and polygons and provided mapping functionality, they are never directly connected to relational databases through any integration. The research questions behind the *VeNiss* project made working with this software configuration exceedingly

18 https://carto.com/.

19 The project *Venice's Nissology. Reframing the Lagoon City as an Archipelago: A Model for Spatial and Temporal Urban Analysis (16th-21st centuries)* has been funded with a five-year grant by the European Research Council (ERC-2021-StG, n. 101040474). https://www.unipd.it/en/sites/en.unipd.it/files/Galeazzo_VeNISS.pdf.

20 https://dh.itatti.harvard.edu/#Archipelago.

cumbersome and insufficient. One could not, for example, visualize textual or iconographic sources describing historical urban places with digital representations of their ancient configuration, or connect certain geographical modifications to their related urban events and agents.

The software architecture of *VeNiss* echoes the requirements of many other projects in the field: a necessity to integrate geospatial data with historical sources in the form of structured data, and to provide a research and publishing platform that seamlessly integrates these two in a web-based environment made available to non-technical users. Following this extensive analysis of existing software, researchers at the Digital Humanities Lab together with numerous other stakeholders determined that it would be impractical to attempt to recreate a single web-based tool that could match the robust functionality of desktop software such as ArcGIS or QGIS. They decided that the most effective path forward would be to extend the rich data management and publishing features of ResearchSpace with functionality to support geospatial data, search, annotation, and analysis. A seamless software integration with QGIS would then allow scholars to perform the complex work of geocoding and drawing in the more powerful desktop software. Beginning with the research questions, the team then started to map out the architecture of the platform. Before working on any software development, the group spent three months designing extensive mock-ups that laid out the full user interaction and design of the platform, in order to ensure a user experience that was intuitive and productive for non-technical users **[figs 1-2]**.

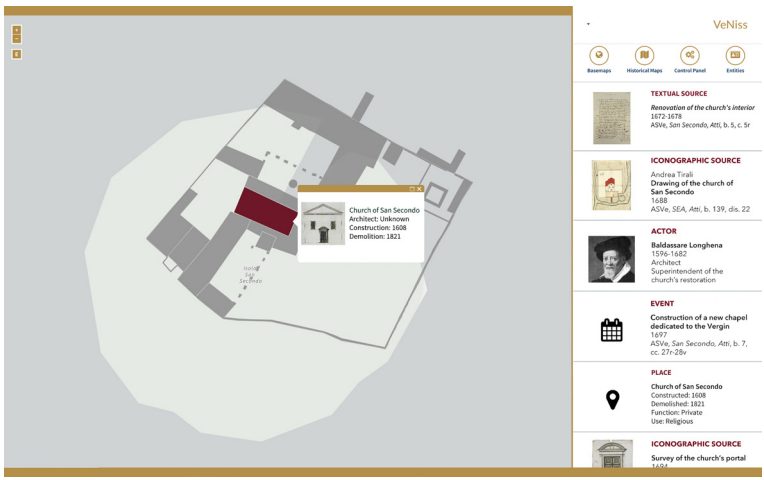


Figure 1 Feature popover opens when clicked

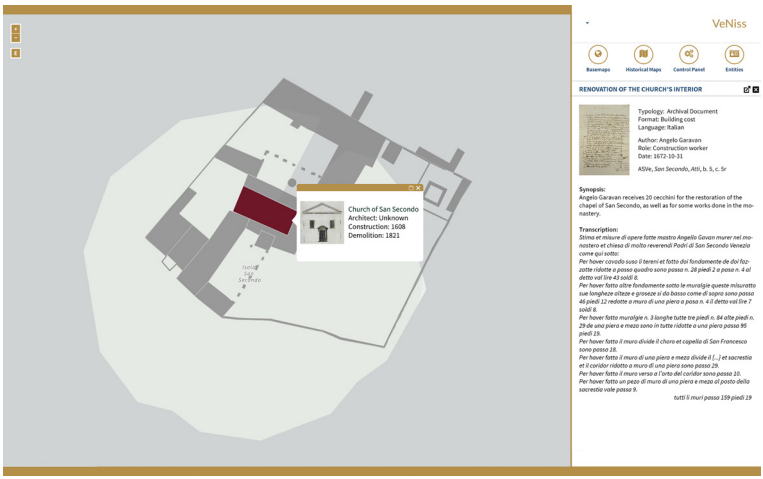


Figure 2 Navigator overlay with specific information of the selected entity

In order to ensure the overall technical feasibility of the project architecture, the team also began developing one critical software component that would allow for the communication between the relational geospatial database back end of QGIS, with the graph database back end of ResearchSpace (Blazegraph).[21] Following months of software development, this critical first step was successfully completed in November of 2021, forming the backbone for integrating GIS-based systems with a Linked Open Data research platform capable of connecting, enriching, and publishing archival data in a geospatial research environment.

4 Spatializing Histories

An extensible architecture for the geospatial mapping components is critical to support a range of research questions into the history of a given space. These components allow for interaction between the map and existing data management and publishing features of ResearchSpace.

Functionality to be able to change the basemap tiles and mapping providers allows scholars to choose a map background that best serves the needs of their individual research. This will permit them to use a wide range of map tile providers: Mapbox, MapTiler, ArcGis, or Open Street Maps. Georeferenced images can also be uploaded, superimposed, and used as a reference to draw geospatial features

21 https://blazegraph.com/.

on the maps in the form of points and polygons, as well as annotation functionality that enables researchers to connect these features to data in the knowledge graph. Visualization tools include the ability to add any number of historical maps as separate layers **[fig. 3]**.

Figure 3 Georeferenced historical maps

Functionality to set transparency levels on these layers that allow researchers to compare and contrast them, together with 'swipe'[22] and 'eyeglass'[23] functionality, allows these layers to be compared and contrasted at both micro and macro levels. The styling (through colors and text labels) of geospatial features based on type, allows users to visualize their historical functions and use. They also allow for the visualization of the movement of objects or goods across space-time, enabling researchers to map their displacement. Most importantly, a timeline slider allows users to limit their visualization of geospatial features and archival sources to a particular point in time. Non-technical users are able to upload images of archival documents to the platform and publish them through an International Image Interoperability Framework (IIIF)[24] server embedded in ResearchSpace. Web forms allow researchers to transcribe data from these sources, publishing them as Linked Open Data. These data can be properly attributed to multiple researchers and can be coupled with textual narratives that provide additional

22 https://openlayers.org/en/latest/examples/layer-swipe.html.

23 https://openlayers.org/en/latest/examples/layer-spy.html.

24 https://iiif.io/.

levels of interpretation. A seamless integration with external knowledge bases allows users to query the title, author, ISBN, or DOI of a particular book or article and seamlessly cite secondary literature.

Figure 4 Dynamic feature styling and labeling

Robust search and faceting functionality permit any user to fine-tune the visualization of archival sources and geospatial features based on their own research interests **[fig. 4]**. The ability to annotate existing entities and text within a web-based editor, allows for the cross-pollination of geospatial data, textual narratives, and data from archival sources to enable a cohesive visualization of these histories.

5 Data Sustainability

Libraries and archives have come to accept the fact that while the underlying software that supports a digital project is inherently ephemeral, the data and research that go into these projects need to be portable from one system to another. SQL structures where the metadata model is built into the database or software are difficult to interpret without the underlying software. By employing graph structures, data is stored in the Resource Description Framework (RDF) model,[25] ensuring greater portability to other systems in the future. Because the ontology that describes both the meaning and context of the data is built into the model itself, it can easily be ported to any other graph data

25 https://www.w3.org/TR/rdf-schema/.

structure that supports RDF, and a user interface can be developed on top. Well-documented ontologies like the Conceptual Reference Model (CIDOC CRM)[26] allow for an extremely detailed description of data, its meaning, and context. This approach is extremely attractive to libraries and archives as the raw data could be retrieved hundreds of years into the future and one would be able to understand and decipher its meaning without any kind of specialized software.

6 Conclusion

The *Metapolis* project, through its integration of advanced geospatial capabilities with the open-source ResearchSpace infrastructure, marks a significant leap forward in the field of digital humanities. Our work has addressed existing gaps in the availability of comprehensive, user-friendly digital tools that enable in-depth research and interactive publishing in humanities disciplines. *Metapolis* is designed to move beyond the limitations of traditional Geographic Information Systems, empowering scholars to seamlessly explore the interplay between historical data and spatial phenomena. It brings the capacity for novel historical inquiry to the forefront by incorporating archival documents, visual representations, and secondary literature directly into its spatial framework, rather than relegating these crucial sources to footnotes or separate databases.

Building on the established functionalities of ResearchSpace, *Metapolis* delivers a user-centric interface that enables not only the visualization but also the dynamic interpretation of spatial-temporal data. By allowing users to integrate georeferenced media and link these to the underlying sources, *Metapolis* expands the historical record and provides fresh perspectives on cultural patterns and developments that might otherwise remain hidden.

The project source code is open-source and freely available on GitHub, allowing for contributions from a global community.

Finally, *Metapolis* aims to contribute to the global shift in publishing models by promoting a collaborative and open-access approach. This departure from traditional proprietary databases and publishing platforms allows for a more transparent and integrated scholarship, paving new paths for scholarship working with historical reconstructions.

The project encapsulates an ambitious vision for the future of digital humanities research and education. By providing a familiar Google Maps-like interface to users with a time slider and direct access to archival sources, *Metapolis* allows new paths of scholarly

26 https://www.cidoc-crm.org/.

inquiry to be tested and defined. As we continue to refine and expand this platform, we anticipate that it will stimulate innovative scholarship, foster global collaboration, and reshape our understanding of the world's cultural and historical landscapes.

Bibliography

Abbott, A. (2017). "The 'Time Machine' Reconstructing Ancient Venice's Social Networks". *Nature News*, 546(7658), 341.
https://doi.org/10.1038/546341a

Gaehtgens, Th.W. (2013). "Thoughts on the Digital Future of the Humanities and Art History". *Visual Resources*, 29(1-2), 22-5.
https://doi.org/10.1080/01973762.2013.761110

Galeazzo, L. (2021). "Autorità ecclesiastica e civile nell'iconografia dell'arcipelago veneziano tra XVI e XVII secolo". *In_Bo*: *Dominio del Sacro. Immagine, cartografia, conoscenza della città dopo il Concilio di Trento*, 12(16), 186-97.

Galeazzo, L. (2022). "Analysing Urban Dynamics in Historic Settlements Using a Geo-spatial Infrastructure. The Venice's Nissology Project". *Journal of Art Historiography*, 27, 1-13.

Gregory, I.N.; Geddes, A. (eds) (2014). *Toward Spatial Humanities: Historical GIS and Spatial History*. Bloomington: Indiana University Press.

Huang, W.; Harrie, L. (2020). "Towards Knowledge-Based Geovisualisation Using Semantic Web Technologies: A Knowledge Representation". *International Journal of Digital Earth*, 13(9), 976-97.
https://doi.org/10.1080/17538947.2019.1604835

Huffman, K.L.; Giordano, A.; Bruzelius, C. (eds) (2018). *Visualizing Venice: Mapping and Modeling Time and Change in a City*. London: Routledge. Routledge Research in Digital Humanities.

Klic, L. (2019). *Digital Publishing and Research Infrastructure for Cultural Heritage: an Institutional Roadmap* [PhD dissertation]. Venice: Ca' Foscari University of Venice.
http://hdl.handle.net/10579/15587

Knowles, A.K. (ed.) (2008). *Placing History: How Maps, Spatial Data, and GIS Are Changing Historical Scholarship*. Redlands (CA): ESRI Press.
https://doi.org/10.14714/cp63.160

Oldman, D.; Tanase, D. (2018). "Reshaping the Knowledge Graph by Connecting Researchers, Data and Practices in ResearchSpace". Vrandečić, D. et al. (eds), *The Semantic Web – ISWC 2018*. Cham: Springer, 325-40. Lecture Notes in Computer Science.
https://doi.org/10.1007/978-3-030-00668-6_20

Oldman, D.; Tanase, D.; Santschi, S. (2019). "The Problem of Distance. A ResearchSpace Case Study on Sequencing Hokusai Print Impressions to Form a Human Curated Network of Knowledge". *International Journal for Digital Art History*, 4, 5.29-5.45.

Terpstra, N.; Rose, C. (eds) (2016). *Mapping Space, Sense, and Movement in Florence GIS and the Early Modern City*. London: Routledge.

Waters, D.J. (2023). "The Emerging Digital Infrastructure for Research in the Humanities". *International Journal on Digital Libraries*, 24(2), 87-102.
https://doi.org/10.1007/s00799-022-00332-3

Experiences: Historical Archives, Database and Online Publication

Models of Data Extraction and Architecture in Relational Databases of Early Modern Private Political Archives
edited by Dorit Raines

Including the Archival Context in the Historian's Materials: The Advantages of Archival Standard Databases in Historical Research

VINCULUM Project Database and Information System Guide

Maria de Lurdes Rosa
NOVA.FCSH; IEM – NOVA. FCSH; PI of project VINCULUM, ERC Grant Agreement 819734

Abstract This paper examines the challenges and solutions in designing historical databases, focusing on maintaining archival context and relational integrity. Databases often prioritize specific inquiries, neglecting the comprehensive relationships and temporal transmission crucial for accurate historical analysis. The VINCULUM Project's database demonstrates the benefits of integrating international standards in order to preserve the integrity of information and enhance its reliability through design rationale, practical implementation, and the importance of an accompanying 'Information System Guide' to aid researchers. The conclusion underscores the necessity of providing comprehensive guidelines to ensure effective information-seeking and knowledge organization.

Keywords VINCULUM Project. Information System Guide. Entail. Historical-archival databases. Archival footprint.

Summary 1 Introduction. – 2 VINCULUM Project: A Presentation. – 3 Gathering Archival Records as a Scientific Problem. – 4 Building the Database. – 4.1 Characterization of the Data. – 4.2 Information Structuring. – 4.3 Data Architecture. – 4.4 Integration of the Documentary Information in the Global Information System. – 5 Constructing the (VINCULUM) Information System Guide. – 6 The Construction of a Database. A Challenge for Transparency and User Education.

Edizioni Ca'Foscari

Studi di archivistica, bibliografia, paleografia 8
e-ISSN 2610-9093 | ISSN 2610-9875
ISBN [ebook] 978-88-6969-919-1 | ISBN [print] 978-88-6969-920-7

Peer review | Open access
Submitted 2024-05-23 | Accepted 2024-09-02 | Published 2025-05-22

DOI 10.30687/978-88-6969-919-1//005

1 Introduction

Databases developed in response to historical inquiries often exhibit a tendency to diminish the archival context of the data, limiting it to a mere identification of provenance primarily conveyed through reference or access numbers. Furthermore, such databases tend to organize archival information based on specific inquiries rather than the comprehensive relational foundation of the empirical basis of the inquiry. Consequently, there is a potential risk of neglecting both the horizontal dimension of data, which encompasses the intricate web of relationships between data points, and the vertical dimension, which represents their temporal transmission over time. The significance of these dimensions in the context of historical analysis, particularly in ensuring the accurate sequencing of events to construct historical narratives, tends to remain obscured, if not entirely disregarded.

The judicious utilization of archival description databases, which adheres to the integrity of archival fonds and to international archival standards and the principles of diplomatics concerning document transmission, offers a viable solution to address the issues encountered in historical databases. By upholding the comprehensive nature of information and adopting 'neutral' structures rooted in information and document production, archival description databases can attain heightened permanence and reliability. Firstly, they achieve this by carefully analysing the available historical information. Secondly, their capacity to elucidate the transmission of information across centuries and the potential alterations in the transmission process further contributes to their durability.

Additionally, these databases can furnish researchers with access to materials through diverse categories of indexes, encompassing anthroponomical, toponymical, and subject-based classifications, among others, employing controlled vocabularies. These indexes may also undergo enrichment by various users, based on specific research inquiries.

The relational aspect of archival databases is strengthened through the synergistic utilization of various standards, such as ISDIAH, ISAAR-CPF, and ISAD-G.[1] An illustrative example of the

1 It is important to note that the VINCULUM project was conceived in 2018, during a period when several developments in the field of archival description were underway but not yet completed or publicly available. We recognize the significance of models such as Records in Contexts (RiC), as well as the various knowledge organization tools and theories developed over the past decade in the field of Archival Science. Indeed, we have made concerted efforts to promote their adoption by historians, who, on the whole, remain largely unfamiliar with them, often unaware even of the International Standard Archival Description (General) [ISAD(G)] and other related standards employed in the VINCULUM database (Rosa 2024). It is important to emphasize, however, that this text is written from the perspective of a historian who relies on archival

potential benefits derived from the integration of multiple standards lies in the introduction of authority records for creators that can be interconnected in multiple ways. Additionally, it is imperative to accompany these databases with an 'information system guide' that provides a comprehensive account of the institutional history of the creators, the methodologies employed in the production and documentation of information, as well as the historical evolution of document transmission and the constitution of records within historical archives, including custodial and archival histories.

This text endeavours to delve into these themes with a primary focus on the experiences and developments within the VINCULUM Project's database, which has been in progress since 2019. The discussion will commence with a succinct introduction to the project, followed by an elucidation of the underlying rationale behind the database's design. Subsequently, the second part of the text will concentrate on practical aspects and the presentation of outcomes, encompassing the database's construction and an in-depth exploration of the characteristics of the associated 'Information System Guide'. The discourse will culminate in a contemplation of the challenges associated with the selection of an archival database as an option, underscoring the critical need to equip researchers with comprehensive guidelines that draw upon insights from studies on information-seeking behaviour and knowledge organization.

2 VINCULUM Project: A Presentation

Few social phenomena have been as pertinent to premodern societies in southern Europe as entails (*morgadios*, *mayorazgos*) and chantries. These institutions, the subjects of analysis within the VINCULUM project, are collectively examined under the term 'entailment'. They evolved as a means of preserving property within particular family structures by establishing a legal entity managed by carefully chosen successors in perpetuity. This entity constituted a corporate body, overseen by the designated successors, operating within the

description, believing in its fundamental importance for historical research, yet must contend with the specific challenges posed by data from the fourteenth to the seventeenth centuries. This involves dealing with handwritten documents (which require strong palaeographical skills) produced by non-contemporary institutions, which have entered public archives through long and disorganized processes of incorporation, and which have often been poorly and anachronistically described by archivists. The VINCULUM project reviewed thousands of documents, characterized dozens of information-producing institutions, and identified and defined approximately 150 documentary typologies. The application of the latest and most sophisticated models would have been extremely challenging while the project was already in progress; however, we hope that these models may serve as the foundation for future projects.

framework of enduring continuity. It encompassed the living, the deceased, and future family members, holding significant authority. Rooted in the deeply ingrained cultural figure of the *founder* – whose will was regarded as law in its most comprehensive sense – the corporate body governed interpersonal relationships within and outside the family sphere. It established distinct connections with property and the economy, negotiated tradition, and controlled change. To a large extent, the corporate body served as the social agent dictating the actions of its human members, their circumstances, and strategies.[2]

Between the fourteenth and seventeenth centuries, approximately 7,000 entails and chantries were established across what is now continental Portugal, as well as in the Atlantic regions colonized during this period of institutional expansion and consolidation. These institutions played a vital role in enabling the nobility to maintain their social standing and facilitating social mobility among other social groups. In the island communities of Madeira and the Azores, entailment rapidly organized society and property within a few decades, imbuing them with distinct characteristics. Meanwhile, in Cape Verde and the territories occupied along the Brazilian coast, entailment was primarily introduced through chapels, gaining significance at a more gradual but consistent pace. Existing studies suggest that this trend mirrored developments in other Iberian kingdoms, particularly Castile. Resembling the peninsular *morgadio* in many aspects, during the modern era, entailment evolved into a defining characteristic of numerous regions in southern Europe and served as a fundamental model for elite reproduction. Consequently, it exerted a profound influence on various facets of society.

Although entailment gradually waned and eventually vanished in later centuries – in the Iberian territories, it became extinct in the second half of the nineteenth century – its concluding phases were not without turmoil. The discussion of its legal abolition spanned decades. On the eve of its demise, numerous entail owners officially chose to preserve it, despite the associated high administrative costs. Even after its formal abolition, it persisted on a significant scale through private agreements until at least the mid-twentieth century. An influential outgrowth of the socio-political relationships shaped by entailment was the figure known as the *cacique*, a pivotal presence in post-liberal, rural societies. Remarkably, entailment was formally resurrected, in various configurations, by the fascist regimes that gained prominence in twentieth-century southern Europe. In reality, the framework of entailment transcended mere social solutions; it constituted a cultural phenomenon, deeply ingrained in the societies

2 See a recent state-of-the-art in Rosa 2020.

under examination. Key elements such as the paternal figure's prominence, the allure of paternalism as a political solution, the role of the extended family (always present though in the background), and the influence of familial obligations (both material and spiritual), are recurrent themes in sociological portrayals. Furthermore, the subject resurfaces in the literature and other cultural expressions of southern Europe, particularly in Portugal and Brazil. An illustrative example is the *morgadio*, which remains a central figure in nineteenth-century romantic medievalism and its extensions.

Historiography has not overlooked this profoundly significant topic. Prominent historians, encompassing both medievalists and modernists, have long underscored the impact of entailment on the formation of Iberian societies. In recent decades, scholars specialising in medieval and modern periods alike have reexamined the timing and forms of lineage transformations in these societies, emphasizing the pivotal role of the *morgadio* in consolidating such processes. This feature appears to be characteristic of Iberian societies, both in terms of the institutional element's significance and the early adoption of primogeniture legal systems (Sottomayor-Pizarro 2011; Monteiro 2001). While these historiographical perspectives lay the groundwork for scholarly contemplation, they also underscore the absence of a comprehensive study of this phenomenon.

The project focuses on a vast and intricate subject, one in which multiple social forces converged within a legal framework. This framework conferred a unified identity as a corporate legal entity, shaped around the figure of the founder and governed by internal laws. However, the study of entailment thus far has been approached from one-sided perspectives, whether they be legal, social, or cultural, which tend to overlook its holistic nature. Alternatively, it has been explored through site-specific monographs that seldom engage with broader issues. The project is grounded in the premise that a comprehensive understanding of this phenomenon can only be achieved by considering it in its entirety, especially because the autonomy of analytical domains such as politics, economy, society, and religion corresponds to an epistemological paradigm of a different era (Guerreau 1980; 1990).

Starting from the Portuguese-Iberian context, 'entailment' (*vinculação*) is examined as a multifaceted, yet pivotal practice deeply embedded in law, aristocratic discourse, and kinship-based organization. The aim is to formulate a definition that accommodates its global nature. The research program seeks to comprehend and interpret this historical phenomenon in a global and innovative manner, proposing a new paradigm for analysing entailment with the goal of arriving at a novel conceptualization that allows for a comprehensive interpretation of this social practice.

Although the term 'entailment' already possesses a suitable definition in English, referring to "the restriction of property by limiting the inheritance to the owner's lineal descendants or to a particular class thereof", its historiographical usage requires further elaboration and comparative analysis. The VINCULUM project is grounded in the analytical significance of the idea that entailment, as a corporate body, played a pivotal role as a social agent. The project endeavours to elucidate how this agency was established, how it operated, and why it endured for numerous centuries. By assuming that this distinct legal entity acted as a social agent within a society where non-personal legal entities were commonplace, it becomes possible to illuminate and scrutinize the historical problem, comprehending entailment in its entirety, including its various functions related to people, kinship, suffrages, community integration, economic management, and archival production (Rosa 2012).

Consequently, the core structure of the research program has been crafted around what are identified as the central elements of the agency of the entailment corporate body, namely kinship, power, and identity. The ultimate objective is to propose a fresh conceptual framework for this phenomenon.

3 Gathering Archival Records as a Scientific Problem

The decision to propose a comprehensive database as the central outcome of this project was not made lightly. We held a firm conviction that databases have the capacity to shape realities, which can be potentially hazardous if not thoroughly and transparently elucidated for users. Furthermore, our awareness of numerous databases that had relatively short lifespans, often designed for highly specific historical inquiries, gave us pause. However, in this particular instance, it was not merely a matter of choice but rather a necessity driven by the complexity of the issues at hand, a viewpoint we shall briefly elaborate upon.

In the realm of Iberian historiography, the first decade of the twentieth century witnessed a growing interest in this subject, emerging within both historiographical traditions. This burgeoning interest was fuelled by distinct institutional circumstances, notably the significant activity of the *Seminario Familia y Elite de Poder* at the University of Murcia, along with the effects of modernization and increased openness within scientific environments in Iberian universities. These factors facilitated the modernization of historical approaches, marking a significant qualitative advancement from the predominantly erudite and monographic studies of the past. For instance, scholars began to delve into more specific facets of *mayorazgos* (primogeniture) – such as their economic constraints – which

were perceived as being rather adaptable (Quintanilla Raso 2004). There was also a heightened focus on the religious dimension of entailment, particularly evident in funerary chapels (Roth 2007; Irigoyen 2004). Nevertheless, even amidst these positive developments, figures like Soria Mesa, a prominent authority in the history of Iberian nobility, emphasized the pressing need for holistic perspectives on entailment.

In fact, following his initial attempts at extensive data collection (Soria Mesa 2007), the scholar soon concluded, through a comprehensive historiographical review of research on the Spanish Modern period nobility, that the

> *mayorazgo* was conspicuous by its absence from the historiographical scene, despite the fact that it had been, paradoxically, the main protagonist of the evolution of Spanish nobility, at least from the fourteenth to the nineteenth century. (Soria Mesa 2009, 225)

This assessment resonated closely with my own research experiences within the Portuguese context. Consequently, the imperative for 'large data gathering' was recognized as a fundamental requirement for the VINCULUM project from its inception.

The reasons behind the rapid proliferation of entail foundations in the Iberian Peninsula, particularly in the realms of Portugal and Castile, remain largely unknown, partly owing to the inherent challenges associated with their documentation. This predicament is multi-faceted. Firstly, the registration of entails within Crown administrative offices was a development that occurred relatively late. Secondly, invaluable repositories for research, such as notary offices and family/house archives, have suffered significant losses, with the latter yet to be comprehensively catalogued or exploited. Thirdly, archives directly linked to entails experienced substantial losses following the dissolution of the legal institution, accompanied by the subsequent disbandment of corporate families.

Within the ARQFAM program[3] and the INVENT.ARQ project,[4] we have identified approximately thirty private archives that have become accessible to our project, either through the documents themselves or via antiquated inventories containing dated and substantial summaries (Rosa, Head 2015). Some family archives are also housed in state archives, and recent efforts have been made to access this material. However, contemporary archival theories regarding family archives have had limited influence, necessitating

3 https://arqfam.fcsh.unl.pt/.

4 https://inventarq.fcsh.unl.pt/.

a rigorous analysis of the organisational structure before utilizing the fonds. Furthermore, the voluminous collections maintained by Crown/State and Church institutions responsible for overseeing entails and chantries are predominantly either undocumented or reliant on archival treatments from the nineteenth century. Accessing these fonds demands an intimate understanding of the intricacies of the entail system.

It is imperative to underscore that the significance of this undertaking extends far beyond the mere compilation of concrete information pertaining to entails. It encompasses the reconstruction of the lost archives of entails, or, from a broader perspective, the production of information related to entails. Our decision was to reconstruct the informational and documentary landscape surrounding entailment, whether it be through direct sources (family archives) or indirect ones (Crown/Church institutions connected to entailment). This reconstruction spans from the generation of information upon its initial entry into archival institutions, including the organization of records at that stage (and their enduring presence within family archives, often entailing emotional and social distinctions) (Rosa 2017). This approach holds particular relevance in the study of entails since the institution itself *created archives*. Remarkably, a vast majority of the central documentation found in family archives traces its origins back to entails – a fact that has been largely overlooked by historians and yet bears immense significance. For this project, the archives of entails, replete with their social implications, are considered objects of study in their own right, not merely sources of information for historical investigations.

Hence, among the novel theoretical pathways advanced in this project, the database assumes a central role. It is not merely a means to an end, although historical research employing the database is both planned and underway.[5] Rather, the database constitutes an end in itself. The primary objective is to construct a theory-driven collection and organization of empirical materials, adhering to specific methodological approaches, which we shall elucidate in the following paragraphs.

The central premise revolves around the notion that reconfiguring entailment heuristics necessitates the reconstruction of the entailment's information system, as opposed to *relying solely on historical archives for data collection*. In advocating for the convergence of novel epistemological sources that encompass the scientific scrutiny of historians' materials, proposals drawn from archival science, and the emerging field of the history of archives and information, this project pledged a substantial commitment to reconstructing the

5 https://www.vinculum.fcsh.unl.pt/about/.

contemporaneous landscape of information production, documentation, utilization, preservation, and transmission (Rosa 2017).

This approach encompasses the consideration of existing documents in conjunction with those that have been lost, with the objective of forming a comprehensive overview. Each piece of information is meticulously linked to its respective producer, thereby imbuing it with its full contextual significance. The coexisting centres of information production, namely the entailments themselves, along with entities related to them such as Crown offices, church institutions, and municipalities, are conceptualized as direct or interconnected information producers. The entailment is scrutinized as the generator of official and legally acknowledged information, given that record creation and management are inextricably bound by the foundation charter (Rosa 2019; 2022; Iranzo Muñio 2009). A comprehensive exploration of *jus archivi* underscores the operation, as the inherent legality of entailment archives had previously been overlooked (Rosa 2022).

Beneath these tasks lies the fundamental notion that it is imperative to trace the path from *where things are kept* to *where things have come from, why and how*, elucidating the reasons and mechanisms involved. To achieve this, contemporary theoretical concepts concerning the reconsideration of sources will be brought into play. In accordance with Kuchenbuch's formulation (2004), every archival record, prior to being categorized as such, existed as a document of its own era. In the context of archival science, the informational act precedes the creation of the record itself. Finally, adhering to a conception of organizing and utilizing documentary materials that respects the structure of information and documentary production (Cammarosano 1991; Lodolini 1991), it becomes indispensable to have a comprehensive understanding of the institution responsible for generating the record and its administrative operational framework.

Ultimately, the overarching objective is to *engage with organisational archives*, recognizing their existence prior to the establishment of each currently available historical archive. Consequently, this project adheres to a methodology centered on the reconstruction of the institutional system engendered by entailment, both internally and externally. This encompasses the design of an institutional network, exploration of various forms of information production, the process of documentalization, and the constitution of archives.

Given that we were delving into the realm of premodern information production, the project also places significant emphasis on the novel perspective offered by the *anthropology of archives* (El-Leithy 2011). Comprehending the entailment information system necessitates a comprehensive view encompassing diverse methods of information gathering, utilization, and preservation. Integral components of entailments included armour, tombs, genealogical records, and prestigious symbolic objects (both tangible and intangible, such as evidence of

participation in prestigious battles). These elements were considered a core part of the entails and intrinsic to the foundation document, commemorated on specific occasions or during generational ceremonies, and prominently displayed in the social spaces of the properties (Rosa 1995; Martínez Perera 2010; Contreras Jiménez 2016). Ritual, liturgical, and charitable acts were obligatory for successive administrators and constituted an integral part of the discourse supporting their authority (Clanchy 1980-81; Cook 2013; Rosa 2021).

Furthermore, this perspective necessitates the recognition of archives of a distinct nature, akin to contemporary community archives (Rosa 2020). To fathom this milieu, contemporary perspectives from the field of archival science assume significant importance. The establishment of national archives has fragmented the premodern archival landscape and evolving societal contexts have relegated family/entail archives to the realm of private collections of amateur historians or symbols of social distinction. The few family archives housed in public archives are rarely organized or are classified as a form of administrative output of modern institutions. Additionally, the comprehensiveness of the information system encompassing documents, physical spaces, objects, and traditions was dismantled during the establishment of the modern taxonomy of memory and heritage institutions, demarcated by the triad of archives-museums-libraries (Stauffer 2021, chs 10 and ff.). Consequently, a comparative analysis with currently extant family information systems holds relevance, along with other observations concerning the continuity of entailment in modern societies.

Translating these historiographical suggestions into the conception of the database necessitated a strategic plan. The guiding principle was that the dismantling of Ancien Régime society, particularly the restructuring of its records during the establishment of the National Archives, obliterated prior information production and the archival landscape fundamentally. Therefore, to facilitate effective record retrieval and organization, the initial step involved reconstructing the informational landscape through:

1. an institutional examination of the entailment information system, encompassing archives related to entails and family archives, Ancien Régime control institutions (Crown and Church), and state institutions responsible for the abolition of entails;
2. an exploration of the custodial history of the archives of these institutions, considering their inclusion in either public or private archives, along with the types of descriptions applied;
3. the tangible reconstruction of each institution's information production, documentation, and preservation, utilizing surviving documents and ancient inventories that were subject to previous documentary criticism.

4 Building the Database

4.1 Characterization of the Data

The database aims to furnish comprehensive and accurately structured information on all existing entails (referred to as *vínculos*, encompassing both *morgadios* and *capelas*) within Portuguese territories (excluding the 'Estado da India') from the fourteenth to the seventeenth century, featuring lay familial administration. Preliminary estimates suggest that the database will encompass approximately 7,000 entail institutions. The scope of information and available documents for each institution will vary widely, ranging from mere mentions to including dedicated archives, some of which may contain hundreds of records. All document types produced or received by the entail will be included, with emphasis on particularly significant and voluminous ones, such as foundation deeds, legal judgments, inventories, and narrative documents. The database aims to incorporate descriptions of existing documents, as well as those that have been lost, utilizing information from ancient inventories when deemed suitable for inclusion.

4.2 Information Structuring

Beneath the development of the database lies a central tenet of the project, namely the reconstruction of the information structure under the concept of an 'entailment information system'. The project posits that information and documents related to entails are sporadically present in archives, with archival fonds often being inaccurately arranged and described. This predicament poses a fundamental question requiring resolution prior to any historical research. The inquiry itself constitutes a form of research – how to deal with a historically fragmented archival landscape, exacerbated by problematic archival and historiographical practices.

This misleading heuristic landscape can be attributed, primarily, to the oversight of the consequences of a protracted process of archival transformation, fragmentation, and dispersal initiated by the nineteenth-century 'Liberal' regime's (1863) abolition of entails. This process was preceded by the Enlightened Reformism's pressure to curtail new foundations and succeeded by the proclamation of the Republican regime in 1910, signalling the end of the nobility. Over the period from approximately 1750 to 1910, entail archives underwent significant changes that substantially altered their nature. Ultimately, these archives faced fates such as immediate destruction, transfer to buyers of the newly 'free estates' who subsequently destroyed them after aligning with the new legal system, or

preservation within families for sentimental reasons, subject to varied destinies throughout the twentieth century, including division or sale in auctions.

Moreover, the establishment of the National Archives significantly contributed to the fragmentation of archives from the two central institutions with which entails interacted – the Crown and the Church. Another contributing factor lies in 'classical' historiographical practices, such as constructing prosopographical databases, which often overlook or neglect the origin of archival information. Additionally, these practices are sometimes designed without consideration for the institutional nature of entails and the formal production of information, records, and archives. The use of anachronistic or simplistic proposals for the organization and classification of materials by archivists, coupled with a failure to study the institution and acknowledge the institutional distinctiveness of societies like the Ancien Régime, further compounds the issue.

In response to these challenges, VINCULUM has set as its primary and essential task the reorganization of information, facilitated by the AtoM database. Through the meticulous reorganization of documents associated with each entail – considered as both the creator institution and the archival custodian – while transparently documenting the provenance of each document, VINCULUM seeks to mitigate the dispersion of historical information.

4.3 Data Architecture

The database is constructed using AtoM, an archival database employing open-source software with a long-standing and widespread use across various countries, ensuring continuity. This platform adheres to archival standards for record treatment and facilitates the inclusion of digital copies of the documents.[6] AtoM is structured around several interconnected internal databases, namely, 1) Archival institutions (renamed for clarity as 'Entail/vínculo'); 2) Authority records; 3) Archival descriptions.

The single unit of data entry and organization adheres to both archivistic principles and the project's theoretical framework, considering entailed entities as 'Archival Institutions' and describing them based on a simplified version of ISDIAH.[7] Leveraging AtoM enables the tracking of documents back to their creators, regardless of their eventual storage location. The associated 'Authority Records'

6 https://www.accesstomemory.org/; https://accesstomemoryfoundation.org/.

7 https://www.ica.org/en/isdiah-international-standard-describing-institutions-archival-holdings.

database, aligning with archival standards (ISAAR-CPF), accommodates prosopographies of individuals and institutions linked to entailed entities.[8] As the project's historical research progresses, these prosopographies may expand to include supplementary information on people and institutions, facilitated by the inclusion of familial relationships within the methodology of Authority records, aiding kinship analyses (Gago 2017; Dryden 2007).

Document descriptions, following ISAD (G) and diplomatic description rules, are stored in the 'Archival Descriptions Database'.[9] A foundational project principle involves describing all documents according to the chain of diplomatic and archival transmission over centuries. This entails documenting not only the document conveying information about the historical subject but also its entry into the archive – successive copies, and/or the institutional path that produced it. Given that a significant majority of entailment documents exist only in later copies, this approach introduces additional workload and necessitates numerous description decisions. However, it offers the advantage of situating information within its authentic context, encompassing transcription errors, truncations, and previous archival choices.

Documents of specific typologies in the database, such as institution charters, wills, family partitions, and court sentences, undergo indexing by subject and documentary expression (the semantic of entailment). This indexing employs controlled terminology and thesauri, and the entire dataset is categorized geographically.

4.4 Integration of the Documentary Information in the Global Information System

In addition to finalizing the various 'entail archives', the plan for future years involves establishing connections between the documentary information compiled in the database and information available in other formats, such as images and architectural elements. Written documents of different natures will also be correlated with the entailed entities that produced them, whenever applicable. This second phase of information compilation and integration is likely to be conducted through a sampling method, focusing on selected case studies.

8 https://www.ica.org/resource/isaar-cpf-international-standard-archival-authority-record-for-corporate-bodies-persons-and-families-2nd-edition/.

9 https://www.ica.org/en/isadg-general-international-standard-archival-description-second-edition. Regarding diplomatic description, we followed the Portuguese rules, adapted for International standards: Costa 1993.

5 Constructing the (VINCULUM) Information System Guide

The database includes a complementary component, the 'Information System Guide', which represents another significant output of the project. This guide systematically presents the materials collected and structured within the database based on the institution that created them and the flow of information. Additionally, it provides an overview of the current archival arrangement and description of the records in their respective archives. Moreover, it aims to elucidate the processes of document creation and subsequent conservation of social information, considered fundamental for initiating any research.

The 'Information System Guide' endeavours to reconstruct the institutional, administrative, and information production system that evolved through the practice of entailment from the fourteenth to the seventeenth centuries, both internally and externally. It comprises two primary sections: the design of the institutional network and the modalities of information production, record creation, and archive establishment.

The central research problem involves challenging conventional notions regarding archives associated with entails, commencing with historical archives as understood in contemporary terms and posing questions such as "What are their origins? Why and how were they created?". This approach aligns with the latest theoretical reconceptualization of the role of 'sources' in historical writing, emphasizing that all archival records were initially documents of their respective eras. According to the tenets of Archival and Information Science, the act of information generation precedes both 'documentalization' and the record itself.

In a framework of record organization and utilization aimed at respecting the structure of information and document production, it is imperative to comprehend the institutions responsible for generating that information and their administrative operations. Subsequently, it becomes crucial to provide comprehensive descriptions of the documents and delineate the legal boundaries that governed the existence of entails. Accordingly, the 'Information System Guide' is structured into three principal sections, facilitating a comprehensive understanding of the multifaceted aspects underlying the information system of entails:

1. 'institutions' involved in producing, receiving, and/or preserving information pertaining to entails, including the Crown, the Church, and the entails themselves. Information regarding each institution is categorized based on: Chronological span; Normative documents; Competencies; Institutional

structure and the roles of its agents in relation to entails; Interactions with other institutions concerning entails;[10]

2. 'documents' produced by entailment, namely the document types identified and their diplomatic analysis.[11] The typological analysis of the documentation regarding entails, provides for each type, whenever applicable: name, specific bibliography, definition, author/creator, addressee(s), legislation, processing, organization of documents within the typology/related documents, administrative validity, and registration and archiving conditions. It was deemed necessary to conduct diplomatic analysis of the identified typologies of documents related to entails rather frequently. This involves presenting, for each typology and whenever applicable: reference and date of the analysed documents, parts of the documents/elements, protocol (*invocatio*, *intitulatio*, *inscriptio*, and *salutatio*), text (*notificatio*, *expositio*, *dispositio*, and final clauses), and eschatocol (the closing protocol noting the topical and chronological date, the validation and secretarial notes). The documents were selected by the project team and are indicated in each diplomatic typological framework;
3. 'civil and ecclesiastical legislation' comprises the collection of laws that regulated the operation of entails between the fourteenth and seventeenth centuries, produced by the Crown and the Church, coexisting in a dynamic relationship with the central law of entailment, the 'will of the founder'.[12]

To these three sections are added the bibliographical references that include sources, and working tools used in the preparation of the texts of the 'Information System Guide'.[13]

10 https://visg.vinculum.fcsh.unl.pt/en/instituicao/crown/; https://visg.vinculum.fcsh.unl.pt/en/instituicao/church/; https://visg.vinculum.fcsh.unl.pt/en/instituicao/entail/.

11 https://visg.vinculum.fcsh.unl.pt/en/documentos/.

12 https://visg.vinculum.fcsh.unl.pt/en/legislacao/.

13 https://visg.vinculum.fcsh.unl.pt/en/referencias/.

6 The Construction of a Database. A Challenge for Transparency and User Education

VINCULUM database is now[14] approaching 7,100 'entails' (considered as 'Archival holding institutions'), 24,826 document producers' biographies ('Authority records'), and 33,765 documents' summaries (in English) ('Archival descriptions', of which c. 26,692 are items and 7,073 are fonds). It contains c. 13,150 indexations by subject and c. 5,630 indexations by place. It also contains a high number of errors and mistakes, which will have to be corrected before publishing it online. Furthermore, a significant number of documents remain to be inserted, many of which originate from complex and largely undescribed archival collections. These documents necessitate paleographical skills and diplomatic training for accurate transcription and interpretation.

In the realm of historical research, the construction of extensive documentary databases is a task for which historians are typically unprepared during their academic training at universities and research institutions. The development of vast databases such as VINCULUM presents uncharted territory, especially for medievalists and modernists. The following discussion should be viewed within the context of addressing this educational gap within the profession, which cannot be entirely rectified by technology, regardless of how sophisticated and modern the databases may seem.

One primary concern is the necessity for standardization in data collection and processing. In the VINCULUM case it was imperative to establish a set of protocols and guidelines for material collection and input, with ongoing monitoring of their implementation. Initially, a weekly database correction system was implemented, whereby errors were identified by researchers not directly involved in data entry and notified to the members working on the database. However, it became apparent that errors were only effectively addressed when corrected by their creators. Subsequently, a more time-consuming but significantly more effective correction system was introduced. Additionally, a comprehensive set of working documents is being prepared, containing information on examined archival collections, documentary analysis, and rejected documents due to dubious or incomplete content. Weekly meetings and written reports have been conducted since the project's inception to discuss document options, all of which are meticulously documented and will be made available on the project's website.

Another critical issue necessitating specific training for historians is the understanding of standardized document description and

14 At the date of April 2024.

the inclusion of the document transmission chain. The modernization of research through database creation alone is illusory, particularly concerning prosopographical databases. The notion of collecting 'one more document' about individuals to be studied often obscures the importance of obtaining correct and comprehensive information about data provenance, resulting in inevitable distortions in samples. Moreover, the historians' lack of archival education and their reliance on footnotes, often incomplete or incipient, for document identification is problematic and more so if multiplied in order to show the author's erudition. Instead, it is essential to heed Marc Bloch's call, lately taken up by some medievalist historians, for a rigorous presentation of sources and their treatment at the outset of research endeavours (Anheim 2012).

Archivists, amidst their postmodern self-criticism and elsewhere, now deem it essential to present the 'archival footprint', commencing with a relativization of their neutrality (Cook 2009, 533). Historians, despite decades of acknowledging the subjectivity inherent in their profession, often still rely on a simplified positivist approach regarding their relationship with sources, which, in reality, constitute a complex and diverse world requiring diligently undertaken exploratory endeavours. Regardless of the potential price of this approach being the relativization of documentary evidence, the notion of historicizing the archive has long been advocated by anthropologists, yet not systematically embraced by historians, despite the increasing number of studies on the subject (Dirks 2002).[15]

One of the main archival challenges are the vast holdings of royal and ecclesiastical institutions from the Ancien Régime, particularly in Portugal, which remain largely unclassified and undescribed. Their integration into the National Archives was often disorderly, reflecting a lack of institutional and administrative history documentation. The documents were thus transferred between offices, warehouses, and, regrettably, discarded before finding their way into state archives. Consequently, the institutional and administrative history conducted within the framework of the 'Information System Guide' evolved from a theoretical principle into a practical necessity. Despite the technological sophistication of archival databases, they may exhibit flaws akin to those found in historical databases, including a dearth of information on the institutional and collection history data (the pivotal 'custodial history'), significant variability in the

15 The increasing historical approach can be seen though a sequence of journals' special numbers: "Archival Knowledge Cultures in Europe, 1400-1900", *Archival Science*, 10 (September 2010); "Archival Transformations in Early Modern Europe", *European History Quarterly*, 46(3) (July 2016); "The Social History of the Archive: Record-Keeping in Early Modern Europe", *Past & Present*, 230(11) (November 2016). More recently the trend extends to "Information", cf. Blair et al. 2021.

texts, and incomplete and/or inconsistent descriptions of the fonds. Additionally, the lack of collaboration between the two fields is particularly evident here, with historians largely neglecting institutional and collection history studies, and archivists facing challenges in undertaking this task independently. Now more than ever, collaboration between the two disciplines is imperative.[16]

Lastly, it is crucial to underscore the importance of educating users of historical, archival, or archival-historical databases. Promoters of these databases must provide clear and comprehensive explanations of how they were constructed, elucidating the various options involved. Users, on their part, must familiarize themselves with 'reading the instructions'. Those of us who teach history courses bear a special responsibility in this regard. Young researchers often seek 'content' that can substantiate their initial research ideas, with limited patience for longer timelines to obtain results in today's generation. Simply clicking on 'search' and gathering results may seem easier than delving into detailed explanations of the challenges and complexities inherent in archival collections. However, this problem existed, albeit in different forms, even before the digital era. How many history professors instructed their students that the 'reference room' of archives and their 'access tools' were (and are) artefacts, not neutral portals? Conversely, how many urged 'more diligent' students to explore everything, hinting that novelty and legitimization of their research lies in years of archival work, meticulously examining hundreds of scrolls, bundles, and boxes?

An important contribution to rethinking the research practices of various archival user groups, including historians, university professors, and researchers who may believe they do not require such training, comes from studies on information literacy or information-seeking behaviour. Information scientists in the field of archives have extensively developed this area in recent decades, alongside studies on knowledge organization (Rhee 2014; Torou 2006; Troitiño 2018). Both perspectives are of significant interest to database developers and users. On one hand, they provide insights into what different user groups seek and how they search for it, enabling the teaching of effective search methods and the development of user-friendly yet rigorous information access tools. On the other hand, in database planning, the 'organization of knowledge' underscores the responsibility involved and the situationalities that must be considered or avoided.

As we near the completion of several years dedicated to constructing a database aimed primarily at advancing scientific progress, our hope is that despite any flaws it may possess, it also embodies some of these virtues.

16 For an overall revision cf. Poole 2015 and Blouin 2019.

Bibliography

Anheim, É. (2012). "L'historien au pays des merveilles". *L'Homme*, 203-4, 399-427. https://doi.org/10.4000/lhomme.23239

Blair, A. et al. (eds) (2021). *Information: A Historical Companion*. Princeton: PUP.

Blouin, F.X. (2019). "Epilogue: A Perspective on the Archival Divide". Rosa, M. de Lurdes et al. (eds), *Recovered Voices, Newfound Questions: Family Archives and Historical Research (14th to 19th Centuries)*. Coimbra: Imprensa da Universidade de Coimbra, 367-78.

Cammarosano, P. (1991). *Italia medievale. Struttura e geografia delle fonti scritte*. Roma: Carocci.

Clanchy, M.T. (1980-81). "'Tenacious Letters': Archives and Memory in the Middle Ages". *Archivaria,* 11, 115-25.

Contreras Jiménez, M.E. (2016). *Linaje e transición histórica: los Arias Dávila entre el Medioevo e la Modernidad* [Ph.D. Dissertation]. Madrid: Universidad Complutense.

Cook, T. (2009). "The Archive(s) is a Foreign Country: Historians, Archivists and the Changing Archival Landscape". *The Canadian Historical Review*, 90(3), 497-534.

Cook, T. (2013). "Evidence, Memory, Identity, and Community: Four Shifting Archival Paradigms". *Archival Science,* 13, 95-120.

Costa, A. de Jesus da (1993). *Normas gerais de transcrição e publicação de documentos e textos medievais e modernos*. 2nd ed. Coimbra: FL-UL.

Dirks, N. (2002). "Annals of the Archive: Ethnographic Notes on the Sources of History". Axel, B. (ed.), *From the Margins: Historical Anthropology and Its Futures*. Durham: Duke University Press, 47-65.

Dryden, J. (ed.) (2007). *Respect for Authority. Authority Control, Context Control, and Archival Description*. New York: Routledge.

El-Leithy, T. (2011). "Living Documents, Dying Archives: Towards a Historical Anthropology of Medieval Arabic Archives". *Al-Qantara*, 32(2), 389-434.

Gago, A.B. (2017). "A importância dos registos de autoridade arquivística nos arquivos de família: estudo de caso do Arquivo Almada e Lencastre Bastos". *Boletim do Arquivo da Universidade de Coimbra*, 30, 451-93.

Guerreau, A. (1980). *O feudalismo: um horizonte teórico*. Lisboa: Ed. 70.

Guerreau, A. (1990). "Fief, féodalité, féodalisme. Enjeux sociaux et réflexion historienne". *Annales E.S.C.*, 45(1), 137-66.

Iranzo Muñio, M.T. (2009). "Arqueología del archivo: inventarios de los condes de Aranda". Casaus Ballester, M.J. (ed.), *El Condado de Aranda y la nobleza española en el Antiguo Régimen*. Saragoça: Instituicion Fernando el Catolico, 85-114.

Irigoyen, A. (2004). "La Iglesia y la perpetuación de las familias: clero y mayorazgo en Castilla durante el Antiguo Régimen". Ghirardi, M. (ed.), *Familias iberoamericanas ayer y hoy. Una mirada interdisciplinaria, Programa de Estructuras y Estrategias familiares*. Córdoba: Centro de Estudios Avanzados, Universidad Nacional de Córdoba; ALAP Editor, 113-33.

Kuchenbuch, L. (2004). "Sources ou documents? Contribution à l'histoire d'une évidence méthodologique". *Hypothèses 2003. Travaux de l'École doctorale d'histoire de l'université Paris 1 Panthéon-Sorbonne*, 1, 287-315.

Lodolini, E. (1991). "El problema fundamental de la archivistica: la naturaleza y el ordenamiento del archivo". Gutiérrez Muñoz, C. (ed.), *Materiales de Enseñanza de la Facultad de Letras y Ciencias Humanas*. Peru: Pontificia Universidad Católica de Perú, 30-51.

Martínez Perera, M.Á. (2010). "Ceremonial y rituales de posesión en la formación del mayorazgo". García Luján, J.A. (ed.), *Nobleza y monarquía*: *los linajes nobiliarios en*

el Reino de Granada, Siglos XV-XIX: el linaje Granada Venegas, Marqueses de Campotéja = Actas del Simposio (Huéscar, 16-18 septiembre 2010). Huéscar: Asociación Cultural Raigadas, 229-44.

Monteiro, N.G. (2001). "Trajectórias sociais e formas familiares: o modelo de sucessão vincular". Chácon Jiménez, F.; Franco, J.H. (eds), *Familias, poderosos y oligarquías*. Murcia: Universidad de Murcia, 17-37.

Poole, A.H. (2015). "Archival Divides and Foreign Countries? Historians, Archivists, Information-Seeking, and Technology: Retrospect and Prospect". *The American Archivist*, 78(2), 375-433.

Quintanilla Raso, M.C. (2004). "Propiedad vinculada y enajenaciones: métodos y lógicas nobiliarias en la Castilla tardomedieval". *Historia. Instituciones. Documentos*, 31, 493-510.

Rhee, H.L. (2012). "Modelling Historians' Information-Seeking Behaviour With an Interdisciplinary and Comparative Approach". *Information Research*, 17(4), paper 544. https://informationr.net/ir/17-4/paper544.html

Rosa, M. de Lurdes (1995). *O morgadio em Portugal, séculos XIV-XV. Modelos e formas de comportamento linhagístico*. Lisboa: Estampa.

Rosa, M. de Lurdes (2012). *As almas herdeiras: fundação de capelas fúnebres e afirmação da alma como sujeito de direito (Portugal, 1400-1521)*. Lisboa: IN-CM.

Rosa, M. de Lurdes (2017). "Reconstruindo a produção, documentalização e conservação da informação social pré-moderna. Perspectivas teóricas e proposta de percurso de investigação". *Boletim do Arquivo da Universidad de Coimbra*, 30, 547-86.

Rosa, M. de Lurdes (2019). "The Production of Genealogical Knowledge for the Arrangement of Noble Family Archives in Portugal, 15th-Early 19th Century". Eickmeyer, J.; Friedrich, M.; Bauer, V. (eds), *Genealogical Knowledge in the Making: Tools, Practices and Evidence in Early Modern Europe*. Berlin; Boston: De Gruyter, 245-89.

Rosa, M. de Lurdes (2020). "Preventing Household Failure: Figures of Authority in Familial Corporate Bodies (The Portuguese Morgadio System From the Fourteenth to the Seventeenth Century)". *L'Atelier du Centre de recherches historiques,* 22. https://doi.org/10.4000/acrh.11096

Rosa, M. de Lurdes (2021). "Penser et organiser les archives de famille, entre histoire et archivistique". Lamazou-Duplan, V. (ed.), *Trésor, Arsenal, Mémorial: les archives familiales en péninsule Ibérique et dans l'Occident chrétien (moyen âge, première modernité)*. Madrid: Casa de Velázquez, 63-76.

Rosa, M. de Lurdes (2022). "Ouvertures et fermetures des archives de famille. Portugal et Péninsule Ibérique, XIVe–XXIe siècles". Péquignot, S.; Potin, Y. (eds), *Les conflits d'archives. France, Espagne, Méditerranée*. Rennes: Presses Universitaires de Rennes, 243-58.

Rosa, M. de Lurdes (2024). "História, Ciências sociais e humanas, Ciência arquivística, Ciência da informação. Caminhos para a criação de espaços científicos comuns". *Boletim do Arquivo da Universidade de Coimbra*, Extra 1, 209-73. https://doi.org/10.14195/2182-7974_extra2024_1_7

Rosa, M. de Lurdes; Head, R.C. (2015). "Delineating the Social Complexity of Archival Practices: The Objectives and Results of the INVENT.ARQ Project on Family Archive Inventories". Rosa, M. de Lurdes; Head, R.C. (eds), *Rethinking the Archive in Pre-Modern Europe: Family Archives and their Inventories From the 15th to the 19th centuries*. Lisboa: IEM; NOVA.FCSH; UNL, 9-22.

Roth, D. (2007). "Mayorazgos, capellanías y lugares de memoria como perpetuación del ascenso social de la oligarquía de un centro administrativo de señorío. El ejemplo de Vélez Blanco (1588-1788)". Andújar Castillo, F.; Díaz López, J.P. (eds), *Los*

señoríos en la Andalucía Moderna. El Marquesado de los Vélez. Almería: Instituto de estudios Almerienses, 213-34.

Soria Mesa, E. (2007). *La nobleza en la España moderna: cambio y continuidad*. Madrid: Marcial Pons Historia.

Soria Mesa, E. (2009). "La nobleza en la España moderna. Presente y futuro de la investigación". Casaus Ballester, M.J. (ed.), *El Condado de Aranda y la* nobleza *española en el Antiguo Régimen*. Saragoça: Instituicion Fernando el Catolico, 213-41.

Sottomayor-Pizarro, J.A. de (2011). "Linhagem e estruturas de parentesco". *e-Spania*, 11.
https://doi.org/10.4000/e-spania.20366

Stauffer, S.M. (ed.) (2021). *Libraries, Archives, and Museums. An Introduction to Cultural Heritage Institutions Through the Ages*. Washington: Rowman & Littlefield Publishers.

Torou, E. et al. (2006). "Creating an Historical Archive Ontology: Guidelines and Evaluation". *Proceedings of the 1st International Conference on Digital Information Management (ICDM 2006).* Bangalore: IEEE Press, 102-9.

Troitiño, S. (2018). "Different Parameters for Knowledge Organization in Archives". Ribeiro, F.; Cerveira, M.E. (eds), *Challenges and Opportunities for Knowledge Organization in the Digital Age = Proceedings of the Fifteenth International ISKO Conference* (Porto, 9-11 July 2018). Baden-Baden: Ergon-Verlag, 160-6.

Models of Data Extraction and Architecture in Relational Databases of Early Modern Private Political Archives
edited by Dorit Raines

Cracking the Historical Code
From Unstructured Correspondence Corpora to Computational Analysis

Agata Bloch
Tadeusz Manteuffel Institute of History of Polish Academy of Sciences, Poland

Clodomir Santana
Tadeusz Manteuffel Institute of History of Polish Academy of Sciences, Poland

Demival Vasques Filho
University of Luxembourg, Luxemburg

Michał Bojanowski
Kozminski University, Poland

Abstract The chapter addresses a methodological approach to unstructured data and discusses the potential that structured data offers in the field of historical research. The dataset, which initially consists of textual content sourced from digital collections at the Portuguese Overseas Archives in Lisbon, undergoes a preprocessing phase that forms the basis for the extraction of structured data. The authors combine history, social sciences, and computer science to convert the correspondence repository into a machine-processable form. This transformation is supported by an interdisciplinary strategy in which they weave together elements of effective content management, topic modelling, and social network analysis.

Keywords Public correspondence. Colonial Portuguese Empire. Structured data. Digital infrastructure. Historical dataset.

Summary 1 Paper versus Pixel. Archival Practices in the Digital Era. – 2 Navigating Big Data in Colonial Correspondence. – 3 Decoding the Past. From Manuscripts to Metadata. – 4 Interdisciplinary Strategies for the Structured Data. – 5 Breaking Barriers Between 'Digital' and Historians.

Edizioni Ca'Foscari

Studi di archivistica, bibliografia, paleografia 8
e-ISSN 2610-9093 | ISSN 2610-9875
ISBN [ebook] 978-88-6969-919-1 | ISBN [print] 978-88-6969-920-7

Peer review | Open access
Submitted 2023-10-03 | Accepted 2024-01-18 | Published 2025-05-22

DOI 10.30687/978-88-6969-919-1//006

1 Paper versus Pixel. Archival Practices in the Digital Era

Digital history encompasses a wide range of digital methods used by historians to open up new perspectives on the past. These methods have a remarkable capacity to enable new explorations of historical narratives that go beyond conventional approaches and are particularly welcomed by younger generations of historians. Although the purpose of this paper is not to explore the nature of digital historiography, we can agree on one fundamental aspect: the application of the latest technologies in the field requires constant experimentation (Jaillant et al. 2022).

In this paper, we would like to explore the potential of archival collections – especially correspondence – in the application of digital archiving techniques and computer-assisted analysis. Before we do so, it is important to distinguish between three categories of storage modalities for the collections in question. These distinctions are important for potential use by digital historians and provide different opportunities for scholarly work. The first modality is archives that have never been digitised and can be classified as 'endangered'; the second category includes mixed collections consisting of both paper documents and their digital copies in repositories; and finally, born-digital archives originally created in a digital format.

The first category includes archival collections around the world that have not received adequate physical and digital preservation of their documents. They can benefit from the Endangered Archives Programme (EAP), an initiative that allocates financial resources to facilitate the digitisation processes of primary sources.[1] Over the past ten years, more than eleven million images have been digitised under this program. In addition, new collections from South Africa, India, Nepal, and Georgia were made available online through the British Library Catalogue in 2003 (Supple 2015).

The second modality consists of mixed collections that contain both physical and digitised documents. One such example, among many others, is the Portuguese *Biblioteca Nacional Digital*, where users can access well-organised but still unstructured material through basic metadata. Another archive collection in this category is *Projeto*

Authors acknowledge the support of National Science Centre of Poland through the grant *2022/45/B/HS3/00473: Imperial Commoners of Brazil and West Africa (1640-1822): global history from a correspondence network perspective,* National Science Centre of Poland. For additional information about our website *MAPE – Mapping Atlantic Portuguese Empire*: https://www.projectmape.org/.

1 See *The Endangered Archives Programme* of the British Library at: https://eap.bl.uk/#:~:text=The%20Endangered%20Archives%20Programme%20%28EAP%29%20facilitates%20the%20digitisation,in%20danger%20of%20destruction%2C%20neglect%20or%20physical%20deterioration.

Resgate, which is evolving digitally. These collections have moved from the earlier practice of photographing and storing materials on CD-ROMs to being searchable via more advanced search engines. The transition from the first to the second modality is challenging because it requires the funding of expensive electronic tools and is time consuming.

However, the gap between the second and third modality is even wider. We have to remember that digitisation alone will not make historical sources machine-readable. An example of such born-digital archive is the Brazilian project *#MemóriasCovid19*, which collects individual image, text, video, and audio material that was originally intended to exist only in digital formats (Nicodemo, Marino 2022).[2]

When discussing the various storage modalities for the collections, it is important to note that they do not imply a hierarchical progression or mean that one category is superior to the others. Each modality faces its own challenges and considerations. Born-digital archives, for example, while considered promising, still face obstacles such as copyright and privacy concerns (Jaillant et al. 2022). Our approach to digital archival practices places us in the second category, which involves the creation of an integrated, machine-readable[3] relational database derived from non-structured public correspondence. We do not intend to move into the third category, but we are experimenting with a digital approach to extracting, organising, and further analysing relational information from this particular type of archival material.

Regarding materials suitable for digital humanities research, Niels Brügger classifies them into three types: digitised, born digital, and reborn digital. Digitised documents include original archival materials that have been digitised. Born digital material is specifically designed to exist solely in digital format, without a paper equivalent. Reborn digital material, in turn, falls into the same category as born digital material, but has been modified during the preservation process (Brügger 2016, 5-7).

In this paper, we explain our approach, in which we started from a digital extended catalogue of an otherwise analogue collection of public correspondence. We have taken a number of systematic steps to make this collection more suitable for computer processing and interpretation. In the following sections, we introduce our extensive dataset in the context of Big Data and then address the extraction

2 https://memoriascovid19.unicamp.br/.

3 Machine-readable data present structures that allow computers to readily process information. The vast majority of text data is unstructured, i.e. these texts are available in plain text only. Examples of structured (machine-readable) text data are XML transcriptions where entities are tagged and identifiable.

of metadata from the manuscripts. Finally, we explore interdisciplinary approaches to digital methods that can be applied to research projects similar to ours.

2 Navigating Big Data in Colonial Correspondence

The current version of our complex dataset *MAPE – Mapping Atlantic Portuguese Empire*[4] contains 169,221 registers of colonial correspondences (160,892 between 1640 and 1822, which is the period of our interest) exchanged by 34,407 social actors establishing 48,173 relationships across Americas, Europe, Africa and Asia. A total of 1,235 colonial institutions existed, providing colonial individuals with opportunities to occupy 2,077 political roles and acquire 113 noble and land titles.

Working with such a massive amount of data classifies our dataset as 'Big Data'. The concept of 'Big Data' varies across different fields, but in our case, it refers to a historical context in which large-scale information is studied using data-intensive methods (Eijnatten, Pieters, Verheul 2013). Because of its scale, this data exceeds the capabilities of standard analytical processing software (Blaney 2021).

But can numbers alone really tell us anything? While Chris Anderson believed that big numbers speak for themselves (Anderson 2008), Trevor Barnes expressed concern that they generate too much noise and indeed provide little historical insight (Barnes 2013). We see it differently: without the right methods and digital capacity to analyse and interpret them, these numbers remain meaningless. Digital History should focus not only on organising unstructured data, but also on exploring additional possibilities. In this regard, we agree with Eijnatten's argument: "What large outstanding questions can historians hope to address by implementary digital humanities?" (Eijnatten, Pieters, Verheul 2013, 58).

Our aim is to use our relational dataset to explore big questions along three dimensions: firstly, to analyse the events and topics that are recorded within the correspondence, thereby scrutinising their spatial and temporal variability; secondly, to examine the interests and experiences of colonial societies encompassed by the Portuguese Empire, thereby discerning the evolving nature of these facets over time; and finally, to elucidate the roles and functions undertaken by officials within the expanse of the empire.

Regarding the first dimension of events and topics, we propose the application of topical classification to the documents, coupled with the utilisation of available metadata pertaining to temporal

4 https://www.projectmape.org/.

and geographical origins of each document in order to address three big questions:

a. What is the volume of correspondence related to different documents' topics as ascertained through topic modelling and how does it vary by location (e.g. geographic location of the sender) and time?
b. What role do historical events play in shaping the observed dynamics?
c. To what degree can the application of topic modelling and examination of dynamics explain changes in the overall scope of correspondence? Can the proliferation and dynamics of documents of different types and topics be explained by known historical events in the colonies, the Portuguese Empire, or the broader global milieu?

Regarding the second dimension of exploring the interests and experiences of the colonial societies, we adopt the suggestions of Graham, Milligan, Weingart (2015) and François et al. (2016) to create a 'macroscope' of the societies of the Atlantic colonies based on the information available in the correspondence. Using a collective study approach, referred to as "a group biography" (Abbott 2009), we investigate the social characteristics of the colonised societies. The key questions we propose to investigate:

a. Who were the people documented in the correspondence originating from Brazil and West Africa, and how did their representation evolve over time?
b. How did topics addressed in the correspondence evolve with respect to gender, social status, and differences in the discussed issues?
c. Who played the most active role in economic, political, and legal affairs?
d. Which regions of colonial Brazil or West Africa were more involved in economic, political, and legal affairs?
e. Finally, what were the concerns of colonised societies related to private life, public life, family, economy, and social progress within the imperial context?

In the third dimension, our goal is to comprehend the roles and functions of officials within the Portuguese Empire. We analyse their involvement, responsibilities, and positions within the correspondence network. By studying how their participation and roles evolved over time, we aim to gain insights into the recipients and nature of petitions from colonial inhabitants.

Recent publications on the political history of the Portuguese Empire emphasise the involvement of various social actors in

constructing the colonial empire.[5] With this in mind, we explore the key actors to whom inhabitants in Brazil and West Africa turned and furthermore, we examine the similarity of the issues they addressed.

Moreover, we endeavour to comprehend the progression of network roles within the cadre of civil servants, exploring the potential interconnection between this evolution and two key factors: the delegation of responsibilities by the monarch to local authorities, and the increase of regal authority as driven by the aspirational absolutist pursuits of the Portuguese court.

In addition, we examine whether petitions on similar problems consistently reached the same officials and to what extent these patterns varied based on the role or function of each official.

Exploring these big research questions across the three dimensions can greatly enhance our understanding of the colonised societies within the Portuguese Empire. This exploration includes their various attributes, interactions, shared experiences, narratives, expectations, aspirations, negotiation patterns, and social dynamics. It is important to recognise that colonialism had profound effects on the political, social, and economic spheres and still affects more than three quarters of the world's population today (Ashcroft, Griffiths, Tiffin 2002). Studying the history of these societies requires extensive archival efforts and a multidisciplinary approach to data management (Cuartero, Gómes 2012). Our effective approach is to use a structured relational dataset for network analysis. Scholars have combined early modern correspondence with social network analysis to study the relationships between individuals over time and across geographic boundaries. This methodology allows one to trace individual life trajectories at the social level and understand how power is consolidated at the institutional level.[6]

Correspondence analysis can be approached in three ways. First, it involves identifying the most common senders and recipients of letters (McShane 2018; Ahnert, Ahnert 2015). Second, semantic analysis can be conducted to examine rhetorical strategies (Franzosi 1998; McLean 2007). Finally, it determines the relational level between social actors. In addition to these existing approaches, we apply a prosopographical method to explore the discursive patterns of Atlantic colonial societies that are often overlooked when focusing only on individual biographies. By examining their trajectories, including their trans-imperial and intercultural connections, relationships with various social actors, and rhetorical strategies, we can uncover patterns of communication that reveal shared characteristics, narratives, and

5 Ramos et al. 2009; Thornton 2012; Costa et al. 2014; Havik, Newitt 2015; Fragoso, Monteiro 2017; Fusaro, Polónia 2017; Xavier et al. 2018; Domingos 2021.

6 Ahnert, Ahnert 2015; McShane 2018; McLean 2007; Padgett, Ansell 1993.

cognitive processes across the Atlantic that transcend class and gender boundaries. Our perspective considers the Portuguese empire as a dynamic space characterised by networks and grassroots communication on the one hand, and globally significant historical phenomena affecting the actions of colonial societies on the other.

Additionally, apart from studying the sender, recipient, timing, and subject, we aim to investigate general political changes at the macro level. We believe these changes are reflected in the structure of correspondence networks, illustrating how different types of information reached officials, including the monarch. The concept of "connected histories" (Subrahmanyam 2007) serves as a framework for analysing the history of colonial and imperial systems from a global perspective. We apply this theory to uncover not only commonalities in their "connected histories" but also to examine divergent paths within the shared history of the Atlantic empire. Despite political or structural similarities, concepts such as race, violence, and knowledge transfer may have developed differently in these regions. Thus, we aim to highlight the significant role of colonies and the workings of the empire's colonial societies.

The idea of studying local and global events through the analysis of correspondence, reports, or information has gained attention in recent decades. Contemporary political scientists are looking for data and methods to learn about ongoing political events based on processing large volumes of local reports and local media outlets. Examples include the Global Database of Society (GDELT project)[7] and the Integrated Crisis Early Warning System (ICEWS)[8] (O'Brien 2010; 2013). The latter was created by the U.S. government as a kind of 'early warning system' for major events in specific regions (e.g. the Middle East) to identify key actors (individuals and organisations), relationships among them (e.g. collusion or hostility), and events that could reach a global dimension. Other research-oriented projects of a similar nature

7 The GDELT project "monitors the world's broadcast, print, and web news from nearly every corner of every country in over 100 languages and identifies the people, locations, organisations, themes, sources, emotions, counts, quotes, images and events driving our global society every second of every day, creating a free open platform for computing on the entire world" (`https://www.gdeltproject.org/`).

8 The Integrated Crisis Early Warning System (ICEWS) combines a database of political events and a system using these to provide conflict early warnings (`https://www.lockheedmartin.com/en-us/capabilities/research-labs/advanced-technology-labs/icews.html`).

include the CAMS EvOlution project (CAMEO)[9] (Gerner et al. 2002) and the Social Conflict Analysis Database (SCAD)[10] (Salehyan et al. 2012).

Our corpus shares similar characteristics with these projects as the correspondence comes from different regions of the Portuguese Empire and is addressed to Lisbon. It contains information about key local and global events, including the Restoration of 1654 in the Brazilian region of Pernambuco, the 1755 earthquake in Lisbon, the 1789 conspiracy in Minas Gerais, and the 1807-11 French invasion of Portugal. These local events mentioned in the correspondence can be cross-referenced and compared to a calendar of known global events. By employing social network analysis methods and furthermore, by building on the political science projects mentioned above, our analysis of the corpus has the potential to provide valuable insights into the social history of the Portuguese Empire at both the local and global levels.

We hypothesise that the communication patterns of colonial societies were influenced by global events described above. Consequently, these communities shared common characteristics (prosopography) as well as common ideas, narratives, and cognitive approaches to the Atlantic coast (global history). However, the empire's response to these influences may have varied, highlighting inequalities in the Global South. Colonised people not only acted as distinct social groups (indigenous, black, women), but also exhibited cross-class and cross-gender behaviours (group biography). To achieve our goals, we will conduct a comprehensive analysis using a variety of digital methods and network science. By examining relational data, we will be able to reconstruct prosopographic networks and examine the connections between colonial societies and government officials in terms of subject, time, and geographic context.

Based on the conceptual framework described here, we explain our methods for dealing with unstructured data in the following section.

9 The Copernicus Atmosphere Monitoring Service (CAMS) provides consistent and quality-controlled information related to air pollution and health, solar energy, greenhouse gases and climate forcing, everywhere in the world. The CAMEO project proposes to advance in the exploitation of new space observations, data assimilation and inversion techniques for the global and regional CAMS production systems. https://www.cameo-project.eu/.

10 "The Social Conflict Analysis Database (SCAD) includes protests, riots, strikes, inter-communal conflict, government violence against civilians, and other forms of social conflict not systematically tracked in other conflict datasets. SCAD currently includes information [on] social conflicts from 1990-2017, covering all of Africa and now also Mexico, Central America, and the Caribbean" (https://www.strausscenter.org/ccaps-research-areas/social-conflict/database/).

3 Decoding the Past. From Manuscripts to Metadata

For our research, we selected the Historical Overseas Archives (AHU) in Lisbon, which José Curto has called "the richest and most complete single depository of manuscripts relating to the administrative, economic, financial, military, political, and social history of the Portuguese overseas colonies" (Curto 1988, 164). The collections consist of documents from various colonial institutions, such as the *Conselho da Índia* (Council of India), *Conselho da Fazenda* (Treasury Council), and *Conselho da Guerra* (War Council), which preceded the more important and longest-running *Conselho Ultramarino* (Overseas Council), that operated – with some political interruptions – from 1642 to 1822.

After the Carnation Revolution in 1974 and the subsequent democratisation and decolonisation of Portuguese institutions, researchers gained greater access to the collections of the Overseas Council. This development was accompanied by an increased interest in creating inventories and guides. One of the challenges Curto noted in the late 1980s, and which we also encountered in our research, is that many of these repositories were not created by archivists from the *Arquivo Histórico Ultramarino* (Historical Overseas Archives) (AHU), but by different groups of researchers with different academic affiliations (Curto 1988, 165). In addition to the laudable efforts in cataloguing documents, our generation of digital humanists inevitably faces the difficult task of managing collections with hundreds of thousands of documents with non-machine-readable, unstructured text data.

Our research focuses primarily on specific collections within the AHU, arranged chronologically and geographically, which include the following:

1. The *Barão do Rio Branco* – Historical Documentation Rescue Project known as *Projeto Resgate* (Bertoletti et al. 2012; Boschi 2018) which includes 26 catalogues of documents referring to Brazilian regions, catalogued at different times and by different researchers.[11] The *Projeto Resgate* collection is currently managed by the National Library of Rio de Janeiro in Brazil, but is housed in the AHU.
2. The Angola Collection (*Série Angola*), whose cataloguing was financially supported by the Portuguese *Fundação para a Ciência e Tecnologia* as part of the project *África Atlântica:*

11 Catalogues of loose manuscript documents are the following: Brasil-Geral, Alagoas, Bahia, Bahia-CA, Bahia-LF, Ceará, Espírito Santo, Goiás, Maranhão, Mato Grosso, Minas Gerais, Pará, Paraíba, Pernambuco, Piauí, Rio de Janeiro, Rio de Janeiro-CA, Rio Grande do Norte, Rio Grande do Sul, Rio Negro, Santa Catarina, Sergipe d'el-Rei, São Paulo, São Paulo-MG, Brasil-Limites. See more: https://actd.iict.pt/collection/actd:CUF004.

da documentação ao conhecimento, sécs. XVII-XIX (Atlantic Africa: from documentation to knowledge, seventeenth to nineteenth centuries).

3. The Cabo Verde and Guinea Collections (*Série Cabo Verde, Série Guiné*), which were catalogued as part of two separate projects: the aforementioned *África Atlântica* and the *Resgate do acervo histórico de Cabo Verde em Portugal* (Rescue the historical collection of Cape Verde in Portugal) funded by Camões, *Instituto da Cooperação e da Língua* (ICL).[12]
4. The *São Tomé* Collection (*Série S. Tomé e Príncipe*), also catalogued within the *África Atlântica* project.
5. The Mozambique Collection *(Série Moçambique*), which was part of the *Projecto de Microfilmagem de Documentação sobre Moçambique existente em Portugal, destinado ao Arquivo Nacional de Moçambique* (Microfilming Project of Documentation about Mozambique Existing in Portugal, Destined to the National Archive of Mozambique), funded by the Swedish Agency for Research Cooperation with Developing Countries (SAREC).

According to Ruth Ahnert, a British digital historian, it is crucial to gain a deeper understanding of the "landscape of cultural data" and the digitisation policies followed in the countries where these collections are digitised and stored. Ahnert highlights that in England, which has a mixed funding system, the outcomes often involve partnerships between the public and commercial sectors. While these partnerships facilitate digitisation, they also impose certain limitations and copyrights (Ahnert et al. 2023, 23-4). Portuguese public archives, instead, are more accessible to researchers and allow them to work with the materials as long as they are not used for commercial purposes. In the context of the Digital Repository of the Portuguese *Arquivo Científico Tropical*, which functions as a repository for the holdings of the Portuguese Overseas Archives and other invaluable collections, it is crucial to underline that the materials encompassed therein are accessible without any cost to individuals involved in personal, educational, and scientific inquiries. However, for alternative intentions, notably those of a commercial nature, usage is contingent upon specified conditions and may solely be sanctioned with explicit authorisation.[13]

Additional collections housed within the Overseas Historical Archive include a diverse array of records such as accounts, statistical

12 https://www.instituto-camoes.pt/en/.

13 For the terms of use of the Digital Repository of *Arquivo Científico Tropical*, see https://actd.iict.pt/about.php?display=terms&lang=pt, section 2.

data, consultations, official proclamations, cartographic materials, reports, and censuses. Notwithstanding the extensive variety of materials contained within the archival holdings, it is imperative to acknowledge that these resources remain confined to their inherent physical manifestations in the form of original paper documents or microfilm reproductions. At present, our efforts are focused on curating summarised and digitised repositories for the aforementioned five collections, catalogued in Africa and Brazil.[14]

It is important to note that our specialisation does not involve the conversion of historical documents into machine-readable formats through the use of handwriting text recognition (Toledo et al. 2019). We rather follow Dong et al. (2021), Zbíral, Shaw (2022) and Ruth Ahnert et al. (2023) in accessing the data and converting human-readable documents into machine-interpretable data usable for computational analysis. Before our intervention, the digital repositories on Portuguese colonies lacked structure and were incompatible with widely used machine-readable formats such as CSV, RDF, XML, and JSON. It is important to comprehend that the mere adoption of a digital format does not inherently ensure machine-readability. Figures 1, 2, and 3 represent examples of non-machine-readable registers from two collections related to Bahia, São Paulo from *Projeto Resgate*, and one from Angola **[figs 1-3]**. They are non-machine-readable because we can digitally access (read) the text in these documents as plain texts only. That is, they lack tags or metadata with which we could readily extract useful information amenable to computational processing.

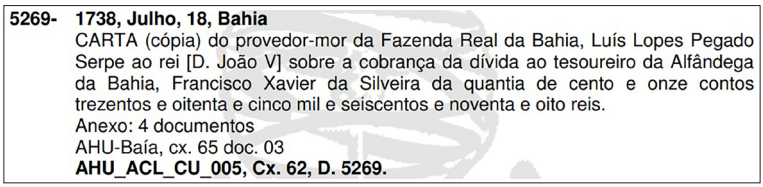

5269- **1738, Julho, 18, Bahia**
CARTA (cópia) do provedor-mor da Fazenda Real da Bahia, Luís Lopes Pegado Serpe ao rei [D. João V] sobre a cobrança da dívida ao tesoureiro da Alfândega da Bahia, Francisco Xavier da Silveira da quantia de cento e onze contos trezentos e oitenta e cinco mil e seiscentos e noventa e oito reis.
Anexo: 4 documentos
AHU-Baía, cx. 65 doc. 03
AHU_ACL_CU_005, Cx. 62, D. 5269.

Figure 1 Example of non-machine-readable document from Collection Brazil, *Projeto Resgate*, Bahia Luísa da Fonseca (1599-1700)

14 For *Conselho Ultramarino - Arquivo Histórico Ultramarino* (arquivos.pt), see https://digitarq.ahu.arquivos.pt/DetailsForm.aspx?id=1119329. For Funds of Portuguese Overseas Archive, see https://actd.iict.pt/collection/actd:CU.

1669, Janeiro, 30, Lisboa
CONSULTA do Conselho Ultramarino, sobre a petição de Francisco Vieira, morador na vila de Vitória, capitania do Espírito Santo, ao (Príncipe Regente D. Pedro), em que diz que em virtude de uma sua provisão, Agostinho Barbalho Bezerra, que foi administrador das minas de São Paulo, lhe concedeu o perdão do crime da morte de um homem, por o ter acompanhado na jornada das ditas minas e por estar inocente daquele crime, pede lhe mande confirmar o citado perdão. Pareceu ao Conselho que o (Príncipe Regente) deve mandar passar a confirmação pedida, a João Falcão de Sousa pareceu que a provisão do (Príncipe Regente) não deve ter lugar e não se deve mandar confirmar o dito perdão, pois é muito grave o crime de morte de homem, que deve ser castigado.
AHU-São Paulo-MGouveia, cx. 1, doc. 26.
AHU_CU_023-01, Cx. 1, D. 26.

Figure 2 Example of non-machine-readable document from Collection Brazil, *Projeto Resgate*, São Paulo Alfredo Mendes Gouveia (1618-1823)

[ant. 1618, Dezembro, 7]
REQUERIMENTO do [contratador de Angola e Cabo Verde], António Fernandes [de] Elvas, ao rei [D. Filipe II] solicitando que lhe fosse passada uma provisão a propósito dos direitos que se deviam pagar dos escravos que iam para as Índias, o Brasil e outras partes, tendo em conta as diferenças de entendimento que existiam entre os feitores da [Fazenda Real].
Obs.: m. est.
AHU-Angola, cx. 1, doc. 98.
AHU_CU_001, Cx. 1, D. 105.

Figure 3 Example of non-machine-readable document from Collection Angola

Consequently, these documents are perceived by computer systems as unstructured text. Digital historians face the challenge of determining what metadata they can extract from such documents for their research purposes. Based on the above examples, we propose to define the following metadata for the correspondence:

a. A unique identifier (ID) assigned to each record that enables linking and referencing individual correspondence.
b. Type of document: crucial to identify different types of correspondence, from royal charter to individual voices, and to analyse the documents based on their nature, such as:
   b.i *Carta* (letter)
   b.ii *Consulta* (consultation)
   b.iii *Requerimento* (petition)
c. Dates shall be put into a standard format for efficient searching and filtering based on specific dates or date ranges, and establishing a chronological order for the thousands of documents, e.g.:
   c.i 1738-07-18
   c.ii 1699-01
   c.iii 1618-12-07
d. Geographic location is another important metadata element, whether it is a village, city, state, or colony, because

it provides valuable information for spatial analysis. We can study communication patterns in different regions and examine the effects of geographic factors on correspondence. Some examples:
  - d.i City: Lisbon
  - d.ii State: Bahia
  - d.iii Colony: Angola
- e. Senders and recipients, with their relevant information about positions or titles, represent relevant metadata for building communication networks, identifying key figures, and analysing the roles and relationships of individuals. The following metadata applies to both senders and recipients:
  - e.i Sender name: André de Melo e Castro, António Fernandes de Elvas
  - e.ii Sender function: *Vice-Rei e Capitão-General, contratador*
  - e.iii Sender title: *conde das Galveias*

Managing the correspondence data in the repositories presented our first challenge. Knowledge management varied not only between territories, but even within the same colony, due to different archiving practices as well as human errors and misspellings. This inconsistency posed a significant obstacle to our research. On the other hand, these challenges highlighted the importance of adopting a digital approach very early in historical research (Reinke 1981; Hedstrom, Kowlowitz 1988). To overcome these obstacles, we first defined the relationship data and then processed the unstructured information to transform it into structured, machine-readable data. This required identifying and categorising the elements through annotations.

We used the SpaCy library,[15] a Natural Language Processing (NLP) tool written in Python, to identify the actors involved in the correspondence and their associated attributes. Since our sources are in Portuguese and the available Named Entity Recognition (NER) libraries for Portuguese have limited accuracy, especially for historical texts, we took the initiative to develop our own NER model. To do this, we had to manually identify and extract entities from a sample of 4,230 letters using programs such as MaxQDA, a software program designed for computer-assisted qualitative and mixed methods data, text and multimedia analysis,[16] and Prodigy, a scriptable annotation tool.[17] In addition to the standard categories, 'Person', 'Lo-

15 https://spacy.io/.

16 https://www.maxqda.com/.

17 https://prodi.gy/.

cation' and 'Organisation', we trained our model to recognise new categories such as 'Role', 'Affiliation', and (nobility) 'Title'.

The goal was to identify the senders and recipients of the correspondence, the location from which the correspondence was sent, and the organisations to which they belong. Successfully identifying the senders and recipients of each letter required searching for patterns in the text. Regular expressions proved useful in dividing the text into three segments: sender information, recipient information, and content. The additional attributes were identified using the NER model we developed. These entities are persons (male and female), noble titles, organisations (civil and secular institutions), military and religious institutions, occupations, and geographic locations. Finally, we created a network dataset in JSON format to represent the extracted information (Bloch, Vasques, Bojanowski 2022). Using our NER model, we achieved a hit rate of 93.1% of accuracy.

Our efforts to convert the examples of Figures 1-3 into a machine-readable JSON format are shown in Figures 4-6, respectively **[figs 4-6]**.


```
       "{
       'doc_id': 157876,
       'doc_type': 'CARTA',
       'date': '1738-07-18',
       'text': ' 5270- 1738, Julho, 18, Bahia CARTA do vice-rei e capitão-general do estado do
Brasil, André de Melo e Castro, conde das Galveias ao rei D. João V sobre o excesso que cometera
o frade do Carmo, Frei António de alcunha Pituba por ter acobertado o criminoso Francisco Gil
Garcia de Araújo',
        'sender': {'aff': [], 'title': ['conde das Galveias'], 'names': ['André de Melo e
Castro'], 'occ': ['vice-rei', 'capitão-general']},
       'recipient': {'aff': [], 'title': [], 'names': ['D. João V'], 'occ': ['rei']}
       }
```


Figure 4 Example of a machine-readable document from Collection Brazil, *Projeto Resgate*, Bahia Luísa da Fonseca (1599-1700)


```
        {
        'doc_id': 286336,
        'doc_type': 'CONSULTA',
        'date': '1669-01-30',
        'text': ' 26-  1669, Janeiro, 30, Lisboa CONSULTA do Conselho Ultramarino, sobre a
petição de Francisco Vieira, morador na vila de Vitória, capitania do Espírito Santo, ao
(Príncipe Regente D. Pedro), em que diz que em virtude de uma sua provisão, Agostinho Barbalho
Bezerra, que foi administrador das minas de São Paulo, lhe concedeu o perdão do crime da morte
de um homem, por o ter acompanhado na jornada das ditas minas e por estar inocente daquele
crime, pede lhe mande confirmar o citado perdão. Pareceu ao Conselho que o (Príncipe Regente)
deve mandar passar a confirmação pedida, a João Falcão de Sousa pareceu que a provisão do
(Príncipe Regente) não deve ter lugar e não se deve mandar confirmar o dito perdão, pois é muito
grave o crime de morte de homem, que deve ser castigado',
        'sender': {'aff': ['Conselho Ultramarino'], 'title': [], 'names': [], 'occ': []},
        'recipient': {'aff': [], 'title': [], 'names': ['D. Pedro'], 'occ': ['Príncipe
Regente']}
        }
```


Figure 5 Example of a machine-readable document from Collection Brazil, *Projeto Resgate*, São Paulo Alfredo Mendes Gouveia (1618-1823)

```
        {
        'doc_id': 148086,
        'doc_type': 'REQUERIMENTO',
        'date': '1618-12-07',
        'text': ' 105.  ant. 1618, Dezembro, 7 REQUERIMENTO do contratador de Angola e Cabo
Verde, António Fernandes de Elvas, ao rei D. Filipe II solicitando que lhe fosse passada uma
provisão a propósito dos direitos que se deviam pagar dos escravos que iam para as Índias, o
Brasil e outras partes, tendo em conta as diferenças de entendimento que existiam entre os
feitores da Fazenda Real',
        'sender': {'aff': [], 'title': [], 'names': ['António Fernandes de Elvas'], 'occ':
['contratador']},
        'recipient': {'aff': [], 'title': [], 'names': ['D. Filipe II'], 'occ': ['rei']}
        }
```


Figure 6 Example of a machine-readable document from Collection Angola

All 31 catalogues were converted into a harmonised and machine-readable form following our extensive computational process. Moreover, they are not only digital, but also accessible for further data processing.[18] The extracted data includes details about senders and recipients, their social and political roles, administrative positions, and geographic locations. These data are now amenable to computational methods, such as social network analysis, that can reveal insights into the key players in the early-modern Portuguese empire and their connections to high-ranking officials. Through the skilful application of NLP techniques, our efforts have transcended the realm of simple digitisation and elevated these repositories into the realm of metadata-rich entities. The following section describes the far-reaching possibilities that structured data can offer.

4 Interdisciplinary Strategies for the Structured Data

In the previous sections, we discussed working with large historical datasets and preparing them for machine readability. By doing so, we have successfully developed a functional digital infrastructure for analysing communication patterns. However, the crucial question remains: what comes next? What historical knowledge and digital capabilities do we need to study events, issues, actors, and places in the colonial territories of the Portuguese Empire? How do we situate these interactions within the global framework of the ongoing change in the Portuguese colonies from 1640 to 1822, spanning the period from the Brigantine dynasty to Brazilian independence? This section focuses on presenting methods and ideas that historians, digital humanities scholars, computer scientists, and others can use in similar projects. The digital projects are not intended to create exclusive research methods. Rather, the goal is to develop universal strategies that can be applied to a variety of projects that use a digital approach.

18 To learn more about structured and unstructured documents read Meroño-Peñuela et al. 2014, 539-64.

We have no doubt that conducting digital history projects requires interdisciplinary approaches. This can be observed in several dimensions of *About Remembering Lincoln*,[19] a digital storytelling project; data visualisation exemplified by *Visualizing Emancipation* (Nesbit 2014);[20] mapping techniques as demonstrated in *The Spread of US Slavery 1790-1860*,[21] using the National Historical Geographic Information System (NHGIS) of the Minnesota Population Center;[22] network analysis in *Dissident Networks Project* (Zbíral, Shaw 2022),[23] or in *Mapping the Republic of Letters* (Edelstein et al. 2017).[24] The largest and most complex digital project to date is *Living with Machines*,[25] where their research experience was that it was "one of the largest investments [in the UK] being made in the arts and humanities, and so it is dealing with a team much larger than normally encountered by researchers from these background" (Ahnert et al. 2023, 4). In our project *MAPE – Mapping the Atlantic Portuguese Empire* we also take a collaborative approach that leverages the expertise of a team with backgrounds in history, social sciences, and computer science for the following steps:

1. Content Management – creating a comprehensive thesaurus composed of entries drawn from the sources;
2. Topical Classification – using topic modelling algorithms such as Latent Dirichlet Allocation (LDA);
3. Social Network Analysis – applying Bipartite networks techniques.

4.1 Content Management

As mentioned, our dataset primarily comprised unstructured textual sources, including letters and petitions. This data required a preprocessing phase that involved systematic treatment to extract the relevant information in a structured form. This preprocessing and standardisation facilitated the transformation of textual data into organised data sets suitable for structured storage and management using formats such as relational databases, graph databases,

19 *Remembering Lincoln*. http://rememberinglincoln.fords.org/. Created and maintained by Ford's Theatre, Washington, D.C.

20 https://dsl.richmond.edu/emancipation/.

21 https://lincolnmullen.com/projects/slavery/.

22 http://www.nhgis.org.

23 https://dissinet.cz/.

24 http://republicofletters.stanford.edu/. More examples of digital projects are accessible here: https://infoguides.gmu.edu/digitalhumanities.

25 https://livingwithmachines.ac.uk/.

or tabular representations. Adopting such structured representations holds significant potential and is tied to the characteristics of what one wants to use the data for. For example, graph-based representations natively are more suitable for densely interconnected data, such as communication letters with relations between people, organisations, locations, and other entities.

Structured representations of large datasets produce great results and offer digital historians a wide range of possibilities for exploration. One of these possibilities is the creation of a comprehensive thesaurus composed of entries drawn from the sources. Our research focuses on cataloguing people, geographic locations, and institutions as they are found in these historical documents. However, this task has its complexities and challenges. Creating such a thesaurus requires the judicious application of data mining techniques at multiple levels. These dimensions include removing duplicate entries to ensure the accuracy and integrity of the thesaurus. In addition, individuals, locations, and organisations with identical names and titles must be distinguished, which requires careful disambiguation methods. Matching titles and occupations of individuals is another complicated task, requiring the detection of nuanced variations and contexts. Finally, tracking changes in place names over time is critical to establishing historical context and accurate representation.

Identifying duplicated entities in the corpora can be approached in various ways, from preprocessing the text to employing robust machine learning models. The preprocessing stage involves tasks such as lowercasing, removing punctuation, tokenisation (i.e. breaking down the text into smaller units, known as tokens, that represent meaningful units of text), and possibly reducing words to their base or dictionary form using methods such as stemming or lemmatisation. After the preprocessing, entity normalisation or other methods can be applied to identify duplicated entities. Entity normalisation assumes that entities might be mentioned in different forms or variations (e.g. "João V" or "D. João V"). This technique requires an exhaustive mapping of different variations of an entity to a common representation.

In contrast to entity normalisation, which requires the user to create a dictionary of different variations of the entities, the Thresholds and Similarity method operates more automatically. This method leverages similarity metrics to identify entities that are slightly different but still represent the same concept. For example, string similarity algorithms (e.g. Levenshtein distance, Jaccard similarity) can be applied to find entities with minor variations. Machine learning models also represent an alternative to identifying duplicates, particularly in large datasets or more complex duplication patterns. The model is trained using text features (e.g., term frequencies, entity types, and contextual information) to identify duplicated entities automatically.

Even after applying one or multiple duplicated entity detection techniques, a post-processing step can help achieve better results. This step usually concerns manual merging duplicated entities into a single representation or highlighting them for further review.

In our previous research (Bloch, Vasques, Bojanowski 2022), tackling the identification and subsequent elimination of duplicated entities focused solely on rectifying typographical errors within the transcripts. Presently, we are confronted with the necessity of employing more robust matching algorithms to identify name variations. These variations can be due to several factors, including linguistic shifts inherent to the natural evolution of the language. Addressing this endeavour entails using algorithms based on textual similarity (Bilenko, Mooney 2003). Another matter that requires our attention pertains to the differentiation of individuals who might possess identical names and titles (e.g. a son who inherits his father's name and position). In the literature, we can find various techniques for person name disambiguation based, for example, on clustering (Khabsa, Treeratpituk, Giles 2013) and graphs (Wang et al. 2011). Also, incorporating contextual analysis can help disambiguate entities in such cases. This analysis involves considering the surrounding words or phrases to determine whether the entity is duplicated. In the context of our documents, a straightforward strategy is a disambiguation based on the document date. Besides distinguishing between individuals, we will enrich their information by attributing their titles and occupations. This information, extracted with the NER algorithm, for example, will be obtained from the documents associated with an individual.

Lastly, we need to track name changes of place units (toponymic). Since many names of places (towns, villages) and institutions or organisations are no longer in use in modern Portuguese, we must first determine each name's geographic location and historical development using early-modern dictionaries.[26] All the mentioned steps aim to clean, standardise and structure the data, which is critical to establish a solid foundation to build our data model. A well-designed data model will ensure consistent treatment of entities and relationships in this context leading to a more generalised and scalable framework. For instance, this framework should efficiently allow the inclusion of new document types and the expansion for different time periods, locations, and data sources.

26 Such as, for example, online available dictionaries: *Diccionario da lingua portugueza* by D. Rafael Bluteau; *Diccionario Bibliographico Brasileiro; Diccionario Bibliographico Portuguez*; *Diccionario da lingua brasileira*; *Diccionario topográfico, histórico, descriptivo da comarca do Alto-Amazonas*; or the *Cambridge Encyclopaedia of Latin America and the Caribbean* (Collier, Blakemore, Skidmore 1985).

4.2 Topical Classification

Another way to use structured historical data is thematic classification. Historians' efforts to uncover analogous patterns in correspondence bring the thematic classification to the forefront of digital scholarly interest (Apolinário 2002; 2016; Brauer, Fridlund 2013). A compelling recommendation is topic modelling – a technique used to discover latent themes or topics within a collection of documents. It is an unsupervised learning approach (i.e. it does not require labelled or pre-classified examples) that assumes that every document in the corpus is a mixture of different topics and that a distribution of words characterizes each topic. The goal is to automatically extract these latent topics and determine their word distributions.

In a manner analogous to the data management process, the selection of the method for topical classification varies according to the characteristics of the data and the results to be achieved. For example, depending on the number of documents in the corpora and their average length, some techniques would be more suitable than others. Among the topic modelling algorithms, Latent Dirichlet Allocation (LDA) is one of the most widely used (Blei, Ng, Jordan 2003). LDA assumes that documents are generated based on a probabilistic process involving topic assignments for words. This information enables researchers to interpret and label the discovered topics based on the most representative words. Topic modelling can provide insights into the main themes present in the corpus and allow for exploratory analysis, content organisation, and information retrieval.

In large corpora such as ours, the primary obstacle when utilising LDA topic models lies in the initial parameterisation of the algorithm, explicitly determining the optimal number of topics and customising the stopword list appropriately. Although some techniques give us bounds on the number of topics (e.g. Nikolenko, Koltsov, Koltsova 2017), defining both parameters still depends heavily on a deep knowledge of the corpus and heuristics. Tuning the models and interpreting the results, especially the meaning of the resulting document themes, require contributions from the historian/archivist in the research team.

Brauer and Fridlund postulated that topic modelling algorithms view each document as a "bag of words". This approach allows the algorithms to distil a coherent essence from the words in a document, revealing meaningful thematic clusters (Brauer, Fridlund 2013, 154). Based on our experience, we observed several letter subjects in our colonial corpus. These categories embody different sets of words that encompass a spectrum of information about social, administrative, political, religious, and economic dimensions. In each letter, residents of the Portuguese colonies express their personal memories and expectations as settlers or colonial officials (Boschi 2011; Candido 2011;

Marquez 2022). It is fascinating to observe, however, that in each communication they focus on a particular ensemble of words that embody their intent. Consider, for example, the petitions of enslaved persons, whose linguistic fabric revolves around terms such as slavery, manumission, and freedom. By comparison, widows' letters revolve around key words such as guardianship or inheritance.

Besides the issues with the algorithm parameterisation, the size of each document in our corpora might also impose some challenges to the topic modelling process. The average text length of records in our data is around 360 characters, comparable to the 280 limit of a tweet, for example. In these scenarios, other versions of the LDA, such as the Gibbs Sampling algorithm for the Dirichlet Multinomial Mixture, GSDMM for short, could perform better (Yin, Wang 2014).

In Figure 7, we can see an example of a visualisation of topics identified by the GSDMM **[fig. 7]**. In this case, we used a word cloud to visualise the most significant words in each group. The colour represents different topics, and the word size is proportional to the number of times it appeared in the documents within that topic. Notice the different sets of words in (a) and (b); for example, the term '*Rei*' (i.e. the King) does not appear in (b). It is worth mentioning that this is a preliminary result produced for exemplification purposes.

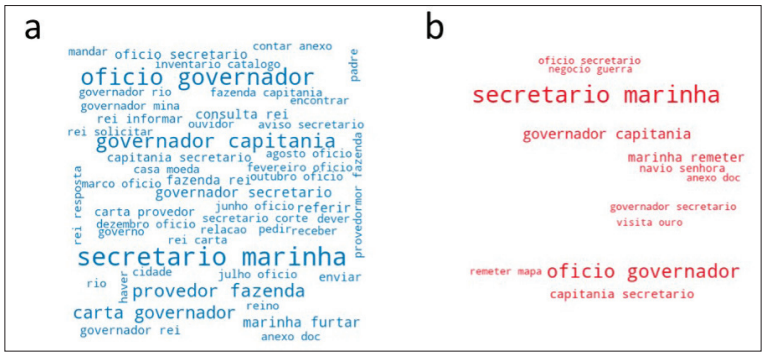


Figure 7 Example of topics identified using GSDMM approach on our corpora

4.3 Social Network Analysis

The structured historical data also provides the opportunity to perform social network analysis. This analytical approach sets to explore relationships and interactions between individuals, groups, organisations, or other social entities. Using this method, historians can gain insight into the patterns of connections, the flow of information, and the dynamics of influence within a given network.

Social networks can be represented as graphs which consist of nodes (representing individuals or entities) and edges (representing relationships or connections between them). Nodes can have various attributes such as demographics, behaviour, or interests, and edges can have different properties like strength, directionality, or type of relationship. Also, one can employ concepts and metrics from graph theory, such as centrality, clustering, connectivity, and community detection.

The use of various analytical techniques such as positional analysis methods (Borgatti, Everett 1992), centrality measurement, block modelling (Doreian, Batagelj, Ferligoj 2005), and Exponential-family Random Graph Models (Handcock et al. 2008; Lusher, Koskinen, Robins 2013) serves as a powerful set of tools to investigate and visualise the complex dynamics and influence of individuals in networks.

Notably, these techniques provide a lens through which one can examine the central role that particular individuals play in the structure of the network. In addition, they are ideal for visualising the fine-grained blocks that emerge in complex networks and are oriented around specific attributes. This ability to visualise the smaller components of networks contributes to a more comprehensive understanding of the architecture of the network.

How can network historians approach their investigations using these techniques? They could examine the dynamic evolution of interactions between colonial subjects and the monarchy over time, especially in terms of their position. Similarly, they could examine the key roles that different officials held within the colonial hierarchy at different levels and the contribution of these roles to shaping the network configuration (positional analysis). Another avenue of inquiry is to identify central figures or roles within the network and determine how their centrality changed over different time periods (centrality measurement). In addition, the researchers could explore whether the network can be divided into subgroups defined by the nature of correspondence and connections between officials and colonial residents and focus on the interplay between these subgroups (block modelling). Finally, network historians could examine the factors that influenced the emergence and evolution of the correspondence network (Exponential-family Random Graph Models).

The networks can provide not only an intuitive way to visualise complex relations but also give insights into the data. Figures 8 and 9 depict two examples of creating these networks. Figure 8 shows an example of a correspondence network where the nodes represent people, and the edge indicates a message exchange between them. In this example, the edges are directed, and the arrow points toward the message's recipient. The node's size is proportional to its degree (i.e. the number of connections), and large nodes represent influential individuals. Figure 8 illustrates the King D. João IV as the most notable hub of connections **[fig. 8]**.

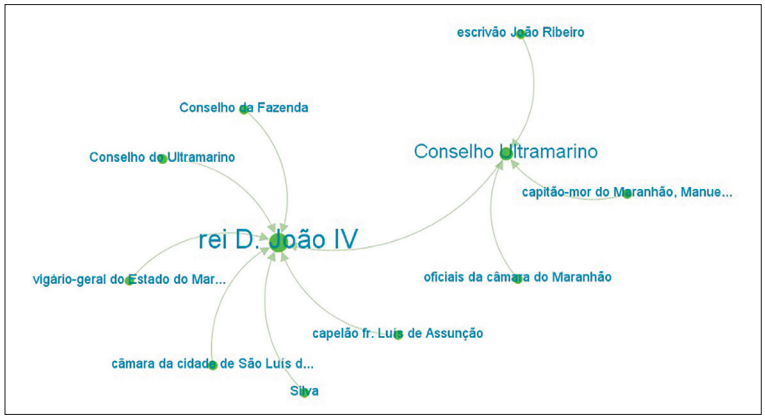


Figure 8 Example of a network connecting senders to recipients of correspondences from our corpus

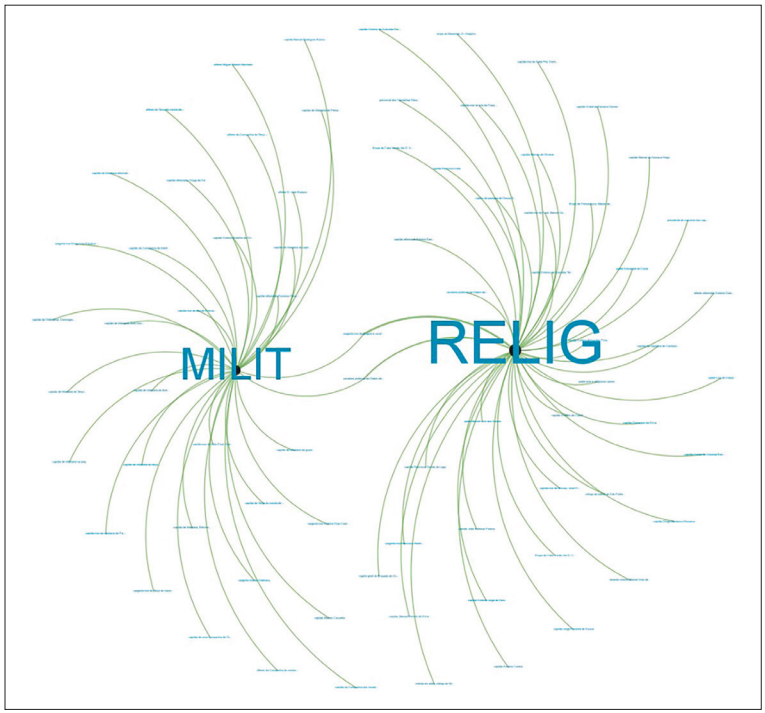


Figure 9 Example of a bipartite network of people and their respective organisation type

Bipartite networks are another way of creating networks **[fig. 9]**. A bipartite network, also known as a bipartite graph or a two-mode network, is a type of graph that consists of two distinct sets of nodes, where nodes in one set are only connected to nodes in the other set and not to nodes within the same set (Curran et al. 2018). In other words, there are no connections between nodes within the same set. These networks can be projected into one-mode networks by creating connections between nodes of the same type that share common neighbours. The projections allow these networks to represent various relationships, such as user-item interactions (e.g. users and products in recommender systems), author-paper relationships (e.g. authors and papers in academic collaboration networks), and more. Figure 9 portrays an example of bipartite networks of people and the type of organisation they are associated with (e.g. military – *MILIT* or religious – *RELIG*). The two nodes in the middle are individuals that occupied a military position but also expressed their association with a religious group or institution. These are perhaps the most interesting nodes in the graph as they express a betweenness, i.e. a brokerage power, that the other individuals in the graph do not possess as they are members only of one organisation type. This simple example demonstrates the potential digital humanities tools have in unveiling hidden structures not otherwise detectable, which have the ability to shed light on certain unexplained phenomena by traditional historiographical tools.

5 Breaking Barriers Between 'Digital' and Historians

In conclusion, we would like to sum up our experience in the realm of digital humanities. First, we emphasise the crucial role of document collection preparation in the initial stages of digitisation, which is essential for all scholars in the digital humanities domain. Improving collaboration between archivists and digital researchers is necessary to facilitate the creation of machine-readable documents. The traditional practice of summarising or transcribing documents and storing them in PDFs or other formats falls short. Research projects in the digital humanities should involve experts from a variety of fields. These include various areas within the humanities, disciplines such as history, archival sciences, and sociology, as well as the dynamic field of computer science.

Second, historians should prioritise the inclusion of metadata when analysing historical documents and carefully consider what types of metadata can be derived from the available sources.

Third, it is necessary to explore and use various software tools from the beginning. Among these tools, Tropy,[27] a free and open-source desktop knowledge organisation application, stands out as user-friendly software that allows describing historical sources, creating metadata, and exporting to JSON-LD formats (Takats 2017). Undoubtedly, JSON, CVS, and XML are the optimal formats to ensure machine readability of historical documents.

Finally, historians need to overcome their reservations about using digital tools and embrace the opportunities they offer. These tools should be viewed as aids that streamline the research process, and certainly not as a substitute for the researcher's expertise and deep understanding of the field. By taking full advantage of existing and rapidly evolving technologies, historians can gain varied interpretations from historical documents.

Bibliography

Abbott, J.M. (2009). *The Angel in the Office*. Durham: British Sociological Association. https://doi.org/10.4324/9780080506999

Ahnert, R.; Ahnert, S.E. (2015). "Protestant Letter Network in the Reign of Mary I. A quantitative approach". *English Literary History*, 82(1), 1-33. https://doi.org/10.1353/elh.2015.0000

Ahnert, R. et al. (2023). *Collaborative Historical Research in the Age of Big Data. Lessons from an Interdisciplinary Project*. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781009175548

Anderson, Ch. (2008). "The End of Theory. The Data Deluge Makes the Scientific Method Obsolete". *Wired magazine* 16(7), 27 June.

Apolinário, J.R. (2002). "Resignificando as Fontes Históricas para o Estudo da Escravidão Negra na região do Tocantins dos séculos XVIII e XIX". *Revista Fontes*, 1(1), 1-117.

Apolinário, J.R. et al. (2016). *Catálogo Geral dos Manuscritos avulsos e em códices referentes à Escravidão Negra no Brasil existentes no Arquivo Histórico Ultramarino*. Campina Grande: EDUEPB.

Ashcroft, B; Griffiths, G.; Tiffin, H. (2002). *The Empire Writes Back: Theory and Practice in Post-colonial Literatures*. London: Routledge. https://doi.org/10.4324/9780203426081

Barnes, T.J. (2013). "Big Data, Little History". *Dialogues in Human Geography*, 3(3), 297-302. https://doi.org/10.1177/2043820613514323

Bertoletti, E.C. et al. (2012). "O projeto resgate de documentação histórica Barão do Rio Branco: acesso às fontes da história do Brasil existentes no exterior". *Clio – Revista de Pesquisa História*, 29(1), 1-26.

Bilenko, M.; Mooney, R.J. (2003). "Adaptive Duplicate Detection Using Learnable String Similarity Measures". *Proceedings of the ninth ACM SIGKDD International Conference*

27 https://tropy.org/.

on Knowledge Discovery and Data Mining. New York: Association for Computing Machinery, 39-48.
https://doi.org/10.1145/956750.956759

Blaney, J. et al. (2021). *Doing Digital History. A Beginner's Guide to Working with Text as Data*. Manchester: Manchester University Press.
https://doi.org/10.7765/9781526157713

Blei, D.M.; Ng, A.Y.; Jordan, M.I. (2003). "Latent Dirichlet Allocation". *Journal of Machine Learning Research*, 3(4-5), 993-1022.

Bloch, A.; Vasques Filho, D.; Bojanowski, M. (2022). "Networks from Archives. Reconstructing Networks of Official Correspondence in the Early Modern Portuguese Empire". *Social Networks*, 69, 123-35.
https://doi.org/10.1016/j.socnet.2020.08.008

Borgatti, S.P.; Everett, M.G. (1992). "Notions of position in social network analysis". *Sociological methodology*, 22, 1-35.
https://doi.org/10.2307/270991

Boschi, C.C. (2011). *O Brasil-Colônia nos arquivos históricos de Portugal.* São Paulo: Alameda.

Boschi, C.C. (2018). "Projeto Resgate: história e arquivística (1982-2014)". *Revista Brasileira de História*, 38(78), 187-208.
https://doi.org/10.1590/1806-93472018v38n78-09

Brauer, R.; Fridlund, M. (2013). "Historicizing Topic Models, a Distant Reading of Topic Modeling Texts Within Historical Studies". Nikiforova, L.V.; Nikiforova, N.V. (eds), *Cultural Research in the Context of "Digital Humanities" = Proceedings of International Conference* (St. Petersburg, 3-5 October 2013). St. Petersburg: Herzen State Pedagogical University & Publishing House Asterion, 152-63.

Brügger, N. (2016). "Digital Humanities". Jensen, K.B. et al. (eds), *The International Encyclopedia of Communication Theory and Philosophy,* 1-8.
https://doi.org/10.1002/9781118766804.wbiect228

Candido, M.P. (2011). "African Freedom Suits and Portuguese Vassal Status: Legal Mechanisms for Fighting Enslavement in Benguela, Angola, 1800-1830". *Slavery & Abolition*, 32(3), 447-59.
https://doi.org/10.1080/0144039X.2011.588481

Collier, S.; Blakemore, H.; Skidmore, T.E. (1985). C*ambridge Encyclopaedia of Latin America and the Caribbean.* Cambridge: Cambridge University Press.

Costa et al. (2014). *História da expansão e do Império Português.* Lisboa: A Esfera dos Livros.

Cuartero, I.A.; Gómes, J.S. (eds) (2012). *Visiones y Revisiones de la Independencia Americana. Subalternidad e Independencia*. Salamanca: Ediciones Universidad de Salamanca.

Curran, B. et al. (2018). "Look Who's Talking. Two-Mode Networks as Representations of a Topic Model of New Zealand Parliamentary Speeches". *PLoS One*, 13(6), e0199072.
https://doi.org/10.1371/journal.pone.0199072

Curto, J.C. (1988). "The Angolan Manuscript Collection of the Arquivo Histórico Ultramarino; Lisbon: Toward a Working Guide". *History in Africa*, 15, 163-89.
https://doi.org/10.2307/3171858

Domingos, N. (2021). *Cultura popular e império: as lutas pela conquista do consumo cultural em Portugal e nas suas colónias*. Lisboa: Imprensa de Ciências Sociais.

Dong, Z. et al. (2021). "Transformation from Human-readable Documents and Archives in Arc Welding Domain to Machine-Interpretable Data". *Computers in Industry*, 128, 103439.
https://doi.org/10.1016/j.compind.2021.103439

Doreian, P.; Batagelj, V.; Ferligoj, A. (2005). *Generalized Blockmodelling*. Cambridge: Cambridge University Press.
https://doi.org/10.1017/CBO9780511584176

Edelstein, D. et al. (2017). "Historical Research in a Digital Age. Reflections from the Mapping the Republic of Letters Project. Historical Research in a Digital Age". *The American Historical Review*, 122(2), 400-24.
https://doi.org/10.1093/ahr/122.2.400

Eijnatten, J. van; Pieters, T.; Verheul, J. (2013). "Big Data for Global History: The Transformative Promise of Digital Humanities". *BMGN – Low Countries Historical Review*, 128(4), 55-77.
https://doi.org/10.18352/bmgn-lchr.9350

Fusaro, M.; Polónia, A. (eds) (2017). *Maritime History as Global History*. Liverpool: Liverpool University Press.
https://doi.org/10.2307/j.ctt21pxjhv

Fragoso, J.; Monteiro, N.G. (2017). *Um reino e suas repúblicas no Atlântico. Comunicações políticas entre Portugal, Brasil e Angola nos séculos XVII e XVIII*. Rio de Janeiro: Civilização Brasileira.

François, P. et al. (2016). "A Macroscope for Global History. Seshat Global History Databank, a Methodological Overview". *Digital Humanities Quarterly*, 10(4), 1-13.

Franzosi, R. (1998). "Narrative as Data. Linguistic and Statistical Tools for the Quantitative Study of Historical Events". *International Review of Social History*, 43(6), 81-104.
https://doi.org/10.1017/S002085900011510X

Gerner, D.J. et al. (2002). "The Creation of CAMEO (Conflict and Mediation Event Observations). An Event Data Framework for a Postcold War World". *The annual meeting of the American Political Science Association*, *29 August-1 September 2002*.
https://parusanalytics.com/eventdata/papers.dir/Gerner.APSA.02.pdf.

Graham, S.; Milligan, I.; Weingart, S. (2015). *Exploring Big Historical Data. The Historian's Macroscope*. Singapore: World Scientific Publishing Company.
https://doi.org/10.1142/p981

Handcock, M.S. et al. (2008). "M. statnet. Software Tools for the Representation, Visualization, Analysis and Simulation of Network Data". *Journal of statistical software,* 24(1).
https://doi.org/10.18637/jss.v024.i01.

Havik, P.J.; Newitt, M. (eds) (2015). *Creole Societies in the Portuguese Colonial Empire*. Newcastle upon Tyne: Cambridge Scholars Publishing.

Hedstrom, M.; Kowlowitz, A. (1988). "Meeting the Challenge of Machine-Readable Records. A State Archives Perspective". *Reference Services Review*, 16(1-2), 31-40.
https://doi.org/10.1108/eb049007

Jaillant, L. et al. (2022). "Introduction: Challenges and Prospects of Born-digital and Digitized Archives in the Digital Humanities". *Archival Science*, 22(3), 285-91.
https://doi.org/10.1007/s10502-022-09396-1

Khabsa, M.; Treeratpituk, P.; Giles, C.L. (2015). "Online Person Name Disambiguation with Constraints". *Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries*. New York: Publication History, Association for Computing Machinery, 37-46.
https://doi.org/10.1145/2756406.2756915

Lusher, D.; Koskinen, J.; Robins, G. (eds) (2013). *Exponential random graph models for social networks: Theory, methods, and applications*. Cambridge: Cambridge University Press.
https://doi.org/10.1017/CBO9780511894701

Marquez, J.C. (2022). "Afflicted Slaves, Faithful Vassals: Sevícias, Manumission, and Enslaved Petitioners in Eighteenth-Century Brazil". *Slavery & Abolition*, 43(1), 91-119.
https://doi.org/10.1080/0144039X.2021.2022963

McLean, P.D. (2007). *The Art of the Network. Strategic Interaction and Patronage in Renaissance Florence*. *Politics, History, and Culture.* Durham: Duke University Press.
https://doi.org/10.2307/j.ctv11g982p

McShane, B.A. (2018). "Visualising the Reception and Circulation of Early Modern Nuns' Letters". *Journal of Historical Network Research*, 2(1), 1-25.

Meroño-Peñuela, A. et al. (2014). "Semantic Technologies for Historical Research: A Survey". *Semantic Web*, 6(6), 539-64.
https://doi.org/10.3233/SW-140158

Nesbit, S. (2014). "Visualizing Emancipation: Mapping the End of Slavery in the American Civil War". Zander, J.; Mosterman, P. (eds), *Computation for Humanity*. Boca Raton: CRC Press, 427-34.

Nicodemo, T.L.; Marino, I.K. (2022). *Por uma história da COVID-19: iniciativas de memória da pandemia no Brasil*. Jardim da Penha, Vitória: Editora Milfontes.
https://doi.org/10.5007/2175-7976.2021.e80966

Nikolenko, S.I.; Koltsov, S.; Koltsova, O. (2017). "Topic modelling for qualitative studies". *Journal of Information Science*, 43(1), 88-102.
https://doi.org/10.1177/0165551515617393

O'Brien, S.P. (2010). "Crisis Early Warning and Decision Support: Contemporary Approaches and Thoughts on Future Research". *International Studies Review*, 12(1), 87-104.
https://doi.org/10.1111/j.1468-2486.2009.00914.x

O'Brien, S.P. (2013). "A Multi-Method Approach for Near Real Time Conflict and Crisis Early Warning". Subrahmanian, V.S. (ed.), *Handbook of Computational Approaches to Counterterrorism*. New York: Springer, 401-18.
https://doi.org/10.1007/978-1-4614-5311-6_18

Padgett, J.F.; Ansell, C.K. (1993). "Robust Action and the Rise of the Medici, 1400-1434". *American Journal of Sociology*, 98(6), 1259-319.
https://doi.org/10.1086/230190

Ramos, R. et al. (2009). *História de Portugal*. Lisboa: A esfera dos livros.

Reinke, H. (1981). "Towards Standards for the Description of Machine-readable Historical Data". *Historical Social Research/Historische Sozialforschung*, 6(2), 3-10.

Salehyan, I. et al. (2012). "Social Conflict in Africa: A New Database". *International Interactions*, 38(4), 503-11.
https://doi.org/10.1080/03050629.2012.697426

Subrahmanyam, S. (2007). "Holding the World in Balance: The Connected Histories of the Iberian Overseas Empires, 1500-1640". *The American Historical Review*, 112(5), 1359-85.
https://doi.org/10.1086/ahr.112.5.1359

Supple, B. (2015). "Preserving the Past: Creating the Endangered Archives Programme". Kominko, M. (ed.), *From Dust to Digital: Ten years of the Endangered Archives Programme.* Cambridge: Open Book Publishers, XXXIX–XLI.
https://doi.org/10.11647/OBP.0052.21

Thornton, J.K. (2012). *A Cultural History of the Atlantic World, 1250-1820.* Cambridge: Cambridge University Press.
https://doi.org/10.1017/CBO9781139021722

Toledo, J.I. et al. (2019). "Information Extraction from Historical Handwritten Document Images with a Context-Aware Neural Model". *Pattern Recognition*, 86, 27-36.

Wang, X. et al. (2011). "Adana: Active Name Disambiguation". *ICDM '11 = Proceedings of the 2011 IEEE 11th International Conference on Data Mining*. Washington, D.C.: IEEE Computer Society, 794-803.
https://doi.org/10.1109/ICDM.2011.19

Xavier, A. et. al. (2018). *Monarquias ibéricas em perspectiva comparada (séculos XVI-XVIII): dinâmicas imperiais e circulação de modelos políticos-administrativos.* Lisboa*:* Imprensa das Ciências Sociais.

Yin, J.; Wang, J. (2014). "A Dirichlet Multinomial Mixture Model-based Approach for Short Text Clustering". Macskassy, S. (ed.), *KDD '14: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* New York: Association for Computing Machinery, 233-42.
https://doi.org/10.1145/2623330.2623715

Zbíral, D.; Shaw, R.L.J. (2022). "Hearing Voices: Reapproaching Medieval Inquisition Records". *Religions*, 13(12), 1175.
https://doi.org/10.3390/rel13121175

Models of Data Extraction and Architecture in Relational Databases of Early Modern Private Political Archives
edited by Dorit Raines

Methods and Tools of Quantification in Historical Research
Napoleonic Employment Applications as a Case Study

Valentina Dal Cin
Università Ca' Foscari Venezia, Italia

Abstract With the advent of Digital Humanities offering innovative tools for historical research, this chapter evaluates their benefits and drawbacks using the Napoleonic Employment Applications project (NapApps) as a case study. After examining data structuring and categorization, it analyses 800 application letters from candidates for the Napoleonic administration to investigate aspects of professionalization. Focusing in particular on willingness to relocate and associated rhetorical strategies, the study identifies specific trends and distinctive vocabulary patterns. Following an evaluation of various text mining techniques along with their strengths and limitations, it explores methods that integrate text analysis with metadata about the authors. It concludes by arguing that this integrated approach provides a robust means for investigating historical phenomena.

Keywords Digital humanities. Quantitative methods. Text mining. Employment applications. Napoleonic Europe.

Summary 1 Introduction. – 2 Research Questions, Sources, and Sample Selection. – 3 Categorization, Quantification, and Correlation Detection. – 4 Text Mining Methods: An Overview. – 4.1 Counting Words. – 4.2 Topic Modelling. – 4.3 Machine Learning Methods. – 5 Combining Text Mining With Metadata Analysis. – 6 Conclusion

1 Introduction

The recent rising number of projects and publications within the field of Digital Humanities has significantly impacted historical research. Contemporary historical publications have extensively addressed

Edizioni Ca'Foscari

Studi di archivistica, bibliografia, paleografia 8
e-ISSN 2610-9093 | ISSN 2610-9875
ISBN [ebook] 978-88-6969-919-1 | ISBN [print] 978-88-6969-920-7

Peer review | Open access
Submitted 2024-07-08 | Accepted 2024-08-28 | Published 2025-05-22

DOI 10.30687/978-88-6969-919-1//007

both theoretical and methodological dimensions, offering researchers detailed guidance on utilizing new digital tools.[1] This contribution presents methods suitable for individual historical research endeavours, utilizing the Napoleonic Employment Applications project (NapApps) as a case study to analyse the advantages and limitations of different methodologies.

Following an introduction to the research questions, archival sources, and criteria for sample selection, the focus transitions to the considerations involved in categorization choices and the methods for identifying correlations between variables. Subsequently, the discussion explores text-mining techniques, emphasizing those that integrate text analysis with metadata essential for prosopographic research.

2 Research Questions, Sources, and Sample Selection

Before discussing any methods and tools, it is essential to introduce the historiographical questions that this research aims to address. The NapApps project focuses on applications for employment submitted by individuals seeking to join the Napoleonic administration at the beginning of the nineteenth century.[2] While the *Declaration of the Rights of Man and of the Citizen* of 1789 opened public employment to all, regardless of social status, it is assumed that the Napoleonic era not only maintained this principle but also consolidated its meritocratic essence.[3]

Although individuals did not enter the State administration through public competition but rather through sovereign appointments, recruitment criteria considered certain professional attributes. For example, the rule that a prefect (i.e., the head of administration in each department) should not be appointed in his place of residence, to avoid conflicts between public and private interests, implied that candidates had to be prepared to relocate. Additionally,

1 On recent methodological reflections and discussion, see Salmi 2021; Crymble 2021; Lässig 2021, 5-34; Guildi 2020, 327-46; Story et al. 2020, 1337-46; Robertson 2016, 289-307. Examples of recent guides are Lemercier, Zalc 2019; Blaney et al. 2021; Corfield, Hitchcock 2022; Graham et al. 2022.

2 The project *Napoleonic Job Applications: from Personal Pleas to Modern Curriculum Vitae in Early 19th-Century Europe* (NapApps) has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No. 101018470.

3 The close connection between meritocracy and the Napoleonic regime has been extensively debated. For further insights cf. Ellis 1997, 50-1; Grab 2003, 21-3, 43, 59; Forrest 2011, 140-1, 202-3. For an in-depth examination of meritocracy spanning the revolutionary period to the Napoleonic era, particularly within the military context, consult Blaufarb 2002.

since the French Empire included non-French speaking departments, knowledge of the local language (i.e., Italian or German) was a significant factor for the government in selecting suitable candidates. However, as the highest local administrative officials were also expected to represent the government properly and lend prestige to their roles, a comfortable economic situation was an additional implicit criterion. This often resulted in the integration of local elite members into the administration to secure their support.[4]

These various elements can be traced within employment applications to determine whether and to what extent candidates understood and internalized these mechanisms and underlying values.[5] The content of the applications and the lexical choices made by the candidates can be related to their characteristics (i.e., age, place of origin, professional background, the employment requested and its geographical location, and the addressee of the application) to understand which features were influential and to what extent. This aspect of the research lends itself to quantitative analysis. Nevertheless, outlining a background of general trends does not exclude – and indeed fosters – the need to focus on specific cases qualitatively to understand in detail the reasons behind the behaviour of the candidates and the choices of the government.

Since the administrative positions under examination are of high and medium-high profiles, the candidates largely belonged to the upper or upper-middle classes of society, situating this research within the broader field of studies on the Napoleonic notables and their interaction with the government, especially in the departments annexed to the French Empire.[6] For this analysis, the focus has been restricted to the departments in the Italian peninsula and those within the *ancienne France*, to allow a comparative study of the application writing styles of Italian and French candidates.

The sources for the quantitative part of the research project are preserved at the Archives nationales de France in a collection covering the period 1789-1878.[7] The collection consists of 178 boxes ar-

4 For prefects within the French Empire, Karila-Cohen 2021 and Whitcomb 1974. Regarding the practice of appointing prefects outside their native departments, Woloch 2001b, 53 and Broers 2016, 350. For studies examining language skills, cf. Woolf 1991, 73; McCain 2018, 206; Lignereux 2019, 61-3, 184-8. Broers (1996, 138-41) explores the question of professionalism, while Lignereux (2019) provides insights into the imperial dimension of careers during the Napoleonic age.

5 Being oriented toward a goal, which is getting hired, the texts of employment applications reveal the criteria their authors considered relevant to reach it (Cohen 2017, 102; Lignereux 2012, 109).

6 On Napoleonic notables, Bergeron, Chaussinand-Nogaret 1979; Woloch 2001a, 67-74; Dunne 2007, 61-78; Levati 2009, 215-28.

7 Archives Nationales de France (AN), F/1dII, *Demandes diverses*, A1; A3; A4; B1; B2; B5; B6; B9; B11; B13; B15; B19; B26; B28; B29; C1; C2; C3; C4; C6; C8; C10; C11; C12;

ranged alphabetically. To target employment applications from the Napoleonic era, the most straightforward approach is to collect them from a random sample of boxes. However, to account for the candidates' place of origin, instead of choosing the boxes randomly, those containing at least one folder with a surname likely of Italian origin dated between 1799 and 1815 were selected. This resulted in a sample of 85 boxes (48% of the total). This choice was made to account for the smaller sample size of candidates from the Italian peninsula compared to those from France and furthermore to ensure the existence of a sufficient number of candidates from the Italian peninsula for meaningful comparisons.[8] Then, within each box, a choice was made to select only the folders of those who applied for the positions of prefect, subprefect, secretary-general of the prefecture, and prefectural counselor, as these are the main positions within the Napoleonic departmental administration.[9] The outcome was a sample of 330 candidates, authors of 800 job applications.[10] Even if the selection process was not random, the guiding principle does not bias the analysis, since – place of origin excluded – all other features of the candidates found in the 85 boxes are random.

C14; D1; D2; D4; D8; D21; F1; F3; F5; F6; G1; G2; G5; G6; G10; H2; I1; L1; L6; L15; L18; M1; M3; M4; M6; M8; M9; M10; M11; M12; M13; N1; N2; O1; P1; P4; P7; P8; P11; R1; R2; R3; R4; R5; R10; S1; S2; S3; S5; S6; S7; S8; T4; T5; T6; V1; V5; V7; Z1.

8 Oversampling of minorities is addressed by Lemercier and Zalc, who recommend selecting samples of similar sizes (2019, 43). In the case presented here, French candidates constitute 59% of the sample (196 individuals), having submitted 456 employment applications (57% of the total), while Italian candidates make up 41% of the sample (134 individuals), with 344 employment applications (43% of the total). Although the sample sizes are not identical, they are sufficiently comparable. These comparisons can be performed using the chi-square statistical test, which will be discussed later. A preliminary version of the sample, consisting of 300 applicants and 690 applications (analysed in Dal Cin 2023, 53-68), has been supplemented with additional transcriptions. The full text of the transcriptions and the spreadsheet underlying the analyses is accessible in the Zenodo open access archive: `https://doi.org/10.5281/zenodo.13686359`.

9 For a general overview of these positions, and their different duties see Godechot 1985, 586-99 and Lentz et al. 2008, 158-9, 521-2, 569, 593-4.

10 As a rule, a sample size of at least 300 items is considered sufficient to obtain reliable results in historical research. Lemercier, Zalc 2019, 39.

3 Categorization, Quantification, and Correlation Detection

Once the sample for the quantitative analysis had been defined, the next step in the research was to extract information about the candidates and their job applications. In order to duly analyse the applications, a sort of a relational database was created with the help of two spreadsheets: the first encompassing the candidates who were assigned a unique ID and the second the job applications marked by the respective candidate ID, and associated with a progressive number for each application starting from 1 (e.g., a candidate with ID 10, would have a job application numbered 10.1). Each job application was then associated with a series of features, including date and place of writing, the requested job, its location (if specified), and the recipient. Other characteristics relate to the author of the application, including age, place of origin, and professional background. Each of these characteristics, called variables, was allocated its own column. They were associated with a column containing the full text of each job application letter.

In the first phase, the information was ingested into the candidates' spreadsheet in a discursive manner, following what was reported in the source. In the second phase, the information was categorized and ingested into the applications' spreadsheet. Categorization is necessary because, as Claire Lemercier and Claire Zalc explain, "quantification is only possible on standardized, as opposed to 'raw' data". They also add that prosopography – the approach adopted in the project – "always begins with categorization" (Lemercier, Zalc 2019, 34). Rather than categorizing the information in the candidates' spreadsheet, the focus was on the applications' spreadsheet as a primary dataset. This choice was justified by the intention to avoid anachronisms, since many features of the candidates could change over time. For example, a candidate's professional background at the time of the first application could be different from that of two or five years later. The same applies to the desired position, which could change over time. Therefore, using employment applications as the focal point of analysis meant contextualizing each action of a historical actor in a specific timeframe.

Though time-consuming, the process of ingesting and categorizing data is an opportunity to rethink how to interrogate the sources and to reexamine research hypotheses (Lemercier, Zalc 2019, 55-6). Consequently, categorization revision occurred several times while attention was paid to avoid losing information. Refining the granularity depended on the nature of analysis to be carried out. For example, a column was added specifying the department in which candidates requested to be employed (if mentioned) and another indicating its geographic area. Regrettably, these columns, like other

columns within the spreadsheet, contain a number of cases of missing data due to the unavailability of certain information in both primary sources and secondary literature (Lemercier, Zalc 2019, 73). Common to historical research, the issue of missing data does not significantly alter the outcomes of quantitative analyses when its occurrence is restricted. This also applies to errors; while a few errors might have a notable impact on a small subset of cases, they typically do not disrupt broader trends calculated from a larger pool of cases.

Hence, for scholars inclined towards a quantitative approach or those seeking to criticize it, a grasp of statistics is indispensable. Lemercier and Zalc argue that "contingency tables and chi-square tests are the most important tools for historians to learn" (2019, 73). Indeed, contingency tables serve as the primary instrument for quantification, as they concurrently present the values of multiple variables (e.g., age class and desired position), thereby revealing any potential correlations. Illustrating this point with an example from the project is elucidating. As previously noted, the inquiry revolves around discerning the presence or absence of professional inclination among candidates who authored a sample of 800 employment applications. One indicative parameter of this inclination is the candidate's readiness to relocate. The archetype of the Weberian-style professional civil servant is presumed to be prepared to discharge duties wherever the State mandates.[11] Conversely, a preference to remain in one's home department suggests a profile akin to that of a local notable, willing to collaborate with the government only under the condition of retaining ties to the milieu from which his social prominence emanates.

The question is particularly intriguing within the Napoleonic era, as Aurélien Lignereux aptly articulates through the concept of the "marché impérial de l'emploi public", offering a lens to examine the impact of border expansion on social behaviour. The imperial expanse brought forth augmented employment prospects for those amenable to relocation, albeit amidst heightened competition (Lignereux 2019, 22-3). In his prosopographical analysis of the *Impériaux* – French nationals employed beyond their national boundaries – Lignereux delineates four applicant categories: those seeking employment in proximity of their home department, those explicitly seeking roles in distant locales, those vying for employment in newly annexed territories following the Empire's expansion, and those exhibiting no preference in a bid to enhance their employment prospects (105-7).

11 The main features of the professional civil servant are described in Weber 1981, 58-66. The link between mobility and professionalism in the Napoleonic administration has been underlined by Michael Broers 1996, 141; 2005, 199-201; 2016, 356.

In contrast, the present analysis adopts a two-tiered approach: the first tier categorizes applications based on whether the candidate specifies a position within his home department, while the second tier differentiates applications for positions within the candidate's local area (including his home department and neighbouring departments) from others. Applications for positions within the candidate's home department account for 19.5% of the total. This percentage gains greater significance when considered alongside other variables. Consequently, the ensuing contingency tables relate the requested department to the place of origin variable, examining disparities between candidates from the *ancienne France* and those from the annexed departments of the Italian peninsula.

Table 1 Requested location in relation to candidates' place of origin (%)

	French departments	**Italian departments**	**Total**
Candidate's own department	10.1	9.4	19.5
Elsewhere or unspecified	46.5	34	80.5
Total	56.6	43.4	100 (N = 777)

Table 2 Requested location in relation to candidates' place of origin (%)*

	French departments	**Italian peninsula**	**Total**
Candidate's own area (home department and neighbouring departments)	18.5	19.7	38.2
Elsewhere or unspecified	38	23.8	61.8
Total	56.5	43.5	100 (N = 782)

* The variables are significantly correlated at the 5% level, as determined by the p-value from a chi-square test.

Only the disparities outlined in Table 2 are statistically significant according to the chi-square test **[tab. 2]**.[12] This test compares observed frequencies with those expected under the null hypothesis,

12 In Table 1, the total is 777 because precise information about the candidate's department was unavailable for 23 out of 800 applications. In Table 2, the total is 782 because, although the department was not explicitly mentioned in some cases, it was evident that five candidates were from Piedmont. Therefore, it was possible to determine their requests for employment within the same area.

assuming independence of the variables. If the difference exceeds a conventional threshold, indicated by a p-value below 0.05, implying a probability of less than 5%, the variables are deemed correlated. This indicates an imbalance in the distribution of values beyond what chance alone would produce, given the sample size. However, correlation does not imply causation, as two correlated variables could both be influenced by a third variable. Therefore, cautious interpretation is imperative. While the test is necessary, it is insufficient on its own. Since no correlation has been detected in Table 1, this indicates that there was no significant disparity between candidates from *ancienne France* and those from the Italian peninsula in terms of requesting a position within their own department **[tab. 1]**. Table 2 refines this observation, demonstrating that a correlation exists when the local area is considered **[tab. 2]**. Italian candidates exhibited a greater propensity to request positions within their own area, with Piedmontese seeking employment in Piedmont, Ligurians in Liguria, and so forth, reflecting the enduring influence of Ancien Régime Italian statal boundaries on societal mindset.

A similar "psychological barrier" existed for the French (Lignereux 2019, 107). Despite French candidates' greater willingness to relocate beyond their neighbouring departments compared to Italian applicants, a significant portion still specified positions within the historical borders of France.[13] Thus, further investigation into the specific geographic preferences of those explicitly citing distant locations is warranted. Among the French applications, 13.4% mentioned positions outside France, with 6.1% specifying departments in Italy, 4.8% indicating departments in Belgium, Germany, and the Netherlands collectively, and 2.2% referencing departments in Spain, Switzerland or the Illyrian Provinces. Conversely, 11.3% of Italian applications mentioned positions outside the Italian peninsula, with 9% specifying French departments, 1.2% mentioning Belgian or German departments, and 1.7% referencing departments in Spain, Switzerland, or in the Illyrian Provinces. The most notable disparity between the two groups is the French candidates' greater propensity to seek employment in Belgian, German, and Dutch departments compared to their Italian counterparts. However, since this difference arises from small percentages, each below 5%, a closer examination of individual applications is necessary. A detailed analysis reveals that French applications specifying positions in such areas were submitted by fifteen candidates. Of these, only eight could be

13 Among the applications submitted by French candidates, 56% mentioning a department outside their area cited one that remained within the traditional borders of France. Conversely, for Italians, the percentage referring to Italian departments outside their area is 53%. This calculation excludes applications where the desired location was unspecified.

considered genuine relocators, while the remainder sought employment in departments where they could leverage their cultural background – particularly their knowledge of German if they came from the eastern regions of France – and family connections. Among these candidates is Felix Barthélemy, an auditor to the Council of State in 1811, who sought employment in the recently annexed Bouches de l'Elbe department, with Hamburg as its chief town, hailing from the Meurthe department in the East of France, drawing on his prior experience in the Left Bank of the Rhine and his knowledge of German.[14] This familiarity with the requested location is mirrored by Etienne de Falaiseau, who sought to become prefect in the Netherlands.[15] A branch of his family had embraced Protestantism centuries prior and settled in the United Provinces after the revocation of the Edict of Nantes. This old connection was revitalized in recent years when Falaiseau and his wife abandoned France to escape the excesses of the revolution and spent some time in the Netherlands as *émigrés* (De Zuylen de Nyevelt 1893, 350-7). The outcomes of this analysis align with observations made by Lignereux concerning the *Impériaux*, albeit his study encompasses not only officials from the prefectural administration. He noted that the second category among the four he identified – applicants explicitly seeking distant locations – comprised a limited number of individuals. Some candidates within this category justified their requests based on prior familiarity with the mentioned location, with many citing family reasons such as having a relative serving or exerting influence there (Lignereux 2019, 105-6). Therefore, the inclination of French candidates to apply for distant positions more frequently than Italian candidates is nuanced by these considerations.

This example highlights the importance of integrating quantitative and qualitative analyses to discern significant disparities through statistical tests. Such tests are vital for identifying correlations warranting attention and guiding qualitative analysis. Hence, deliberating on the results of these tests, alongside highlighting the absolute numbers underlying the percentages, stands as essential research validation practices akin to citing archival sources. Prior to delving further into the relationship between professionalism and relocation by scrutinizing candidates' vocabulary, it is imperative to acquire a comprehensive understanding of text mining techniques best suited to the corpus and research inquiries.

14 Paris, 27 August 1811. To the minister of the Interior (AN, F/1dII/B5, folder Barthélemy). On his life and career refer to Barthélemy 1885, Lignereux 2019, 34-5, 213, and Van der Burg 2021, 126, 136.

15 Paris, 24 July 1810. To the minister of the Interior (AN, F/1dII/D4, folder De Falaiseau).

4 Text Mining Methods: An Overview

In a recent article, Claire Lemercier (2019) underscored the persistent deferral of the intersection between history and quantitative text analysis, noting that "those who count words continue to do so more and more in the absence of historians". Indeed, historians encounter challenges in familiarizing themselves with a variety of methods characterized by non-uniform terminology (e.g., quantitative text analysis, text mining, natural language processing, distant reading, lexicography, etc.). Furthermore, mastering available tools often proves arduous without specialized backgrounds, as many are tailored for computational linguistics or computer science (Sinclair, Rockwell 2016, 288). However, both collective and individual projects have shown interest in these methods, assessing their relevance to historical research.

4.1 Counting Words

These range from basic word clouds and frequency graphs generated using Google Books Ngram to advanced machine learning algorithms. Each method has its own advantages and limitations, making careful selection crucial. As a result, there are often suggestions to explore multiple tools to ensure comprehensive analysis (Sinclair, Rockwell 2016, 288). While Google Books Ngrams and word clouds offer ease of creation and utility for visualizing broad features in presentations, they lack the robustness required to support rigorous scholarly arguments due to inherent biases.[16] More robust are text mining and machine learning methods. The former falls under unsupervised learning, focusing on exploration and discovery, particularly adept at clustering elements based on shared features. The latter, a supervised learning method, facilitates prediction by classifying elements according to researcher-defined properties (Jockers, Underwood 2016, 293; Nanni, Kuemper, Ponzetto 2016, 66-7).

In text mining document similarity can be computed based on textual attributes, such as the relative frequency of most common words. Word count serves as a fundamental statistic in text analysis. However, beyond merely tallying the most frequent words, identifying the most distinctive words – those significantly overrepresented in specific texts or groups defined by metadata – often yields more insightful results. Words that are infrequently used within the corpus may emerge as salient in texts by a particular author or written

16 Biases of Google Books NGrams are discussed in Romein et al. 2020, 304-6 and in Lemercier, Zalc 2019, 146.

in a specific year. Analysing co-occurrences, or the frequency of two terms appearing together, represents a valuable method for further exploration. Voyant, a web-based reading and analysis environment, provides two tools for this purpose. The Correlations tool enables the exploration of term frequency synchronicity, while the Corpus Collocates tool showcases terms frequently appearing alongside a researcher-defined keyword, elucidating contextual associations.[17] The output generated by this tool resembles the display of keyword in context (KWIC), which is a prominent feature of software such as AntConc (Anthony 2005, 729). Although Voyant and AntConc possess notable advantages – freeware, rapidity, and ease of manipulation – they also entail significant drawbacks (Andersen 2022, 138-9). One such limitation, particularly relevant to historical research, is the inability to incorporate metadata. When it is essential to link texts with temporal, authorial, or geographical metadata, manual corpus segmentation becomes necessary. This process, though crucial, is exceedingly time-consuming, especially when dealing with complex metadata structures that encompass numerous variables for analysis.

4.2 Topic Modelling

Another prevalent unsupervised technique is topic modelling. This method, based on an algorithm that lacks semantic understanding of words but calculates the probability of their co-occurrence, identifies various topics within a corpus and presents them through lists of their most distinctive words. While topic modelling can be executed using various tools, it is often associated with MALLET (Machine Learning for Language Toolkit), a Java-based package based on Latent Dirichlet Allocation (LDA).[18] Despite predominantly being employed by literary scholars, historians have also begun to explore topic modelling.[19] However, prior to its application to a corpus, thorough consideration of its numerous pitfalls is imperative. The number of topics must be predetermined, a parameter devoid of definitive rules, necessitating researchers to iterate the process using different numbers until obtaining satisfactory results. Additionally, researchers must compile a list of stopwords – words that frequently occur

17 https://voyant-tools.org/docs/#!/guide/correlations and https://voyant-tools.org/docs/#!/guide/corpuscollocates.

18 Latent Dirichlet allocation (LDA) was developed in the early 2000s by a group of researchers led by David Blei.

19 A survey on the application of topic modelling in historical research is provided by Brauer, Fridlund 2013, 152-63.

but lack substantive meaning (e.g., articles and conjunctions) – to enhance result significance. This occasionally leads researchers to add words with meaning to this list if their excessive frequency compromises output quality, a decision subject to debate. Moreover, topics generated by the algorithm in the form of word lists lack labels, mandating researchers to assign them.[20] This task is challenging as words often are inconsistent, with only a fraction sharing common features. These considerations underscore the significant role of human decisions and interpretation in a seemingly objective technique.

However, this does not discount the utility of Voyant, AntConc, and MALLET for historians. In her study on the transnational history of psychiatry, Eva Andersen integrated AntConc and MALLET with Histogram to devise a potent search tool aiding navigation through her extensive corpus of over 300,000 pages of psychiatric journals, uncovering unnoticed trends and pertinent segments for close reading (Andersen 2022, 131-57). Heidi Hakkarainen and Zuhair Iftikhar utilized topic modelling on a corpus comprising nearly 100 texts from the years 1829 to 1850 to probe the discourse surrounding the concept of humanism (*Humanismus*) in the German-speaking press. Recognizing the pivotal role of historical context in conceptual understanding, they employed topic modelling dynamically, segmenting their corpus into distinct time frames and organizing keywords chronologically. This approach corroborated Reinhart Koselleck's assertion regarding the *Sattelzeit* – a period wherein previously static phenomena became viewed as dynamic processes due to the heightened significance of temporal terms like *Zeit* (time) or *Zukunft* (future) by mid-century. However, Hakkarainen and Iftikhar cautioned that "the output of a topic modelling process is not a result in itself and needs to be studied further for reliable conclusions" (Hakkarainen, Iftikhar 2020, 269, 272). To furnish a bespoke solution for historical research, particularly for studies transitioning from exploratory analysis to hypothesis-testing endeavours, Federico Nanni, Hiram Kumper, and Simone Paolo Ponzetto advocated for a suite of semi-supervised computational methods. These techniques, both knowledge- and data-driven, facilitate fruitful synergy between algorithmic computational power and researcher domain expertise. Specifically, they advocate applying semi-supervised topic modelling algorithms utilized in natural language processing to historical research. These approaches permit the incorporation of metadata associated with texts into topic detection, as demonstrated by Labeled LDA, or allow researchers to manually define a list of relevant words to "guide the topic model in a specific direction", as exemplified by Seeded LDA

20 For a detailed explanation of the mechanics of topic modelling, Graham et al. 2022, 115-54.

(Nanni, Kuemper, Ponzetto 2016, 69-71). Despite showcasing the efficacy of this approach through its application to a corpus of approximately 1,000 legal books from the seventeenth and eighteenth centuries, the adoption of these more intricate variants of topic modelling appears still limited (Nanni, Kuemper, Ponzetto 2016, 73-4).

In summary, the most widely used unsupervised techniques harbour significant heuristic potential when historians aim to explore large corpora that are impractical to read entirely. However, their utility diminishes when applied to relatively small corpora or to address precise research questions. Even Voyant's most useful feature – statistics on word frequency and co-occurrence – may require to be supplemented with other tools if the integration of metadata is indispensable for addressing research inquiries.

4.3 Machine Learning Methods

Before delving into this aspect further in the subsequent section, a brief overview of the possibilities afforded by machine learning methods, particularly supervised learning techniques, is warranted. These algorithms are trained on annotated samples to make predictions. However, their primary limitation lies in their demand for substantial amounts of training data to construct reliable models, alongside significant computational resources. An example is Word2Vec, an algorithm utilizing a neural network model to generate word embeddings. Since its introduction in 2013, it has gained traction among humanists interested in natural language processing. As elucidated by Melvin Wevers and Marijn Koolen, "a Word Embedding Model (WEM) contains semantic and syntactic information" derived from word distribution based on their co-occurrence frequency (Wevers, Koolen 2020, 226). By discerning relationships between words, the model facilitates contextual analysis (embeddings), synonym identification, and semantic change tracking. Consequently, a word embedding model empowers researchers exploring concepts like democracy to search not only for texts explicitly mentioning the keyword but also for texts contextualizing its usage, significantly enhancing information retrieval capabilities. Moreover, word embedding models prove beneficial for conceptual history and historical semantics due to their capacity to trace shifts in meaning, echoing the ideas of Reinhart Koselleck (Wevers, Koolen 2020, 227, 232, 238).

For instance, diachronic word embeddings were employed by Rocco Tripodi, Massimo Warglien, Simon Levis Sullam, and Deborah Paci to scrutinize the chronological evolution of antisemitic language. Through training the algorithm on a corpus comprising resources containing keywords related to Jews published between 1789 and 1914, digitized in the online library of the Bibliothèque Nationale de

France (Gallica), encompassing 54,403 books and 245,188 periodical issues, they traced the "evolution of antisemitic bias in the religious, economic, socio-political, racial, ethic and conspirational domains". This study not only qualitatively confirmed the "chronological development of antisemitic moments identified by historians" but also unveiled "an unexpected peak in adverse bias between 1855 and 1866, in connection with the French Second Empire" (Tripodi et al. 2019, 115-25). Conducted within the framework of the European project ODYCCEUS (Opinion Dynamics and Cultural Conflict in European Spaces), this study underscores a significant aspect of employing machine learning methods: the formidable challenges historians face in independently harnessing these tools, necessitating substantial support from research centres or large-scale collaborative projects.[21]

Nevertheless, despite their potency, word embedding models demand specific conditions seldom met in historical research. As previously noted, one pivotal condition is access to extensive corpora. It has been posited that, for robust representation, Word2Vec necessitates a corpus comprising at least 100 million words per time slice, a vocabulary of approximately one to two million distinct words, and texts featuring frequently co-occurring words. Furthermore, the corpus should incorporate a substantial number of words of interest; otherwise, the analysis output regarding semantic change may prove unreliable (Wevers, Koolen 2020, 233).

Given that the necessity for a large corpus can be mitigated if the corpus is relatively homogeneous, containing texts within a narrow domain, scholars have attempted to apply the algorithm to limited datasets. This was exemplified by Ekaterina Kamlovskaya, who investigated a collection of Indigenous Australian autobiographical narratives. Her initial findings corroborated qualitative research observations on the prevalence of introduced sports due to early Western cultural imposition and underscored the association between sports and words linked to self-esteem, suggesting that word embedding models represent "a promising method", albeit one that requires cautious utilization due to the absence of an optimal combination of parameters, necessitating researcher-driven choices (Kamlovskaya 2022, 93, 102-3, 105).

Unfortunately, in most cases, historians lack digitized text corpora meeting the necessary criteria for processing with this algorithm. Furthermore, the largest available corpora exhibit biases toward certain source types, predominantly comprising parliamentary debates, newspapers, books, and journals, particularly from the nineteenth

21 https://www.odycceus.eu/.

century onward.[22] For preceding centuries and varied collections, digitized primary sources remain scarce, potentially biasing future investigations. Should an increasing number of history students develop an interest in machine learning methods, their focus would naturally gravitate toward the study of accessible digitized sources, potentially exacerbating the divergence between 'traditional historians' and 'digital historians', whose topics of interest – not just methods – could significantly differ. Consequently, the selection of materials for digitization by public and private institutions will wield a profound influence on the trajectory of historical research.

While their adoption in Digital Humanities remains limited, there is burgeoning interest in applying neural network models to predictive tasks such as text reconstruction, authorship attribution, and sentiment analysis. For instance, sentiment analysis was conducted by Tobias Blanke, Michael Bryant, and Mark Hedges employing recurrent neural networks to analyse 1,882 textual transcripts of interviews with Holocaust survivors from the United States Holocaust Memorial Museum. Following the creation of a training corpus using a blend of supervised and unsupervised techniques – including dictionary-based sentiment analysis and recurrent neural networks – they qualitatively and quantitatively assessed the results, determining that the latter technique outperformed significantly. However, given that "one of the major criticisms of neural networks is that they make it difficult to understand why they arrive at conclusions", the authors combined word statistics and algorithms to elucidate the neural network's decision-making process in tagging positive and negative testimonies. This multifaceted approach enabled them to comprehend certain errors, with the neural network being misled by family relations, erroneously associating them with positive memories even when occurring within a negative context (Blanke, Bryant, Hedges 2020, 26, 29-31). Employing predictive techniques on historical data, an approach termed 'Predicting the Past', remains contingent on the availability of extensive datasets. Moreover, while sentiment analysis is increasingly refined, it continues to attract substantial critique (Samuel, Rozzi, Palle 2022).

22 Recent historical research employing text mining techniques has relied on such corpora. See, for example, Buongiorno et al. 2022 and Bunout, Ehrmann, Clavert 2023. While this study does not address these challenges, it is important to note that even when digitized corpora are available, significant issues must be considered, including OCR quality, alternative word spellings, abbreviations, polysemy, changing word usage, idioms, misspellings, and omissions (Oberbichler, Pfanzelter 2023, 136).

5 Combining Text Mining With Metadata Analysis

Combining text mining with metadata analysis presents a middle ground between simplistic approaches like Google Ngram and advanced techniques based on neural networks. Instead of relying on complex algorithms that might seem like a 'black box', historians could benefit from the statistics available in most text mining software. One particularly useful statistical measure is Term Frequency-Inverse Document Frequency (TF-IDF), first introduced by Karen Spärck Jones in 1972 for information retrieval. TF-IDF assesses the importance of a term within a corpus of documents by assigning higher weight to terms that are rare in the corpus yet highly discriminative for specific texts, and lower weight to terms that are common across the corpus and less useful for distinguishing between texts. In essence, a higher TF-IDF score indicates that a term is more distinctive for a certain text compared to all other texts in the corpus. This measure can be computed for individual texts as well as for texts aggregated based on metadata such as author, year, place, subject, etc. (Guldi 2022, 908). In a recent research project, Jo Guldi utilized a statistical measure called Term Frequency-Inverse Period Frequency (TF-IPF) on British parliamentary debates. This measure helped detect the most distinctive words in temporal divisions spanning from twenty years down to a single day. By altering the scale of analysis, Guldi revealed both long-term trends and previously overlooked short-term concerns, suggesting that TF-IPF can mitigate implicit biases and stimulate new research questions about lesser-known events (Guldi 2022, 895-911).[23]

Both TF-IDF and TF-IPF revolve around the concept of distinctiveness. However, distinctiveness can be approached in various ways, including through the comparison of word frequencies. Relative frequency, which measures the proportion of a word's occurrences relative to the total number of words in a text, allows for comparisons across texts of different sizes. Although the idea of counting words and identifying co-occurrences is not new, it remains valuable in historical research. For instance, Jacques Guilhaumou's studies on revolutionary discourse in the 1980s already employed counts and statistics (Guilhaumou 1986, 27-46). More recently, Luca Scholz analysed over 20,000 legal dissertations from seventeenth-century German universities, identifying trends in topics such as marriage, debt, and property law (Scholz 2022, 297-327). Cesare Vetter, Marco Marin, and Elisabetta Gon studied Maximilien Robespierre's vocabulary and rhetoric, measuring the distinctiveness of the most

23 New insights on this topic and general guidance on text mining techniques suitable to historical research are provided in Guldi 2023.

frequent socio-political words across different time periods (Vetter, Marin, Gon 2015, 96-110).

Drawing on the concept of distinctiveness, the Napoleonic Employment Applications project investigated the correlation between candidates' willingness to relocate and their professionalism. Employing an inductive approach grounded in relative frequencies, the study compared the vocabulary of candidates according to their preferred employment locations.[24] The corpus of 800 applications was divided into two distinct sub-corpora based on metadata provided by the applications' spreadsheet, which associated each application's text with specific details such as the desired place of employment. One sub-corpus comprised applications from candidates explicitly seeking positions within their home department, while the other consisted of applications from candidates who did not specify their home department as their desired place of employment. To facilitate the application of text mining methods, all words underwent standardization to contemporary spellings and lemmatization to reduce morphological variations to a singular lemma (Cortelazzo 2013, 302). To identify distinctive words based on relative frequencies, two key operations were conducted. Initially, only words exhibiting significant overuse in each sub-corpus, as determined through a chi-square test performed on relative frequencies, were retained. Consequently, the resulting list comprised words distinctive to each sub-corpus. Subsequently, this list was divided based on whether the highest relative frequency of each word occurred in the first or second sub-corpus. The following bubble chart illustrates the results of this analysis, with the size of the bubbles representing the higher frequency of words and their distinctiveness indicated by colour. The darker the colour, the more distinctive the word is of that sub-corpus **[fig. 1]**.[25]

24 For a deductive approach centred on word lists that define thematic vocabularies, cf. Dal Cin 2023, 62-8.

25 The relative frequencies of words are calculated per 10,000 words, excluding both names of people and places.

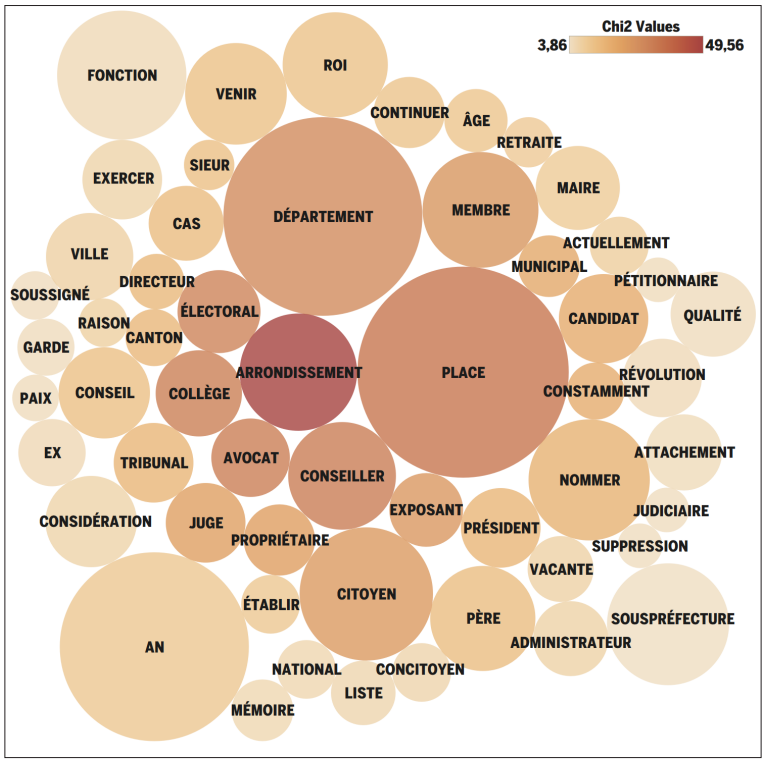


Figure 1 Most distinctive words used by candidates applying for employment within their own department

The larger bubbles indicate that words such as *place*, *département*, and *an* have the highest relative frequency. However, their distinctiveness varies, as indicated by their different colours. To further analyse this set of words, it is beneficial to categorize them according to their respective domains. Many words describe administrative boundaries, such as *arrondissement*, *département*, *canton*, *ville*. Other refer to institutions and positions embedded in specific territorial contexts, such as *collège*, *electoral*, *président* (often referring to the president of electoral colleges), *conseil*, *conseiller*, *municipal*, *maire*, *garde* (referring to the national guard).[26] Some words relate to the candidate's professional situation or background, such as *avocat*, *membre*, *directeur*, *juge*, *judiciaire*, *tribunal*, *candidat* (often referring to the *Corps législatif*), *paix* (referring to the position of *juge de paix*), *sous-préfecture*, *administrateur*, *suppression*, *fonction*, *exercer*.

26 For a description of the role of electoral colleges and other Napoleonic departmental institutions, Lentz et al. 2008, 121-3, 154, 167.

Other words express a temporal dimension, distinguishing between the candidate's past experiences and their present situation. For instance, *an*, *continuer*, *constamment*, *actuellement*, *retraite*, and *ex*. *Propriétaire* and *liste* clearly reference the candidate's wealth and social status, indicating their inclusion on the list of notables or the six hundred largest taxpayers of their department. *Considération* and *concitoyen* suggest the candidate's social standing and the respect they enjoyed as evidence of their notability.

In essence, the words in the list delineate the characteristics of a local notable, highlighting their role as landowners and substantial taxpayers, along with their participation in departmental or municipal institutions. These individuals expressed readiness to engage with the government on the condition that they could uphold their social, familial, and economic connections within their local environment. Further elucidation can be attained by juxtaposing this list of distinctive words with that derived from the other sub-corpus, consisting of applications for positions beyond the candidate's department **[fig. 2]**.

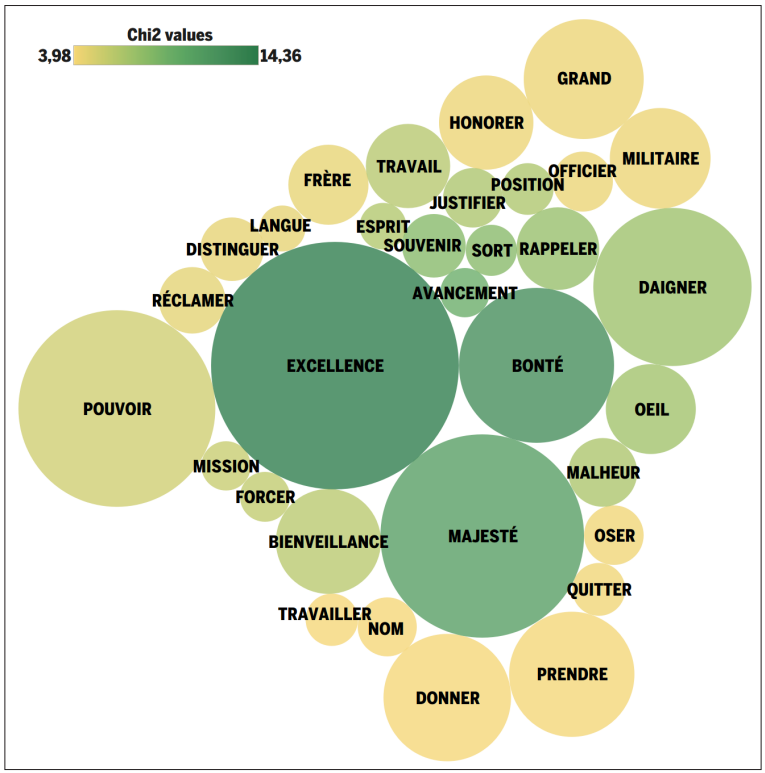


Figure 2 Most distinctive words used by candidates applying for employment outside their own department

In this instance, the word list offers a less distinct portrayal, given the greater heterogeneity of this sub-corpus compared to the previous one. The most frequent words, denoted by larger bubble sizes (*excellence*, *majesté, pouvoir, daigner, bonté, donner, prendre*) offer limited insight into the specific content of these applications. However, one word stands out as particularly distinctive, despite not being among the most frequent: *avancement*. Its absence in the other sub-corpus suggests that no candidate believed they could advance their career within their own department. Other words in the list pertain to duties performed and professional backgrounds, such as *mission*, *travail*, *travailler*, *langue*, *officier*, *militaire*. Yet, there are also numerous words conveying a paternal and benevolent expectation directed towards Napoleon or the Interior Minister, considering the applicant's misfortunes (*bonté*, *bienveillance*, *œil*, *malheur*, *souvenir*, *quitter*, *sort*, *forcer*). Regarding the term *frère*, it refers to either the candidate's brother's role or a recommendation provided by him. Since this term is more frequently used by French applicants, its inclusion in the distinctive words of this sub-corpus is likely due to their slightly higher presence within it (Dal Cin 2023, 63-4).

Within these sub-corpora, two key terms stand out as particularly distinctive: *propriétaire* in the first sub-corpus, and *avancement* in the second. The latter is notable for its reference to career prospects, a significant aspect given the absence of standardized rules for advancement in the French prefectural administration during the nineteenth century (Karila-Cohen 2021, 152-7). The former is significant because it indicates the candidates' intention to emphasize their ties to a specific area, along with their socio-economic status. For example, André-Paul Sain Rousset, the mayor of the Midi district of Lyon, expressed his desire for a position in proximity to the Rhône department, facilitating oversight of his vineyards in Vaux sur Villefranche, which formed a large portion of the estate he was to pass on to his heirs.[27]

The most frequent co-occurrences of *avancement* within a five-word distance include *excellence*, *monseigneur*, *vouloir*, *pouvoir*, *supplier*, *plaire*, *majesté*, *daigner*, *obtenir*, *service*, *administratif*, *place*, *bonté*, *prendre*, *administration*, *carrière*, *solliciter, demander*. The most associated co-occurrences, considering both how often two words co-occur and how often they do not, are *ci-joint*, *administratif*, *carrière*, *oser*, *supplier*, *obtenir*, *solliciter*, *accorder*, *fournir*, *aspirer*, *demander*. Six words appear in both lists: *carrière*, *administratif, supplier, obtenir, solliciter, demander*, indicating the applicant's reference to their career. The most frequent co-occurrences of *propriétaire* within a five-word distance include *département*, *monseigneur*, *an*, *agé*,

27 Lyon, 27 September 1807. To the ministry of the Interior (AN, F/1dII/S2, folder Sain Rousset).

ancien, *général*, *place*, *natif*, *père*, *membre*, *famille*, *excellence*, *plaire*, *fil*, *arrondissement*, *solliciter*, *service*. The most associated co-occurrences are *natif*, *agé*, *département*, *feu*, *père*, *fil*, *commune*, *canton*, *ancien*, *avocat*, *marier*, *arrondissement*. Both lists include terms related to administrative geography (*département*, *arrondissement*, *canton*, *commune*) and family (*père*, *famille*, *fil*, *marier*).[28]

To achieve a comprehensive understanding of the context surrounding both *avancement* and *propriétaire*, it is imperative to analyse the positions sought by applications containing these terms,[29] along with correlations with requested locations and applicants' age classes.[30]

28 Even if they frequently co-occur, names of people and places were disregarded. Hence, TF-IDF was not employed to identify the most distinctive words in the two sub-corpora, as it primarily identified names.

29 Each occurrence was examined to determine its context. Instances where *avancement* referred to Napoleon's rise to power or individuals other than the applicant were excluded from the count. Similarly, instances of *propriétaire* highlighting the opposite of wealthiness were disregarded. Applications containing terms such as *aisance*, *fortune*, *ressource*, *subsistance*, *revenue* (wealth, fortune, resource, subsistence, income), *imposé* (referring to the list of the six hundred largest taxpayers), and *contribuable* (referring to the same context) were added to the count when emphasizing the candidate's wealth.

30 In Tables 3-5, the total is never 800 because a single application could request multiple positions, corresponding to the total number of positions applied for. The number of applications considered is 777, consistent with Table 7. This excludes 23 applications for which the candidate's department is unknown. The total in Table 6 is 539 due to some unknown birthdates. All these contingency tables present values in percentages to facilitate comparisons; however, chi-square tests were performed on row counts. Instead of analysing correlations between pairs of variables, multiple correspondence analysis (MCA), a type of factor analysis, could be employed. MCA is often used in prosopographical research as it provides a general overview of the relationships between all variables, allowing for the detection of correlations without creating numerous contingency tables (Lemercier, Zalc 2019, 88-9, 97). However, MCA was not used in this analysis because the presence of missing data influences the results, as they correlate with each other. Various solutions to this problem exist, but they result in a loss of information.

Table 3 Requested position in relation to mention of the word *avancement* (%)*

	Prefect	Subprefect	Secretary General	Prefectural counselor	Other	Total
Avancement mentioned	2.4 (+)	0.9 (–)	0.7 (=)	0.1 (–)	0.3 (–)	4.5
Not mentioned	19.9 (–)	42.6 (+)	13.9 (=)	5.3 (+)	13.8 (+)	95.5
Total	22.3	43.5	14.6	5.4	14.2	100 (N = 873)

* In this table and the subsequent ones, an asterisk denotes a significant correlation between the variables at the 5% level, as determined by the p-value from a chi-square test. The symbols plus (+) and minus (−) indicate that the observed frequencies are higher or lower than the expected frequencies, respectively. An equal sign (=) signifies no significant difference between observed and expected frequencies. Expected frequencies are calculated under the assumption that the variables are independent. This means that the expected number for each cell is derived from the overall proportions in the table, considering both the row and column totals. If the variables were independent, the observed distribution would match these expected values.

Table 4 Requested position in relation to the mention of the candidate's wealth (%)*

	Prefect	Subprefect	Secretary General	Prefectural counselor	Other	Total
Wealth mentioned	1.8 (–)	7 (+)	1.9 (=)	1 (+)	0.5 (–)	12.3
Not mentioned	20.5 (+)	36.7 (–)	12.7 (=)	4.4 (–)	13.5 (+)	87.7
Total	22.3	43.6	14.7	5.4	14	100 (N = 873)

Table 5 Requested position in relation to requested location (%)*

	Prefect	Subprefect	Secretary General	Prefectural counselor	Other	Total
Candidate's own department	0.8 (–)	9.4 (+)	3.1 (+)	3.2 (+)	1.7 (–)	18.2
Elsewhere or unspecified	21.5 (+)	34.2 (–)	11.6 (–)	2.2 (–)	12.3 (+)	81.8
Total	22.3	43.6	14.7	5.4	14	100 (N = 873)

Table 6 Departments requested by age group (%)*

	20s	30s	40s	50s	60s	Total
Candidate's own department	0.4 (–)	4.8 (–)	3.3 (=)	5.2 (+)	0.6 (=)	14.3
Elsewhere or unspecified	11.1 (+)	37.3 (+)	23.9 (=)	10.9 (–)	2.4 (=)	85.7
Total	11.5	42.1	27.3	16.1	3	100 (N = 539)

Table 7 Department requested in relation to the mention of the candidate's wealth (%)*

	Wealth mentioned	Not mentioned	Total
Candidate's own department	4.4 (+)	15.1 (–)	19.4
Elsewhere or unspecified	7.9 (–)	72.7 (+)	80.6

All contingency tables show a statistically highly significant correlation, with a probability of less than 0.1 percent that the observed distribution of values is due to chance. Table 3 indicates a higher likelihood for applications containing the word *avancement* to be aimed at obtaining the role of prefect. This is attributed not only to its higher position in the hierarchy, prompting subprefects and secretaries general to aspire for promotion, but also to the establishment, beginning in 1810, of various classes of prefectures with differing salary levels (Karila-Cohen 2021, 152-4). Table 4 displays a different trend for applications in which candidates mentioned their wealth, showing a greater inclination towards subprefect and prefectural counselor roles, and less towards the position of prefect. Table 5 complements these findings by illustrating that applications requesting a position within the candidate's department were more likely to target subprefect and prefectural counselor roles, with very few applications for prefect positions specifying the candidate's own department **[tabs 3-5]**. These results align with the differing nature of these positions and governmental policies. As previously mentioned, the government typically refrained from appointing a prefect in their own department to maintain independence from local influences. In contrast, prefectural counselors were often chosen from local notables, offering a position with a low salary and limited career prospects but the advantages of avoiding relocation and gaining recognition under the new regime (Tulard, Tulard 2014, 97-126). Subprefects could be selected from either the local elite or career civil servants (Thoral 2010, 65-6), with the latter group becoming more prominent after 1809 when a decree reserved a quarter of vacant subprefectures for auditors at the Council of State (Durand 1958, 24-5). Table 6 shows that candidates in their twenties and thirties were more inclined to

relocate, while those in their fifties were more likely to seek positions within their own department to enhance their social prestige. Table 7 corroborates previous findings, indicating a greater tendency for applications mentioning the applicant's wealth to specify their own department, whereas those willing to relocate exhibited a limited inclination to do so **[tabs 6-7]**.[31]

A couple of examples illustrate the duality underscored by this quantitative analysis. Charles-Claude Rambaud Brosse, a 53-year-old vice-mayor of the Midi district of Lyon, applied for the position of prefectural counsellor in his own department, Rhône. His application was grounded in his administrative involvement in Lyon, membership in the department's electoral college, and status as a landowner in both the city and the countryside.[32] Conversely, Etienne Charles Garnier, a 32-year-old who had previously served as a bureau chief at the Seine prefecture before becoming a secretary general in the Swiss department of Leman from 1801 to 1809, sought advancement by applying for vacant prefectural or similar administrative positions in Italy (Laharie, Lamoussière 1998, 336). He cited family hardships due to living expenses and climate, but also highlighted his past military service and administrative experience.[33] Garnier explained that he applied for a different role because his current tasks offered no possibility for advancement.[34] His case, coupled with the presence of the word *malheur* in the list of distinctive terms within applications for positions outside the candidate's department, illustrates the potential coexistence of career considerations and personal hardships within the same texts. Contrary to intuition, the inclusion of vocabulary related to 'sufferings' does not necessarily imply the absence of a professional mindset.

Since the subsets of candidates utilizing the vocabulary of wealth and the term *avancement* were selected to illustrate those seeking positions within their own department and those who were not, the correlations shown were expected **[tabs 3-5, 7]**. However, this quantitative analysis reveals additional trends. The position of subprefect emerges as the most sought-after, attracting 43.6% of all applications.

31 The tendency is underlined by the result of the chi-square test, showing which percentages are higher than expected.

32 Lyon, 17 March 1804. To the minister of the Interior (AN, F/1dII/R1, folder Rambaud Brosse).

33 Genève, 24 March 1806. To the minister of the Interior (AN, F/1dII/G2, folder Garnier).

34 "Ses fonctions ne lui fournissent pas dans les attributions qui y sont attachées les moyens de mériter par son travail de l'avancement dans la carrière". Cf. 15 October 1805 (AN, F/1dII/G2, folder Garnier). This letter addressed to the minister of the Interior is written by the candidate's wife. Therefore, it is not included in the sample analysed quantitatively.

Given that candidates rarely used the term *avancement* in this context, it suggests that this role was perceived as an entry-level, not requiring prior administrative experience. For example, Armand-Joseph-Louis Randon Saint-Marcel, a 25-year-old from Isère, applied in 1805 to become subprefect in Vouziers, Ardennes department, citing only his family connection with Napoleon's *capitaine des chasses* and his fervent desire to serve the emperor, following the example of his relatives.[35] Similarly, Claude-François Groshenry d'Emagny, a 26-year-old rentier from Besançon, Doubs department, applied in 1810 for the position of subprefect of Bergerac, Dordogne department, mentioning only his and his brother's military service.[36] Despite the increased prominence given to auditors in subprefect appointments after 1809, aimed at enhancing professionalization, the position continued to attract a variety of applicants from outside the administration.[37] Therefore, it would be beneficial to further investigate the dual nature of this role, sometimes entrusted to local notables and sometimes to career officials, but this analysis is beyond the scope of this study.

In addition to illuminating the heterogeneous profiles of candidates vying for the role of subprefect, the amalgamation of quantitative analysis of prosopographic data and vocabulary usage effectively delineated the profile of local notables inclined to serve the government exclusively within their own department. The distinctive terms they employed primarily related to local administrative positions, representative roles, and their status as landowners. The findings of the contingency tables resonate with analyses conducted on all French notables, indicating that a majority of them fell between the ages of 40 and 60, with 24.55% classified as property owners, 18.12% as local administrators, and 15.76% as civil servants by 1810.[38] Notably, this last percentage reveals an overlap between the profiles of notables and civil servants. However, while a prefect was likely to be a member of his department's electoral college and affluent enough to possess landed properties, a notable did not necessarily

35 3 January 1805. To Napoleon (AN, F/1dII/R1, folder Randon Saint Marcel).

36 Paris, 8 and 11 October 1810. To the minister of the Interior (AN, F/1dII/G10, folder Groshenry d'Emagny).

37 After 1809, applications for the role of subprefect dropped from 35% to 15%. However, this decline is part of a broader trend observed across the entire sample, indicating no correlation with the 1809 decree.

38 These percentages refer to the approximately seventy thousand members of electoral colleges analysed in Bergeron, Chaussinand-Nogaret 1979, 14, 43. Piedmontese and Ligurian members of departmental electoral colleges mostly belonged to the same age class. Cf. Violardo 1995, 85; Beaurepaire-Hernandez 2014, 350-1.

aspire to a career in officialdom.[39] This elucidates why no application for a position within the candidate's own department includes the word *avancement*. This finding is consistent with Lignereux's analysis of the careers of the *Impériaux*, who considered appointments outside France as promotions in 40% of cases. It suggests that candidates rightly perceived career advancement as contingent upon seeking positions beyond their department.[40] Even if this underscores the internalization of certain professional behaviours, it is worth noting that, when mentioned, even departments outside national borders were typically not entirely unfamiliar to applicants requesting them. Furthermore, a rhetorical appeal grounded in hardships was utilized to pursue career progression, alongside references to prior services rendered. The convergence of these factors emphasizes that Napoleonic applications served as a nexus between Ancien Régime pleas and contemporary employment submissions, yet the transition to the latter was far from complete.

6 Conclusion

The various text analysis methods delineated in this study each harbour distinct strengths and weaknesses, rendering them apt for addressing particular inquiries while less suitable for others. Topic modelling and word embedding algorithms, for instance, excel in navigating vast corpora typically encountered in large-scale collaborative ventures. Conversely, statistical analyses anchored in the notion of distinctiveness can yield robust findings even when applied to smaller corpora, a common scenario in individual projects. Moreover, statistical analysis readily accommodates metadata.

Regardless of their level of complexity, the choice between these methods ultimately hinges on a decision between an inductive and a deductive approach. Rather than relying on exploratory techniques, the NapApps project adopted an analytical stance to scrutinize employment applications from the Napoleonic era, specifically probing candidates' inclination to relocate as an indicator of professional behaviour. This inquiry entailed examining correlations among several variables pertinent to both candidates and applications, such as age,

39 "An elite, a nobility cannot be conceived at the beginning of this century without landed property" (Tulard 1975, 222). Members of the departmental electoral colleges were chosen among the six hundred largest taxpayers. However, those to be appointed were not necessary the wealthiest, since their position, social background, and reputation were also considered (Tulard 1975, 224-5).

40 Although this percentage was lower for personnel of the prefectures (1/3 of first appointments), it would increase if identical positions in better-paid, higher-class prefectures were also considered promotions (Lignereux 2019, 199-202).

department of origin, position sought, and desired department. Further analysis of the latter variable involved comparing the most distinctive words used by candidates applying for positions within their own department versus those who did not. Rather than prioritizing visualization, which, despite its strengths in Digital Humanities, can sometimes be misleading,[41] the focus was on statistical analysis, complemented by an examination of individual cases.

Subjected to the critique that targeted quantitative history from the 1960s to the 1980s and its subsequent decline in later decades, statistics should nonetheless be regarded as an asset by historians, particularly within the realm of digital projects. This should be embraced without apprehension of diluting the essence of the historian's craft.[42] As François Furet noted in a 1982 article reflecting on the distinction between narrative history and problem-oriented history, statistical analysis facilitates description rather than interpretation and explanation. Even when history is approached through a problem-oriented lens, historians are tasked with interpreting the mechanisms "through which a probable pattern of collective behaviour – the very one revealed by data analysis – is manifested in individual behaviour during a given period" (Furet 1984, 93-4, 99). This necessitates a dialogue between quantitative and qualitative methods, macro-analysis, and micro-analysis. Numerous studies have showcased the fruitful interaction between these two methods of analysis and scales of observation (Karila-Cohen et al. 2018, 782-3).[43] It is precisely through the integration of general phenomena and individual trajectories that historians can underscore the relevance of their work in the era of 'big data'.

41 On visualization as a narrative technique, Hullman, Diakopoulos 2011, 2231-40.

42 Handbooks tailored for historians seeking to acquaint themselves with statistics encompass works such as those by Haskins, Jeffrey 2011, as well as Hudson, Ishizu 2016. For insights into the trajectory of quantitative history in the United States, Ruggles 2021, 1-25. To explore historical perspectives on criticism and evolving methodologies in quantification, particularly in Europe and France, consult Lemercier, Zalc 2013, 135-64.

43 Examples are present in the same issue of the *Annales*.

Bibliography

Andersen, E. (2022). "From Search to Digital Search. An Exploration Through the Transnational History of Psychiatry". Fickers, Tatarinov 2022, 131-57. https://doi.org/10.1515/9783110723991-007

Anthony, L. (2005). "AntConc: Design and Development of a Freeware Corpus Analysis Toolkit for the Technical Writing Classroom". *Proceedings. IEEE International Professional Communication Conference* (Limerick, 2005), 729-37. https://doi.org/10.1109/ipcc.2005.1494244

Barthélemy, H.-C.-F. (1885). *Souvenirs d'un ancien préfet (1787-1848)*. Paris: E. Dentu.

Beaurepaire-Hernandez, A. (2014). "Un modèle de notable européen? Les 'masses de Granit' des départements liguriens et leur intégration au système impérial". Antoine, F. et al. (éds), *L'Empire napoléonien. Une expérience européenne?*. Paris: Armand Colin, 347-58.

Bergeron, L.; Chaussinand-Nogaret, G. (1979). *Les 'masses de granit'. Cent mille notables du Premier Empire*. Paris: École des hautes études en sciences sociales.

Blaney, J. et al. (2021). *Doing Digital History: A Beginner's Guide to Working with Text as Data*. Manchester: Manchester University Press. https://doi.org/10.7765/9781526157713

Blanke, T.; Bryant, M.; Hedges, M. (2020). "Understanding Memories of the Holocaust. A New Approach to Neural Networks in the Digital Humanities". *Digital Scholarship in the Humanities*, 35(1), 17-33. https://doi.org/10.1093/llc/fqy082

Blaufarb, R. (2002). *The French Army (1750-1820). Careers, Talent, Merit*. Manchester; New York: Manchester University Press.

Brauer, R.; Fridlund, M. (2013). "Historizing Topic Models: A Distant Reading of Topic Modeling Texts Within Historical Studies". Nikiforova, L.V.; Nikiforova, N.V. (eds), *Cultural Research in the Context of "Digital Humanities" = Proceedings of International Conference* (St. Petersburg, 3-5 October 2013). St. Petersburg: Herzen State Pedagogical University & Publishing House Asterion, 152-63.

Broers, M. (1996). *Europe Under Napoleon 1799-1815*. London; New York: Arnold.

Broers, M. (2005). *The Napoleonic Empire in Italy, 1796-1814. Cultural Imperialism in a European Context?*. New York: Palgrave Macmillan.

Broers, M. (2016). "'Les Enfants du Siècle': An Empire of Young Professionals and the Creation of a Bureaucratic, Imperial Ethos in Napoleonic Europe". Crooks, P.; Parsons, T.H. (eds), *Empires and Bureaucracy in World History: From Late Antiquity to the Twentieth Century*. Cambridge: Cambridge University Press, 344-63. http://dx.doi.org/10.1017/cbo9781316694312.015

Bunout, E.; Ehrmann, M.; Clavert, F. (eds) (2023). *Digitised Newspapers. A New Eldorado for Historians? Reflections on Tools, Methods and Epistemology*. Berlin; Boston: De Gruyter. https://doi.org/10.1515/9783110729214

Buongiorno, S. et al. (2022). *The Hansard 19th-Century British Parliamentary Debates with Improved Speaker Names: Parsed Debates, N-Gram Counts, Special Vocabulary, Collocates, and Topics*. Harvard Dataverse, V2. http://doi.org/10.7910/DVN/ZCYJH8

Cohen, D. (2017). "Commis et fonctionnaires, entre service du public et droits de l'individu, de 1792 à l'an IV". *Annales historiques de la Révolution française*, 3, 101-17.

Corfield, P.J.; Hitchcock, T. (2022). "Using Technology Creatively: Digital history". Corfield, P.J.; Hitchcock, T. (eds), *Becoming a Historian. An Informal Guide*. London: University of London Press; Institute of Historical Research, 93-102.

Cortelazzo, M.A. (2013). "Metodi qualitativi e quantitativi di analisi dei testi". *Contemporanea*, 16(2), 299-310.

Crymble, A. (2021). *Technology and the Historian: Transformations in the Digital Age*. Urbana: University of Illinois Press.
https://doi.org/10.5406/j.ctv1k03s73

Dal Cin, V. (2023). "Candidarsi a un impiego in età napoleonica. Riflessioni a partire da una ricerca in corso". *Passato e presente*, 119, 53-68.

De Zuylen de Nyevelt, S.I. (1893). "The Exile of the Marquise the Falaiseau". *The Living Age*, 198(3), 350-7.

Dunne, J. (2007). "Power on the Periphery: Elite-State Relations in the Napoleonic Empire". Dwyer, P.G.; Forrest, A. (eds), *Napoleon and His Empire. Europe, 1804-1814*. Houndmills; Basingstoke; New York: Palgrave Macmillan, 61-78.

Durand, C. (1958). *Les auditeurs au Conseil d'état de 1803 à 1814*. Aix-en-Provence: La Pensée Universitaire.

Ellis, G. (1997). *Napoleon*. London; New York: Longman.

Fickers, A.; Tatarinov, J. (eds) (2022). *Digital History and Hermeneutics. Between Theory and Practice*. Berlin; Boston: De Gruyter.

Forrest, A. (2011). *Napoleon*. London: Quercus.

Furet, F. (1984). "From Narrative History to Problem-oriented History". Furet, F. (ed.), *In the Workshop of History*. Chicago; London: The University of Chicago Press, 54-67. Transl. of: *L'atelier de l'histoire*. Paris: Flammarion.

Godechot, J. [1951] (1985). *Les institutions de la France sous la Révolution et l'Empire*. 3rd ed. Paris: Presses Universitaires de France.

Grab, A. (2003). *Napoleon and the Transformation of Europe*. New York: Palgrave Macmillan.

Graham, S. et al. (2022). "Topic Modeling: A Hands-On Adventure in Big Data". Graham, S. et al. (eds), *Exploring Big Historical Data: the Historian's Macroscope*. 2nd ed. Hackensack (NJ): World Scientific, 115-54.
https://doi.org/10.1142/9789811243042_0004

Graham, S. et al. (2022). *Exploring Big Historical Data: the Historian's Macroscope*. 2nd ed. Hackensack (NJ): World Scientific.

Guilhaumou, J. (1986). "L'historien du discours et la lexicométrie. Étude d'une série chronologique: le 'Père Duchesne' d'Hébert (Juillet 1793-Mars 1794)". *Histoire & Mesure*, 1(3-4), 27-46.
https://doi.org/10.3406/hism.1986.1529

Guildi, J. (2020). "The Common Landscape of Digital History: Universal Methods, Global Borderlands, Longue-Durée History, and Critical Thinking about Approaches and Institutions". Fridlund, M.; Oiva, M.; Paju, P. (eds), *Digital Histories. Emergent Approaches Within the New Digital History*. Helsinki: Helsinki University Press, 327-46.
https://doi.org/10.33134/HUP-5-18

Guldi, J. (2022). "The Algorithm. Mapping Long-Term Trends and Short-Term Change at Multiple Scales of Time". *The American Historical Review*, 127(2), 895-911.
https://doi.org/10.1093/ahr/rhac160

Guldi, J. (2023). *The Dangerous Art of Text Mining*. Cambridge: Cambridge University Press.

Hakkarainen, H.; Iftikhar, Z. (2020). "The Many Themes of Humanism: Topic Modelling Humanism Discourse in Early 19th-Century German-Language Press". Fridlund, M.; Oiva, M.; Paju, P. (eds), *Digital Histories. Emergent Approaches within the New Digital History*. Helsinki: Helsinki University Press, 259-77.
https://doi.org/10.2307/j.ctv1c9hpt8.20

Haskins, L.; Jeffrey, K. [1990] (2011). *Understanding Quantitative History*. Eugene: Resource Publications.

Hudson, P.; Ishizu, M. (2016). *History by Numbers: An Introduction to Quantitative Approaches*. London: Bloomsbury Publishing.

Hullman, J.; Diakopoulos, N. (2011). "Visualization Rhetoric: Framing Effects in Narrative Visualization". *IEEE Transactions on Visualization and Computer Graphics*, 17(12), 2231-40.
https://doi.org/10.1109/TVCG.2011.255

Jockers, M.L.; Underwood, T. (2016). "Text-Mining the Humanities". Schreibman, S.; Siemens, R.; Unsworth, J. (eds), *A New Companion to Digital Humanities*. Malden: Wiley Blackwell, 291-306.
https://doi.org/10.1002/9781118680605.ch20

Kamlovskaya, E. (2022). "Exploring a Corpus of Indigenous Australian Autobiographical Works With Word Embedding Modeling". Fickers, Tatarinov 2022, 87-108.

Karila-Cohen, K. et al. (2018). "Nouvelles cuisines de l'histoire quantitative". *Annales. Histoire, Sciences sociales*, 73(4), 773-83.

Karila-Cohen, P. (2021). *Monsieur le Préfet. Incarner l'État dans la France du XIXe siècle*. Ceyzérieu: Champ Vallon.

Laharie, P.; Lamoussière, C. (1998). *Le Personnel de l'administration préfectorale, 1800-1880. Répertoires nominatif et territorial*. Paris: Centre historique des Archives Nationales.

Lässig, S. (2021). "Digital History. Challenges and Opportunities for the Profession". *Geschichte und Gesellschaft*, 47(1), 5-34.
https://doi.org/10.13109/gege.2021.47.1.5

Lemercier, C. (2019). "L'analisi testuale". Paci, D. (a cura di), *La storia in digitale. Teorie e metodologie*. Milano: Unicopli, 293-4.

Lemercier, C.; Zalc, C. (2013). "Le sens de la mesure: nouveaux usages de la quantification". Granger, C. (éd.), *À quoi pensent les historiens? Faire de l'histoire au XXIe siècle*. Paris: Autrement, 135-64.

Lemercier, C.; Zalc, C. (2019). *Quantitative Methods in the Humanities: an Introduction*. Charlottesville: University of Virginia Press.
https://doi.org/10.2307/j.ctvbqs963.6

Lentz, T. et al. (2008). *Quand Napoléon inventait la France: dictionnaire des institutions politiques, administratives et de cour du Consulat et de l'Empire*. Paris: Tallandier.

Levati, S. (2009). "Les notables napoléoniens: du cas français à celui italien". *Rives méditerranéennes*, 32-3, 215-28.
https://doi.org/10.4000/rives.2969

Lignereux, A. (2012). *Servir Napoléon. Policiers et gendarmes dans les départements annexés (1796-1814)*. Seyssel: Champ Vallon.

Lignereux, A. (2019). *Les Impériaux. Administrer et habiter l'Europe de Napoléon*. Paris: Fayard.

McCain, S. (2018). *The Language Question under Napoleon, 1799-1814*. Cham: Palgrave Macmillan.

Moretti, F. (2013). *Distant Reading*. London; New York: Verso.

Moretti, F. (2022). *Falso movimento: la svolta quantitativa nello studio della letteratura*. Milano: Nottetempo.

Nanni, F.; Kuemper, H.; Ponzetto, S.P. (2016). "Semi-Supervised Textual Analysis and Historical Research Helping Each Other: Some Thoughts and Observations". *International Journal of Humanities and Arts Computing*, 10(1), 63-77.
https://doi.org/10.3366/ijhac.2016.0160

Oberbichler, S.; Pfanzelter, E. (2023). "Tracing Discourses in Digital Newspaper Collections. A Contribution to Digital Hermeneutics while Investigating 'Return Migration' in Historical Press Coverage". Bunout, Ehrmann, Clavert 2023, 126-52.
https://doi.org/10.1515/9783110729214-007

Robertson, S. (2016). "The Differences Between Digital Humanities and Digital History". Gold, M.K.; Klein, L.F. (eds), *Debates in the Digital Humanities 2016*. Minneapolis: University of Minnesota Press, 289-307.

Romein, C.A. et al. (2020). "State of the Field: Digital History". *History*, 105(365), 291-312. https://doi.org/10.1111/1468-229X.12969

Ruggles, S. (2021). "The Revival of Quantification: Reflections on Old New Histories". *Social Science History*, 45, 1-25. https://doi.org/10.1017/ssh.2020.44

Salmi, H. (2021). *What is Digital History?* Cambridge: Polity Press.

Samuel, J.; Rozzi, G.; Palle, R. (2022). "The Dark Side of Sentiment Analysis: An Exploratory Review Using Lexicons, Dictionaries, and a Statistical Monkey and Chimp". http://dx.doi.org/10.2139/ssrn.4000087

Scholz, L. (2022). "A Distant Reading of Legal Dissertations from German Universities in the Seventeenth Century". *The Historical Journal*, 65, 297-327. https://doi.org/10.1017/S0018246X2100011X

Sinclair, S.; Rockwell, G. (2016). "Text Analysis and Visualization: Making Meaning Count". Schreibman, S.; Siemens, R.; Unsworth, J. (eds), *A New Companion to Digital Humanities*. Malden: Wiley Blackwell, 274-90. https://doi.org/10.1002/9781118680605.ch19

Story, D.J. et al. (2020). "History's Future in the Age of the Internet". *The American Historical Review*, 125(4), 1337-46. https://doi.org/10.1093/ahr/rhaa477

Thoral, M.C. (2010). *L'émergence du pouvoir local: le département de l'Isère face à la centralisation napoléonienne (1800-1837)*. Rennes: Presses Universitaires de Rennes.

Tripodi, R. et al. (2019)."Tracing Antisemitic Language Through Diachronic Embedding Projections: France 1789-1914". *Proceedings of the 1st International Workshop on Computational Approaches to Historical* Language *Change*. Firenze: Association for Computational Linguistics, 115-25. https://doi.org/10.18653/v1/W19-4715

Tulard, J. (1975). "Les notables impériaux". Chaussinand-Nogaret, G. (ed.), *Une histoire des élites, 1700-1848*. Paris; La Haye: Mouton éditeur, 218-34.

Tulard, J.; Tulard, M.-J. (2014). *Napoléon et 40 millions de sujets. La centralisation et le Premier Empire*. Paris: Editions Tallandier.

Van der Burg, M. (2021). *Napoleonic Governance in the Netherlands and NorthWest Germany: Conquest, Incorporation, and Integration*. Cham: Palgrave Macmillan. https://doi.org/10.1007/978-3-030-66658-3

Vetter, C.; Marin, M.; Gon, E. (2015). *Dictionnaire Robespierre. Lexicométrie et usages langagiers. Outils pour une histoire du lexique de l'Incorruptible*, t. 1. Trieste: Edizioni Università di Trieste.

Violardo, M. (1995). *Il notabilato piemontese da Napoleone a Carlo Alberto*. Torino: Comitato di Torino dell'Istituto per la storia del Risorgimento italiano.

Weber, M. (1981). *Economia e società*. Vol. 4, *Sociologia politica*. Milano: Edizioni di Comunità. Transl. of: *Wirtschaft und Gesellschaft*. Tübingen: Mohr.

Wevers, M.; Koolen, M. (2020). "Digital Begriffsgeschichte: Tracing Semantic Change Using Word Embeddings". *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 53(4), 226-43. https://doi.org/10.1080/01615440.2020.1760157

Whitcomb, E.A. (1974). "Napoleon's Prefects". *The American Historical Review*, 79, 1089-118. https://doi.org/10.2307/1869564

Woloch, I. (2001a). "The Napoleonic Regime and French Society". Dwyer, P.G. (ed.), *Napoleon and Europe*. New York: Longman, 60-78.

Woloch, I. (2001b). *Napoleon and His Collaborators. The Making of a Dictatorship*. New York; London: Norton & Company.
Woolf, S. (1991). *Napoleon's Integration of Europe*. London; New York: Routledge.

Models of Data Extraction and Architecture in Relational Databases of Early Modern Private Political Archives
edited by Dorit Raines

Gendered Data in Medieval and Early Modern Sources
The *Gendered Networks* and *Digital Edgeworth Network* Projects

Máirín MacCarron
University College Cork, Ireland

Abstract The *Gendered Networks* and *Digital Edgeworth Network* projects applied data-driven approaches and digital analysis to material that might not seem entirely conducive to such methodologies. The process of creating viable datasets for both projects was researcher-led due to the complexity of the data. The resulting datasets highlighted aspects of the source material that had hitherto received little attention and led to the development of new research questions. Our data-driven approach ensured that all individuals were included in the data gathering process and allowed us bring to the fore people who might otherwise have been deemed invisible.

Keywords Data analysis. Gendered data. Social network analysis. Correspondence networks. Bede. Maria Edgeworth. Narrative sources.

Summary 1 Introduction. – 2 *Women, Conflict and Peace: Gendered Networks in Early Medieval Narratives*. – 3 *Digital Edgeworth Network*.

1 Introduction

Data-driven approaches to historical sources have expanded the potential of historical inquiry. The advent of easily accessible computer processing power and the mass digitisation of a whole range of sources have allowed historians to ask questions at a scale that would have been unthinkable for earlier generations. However, the

Edizioni Ca'Foscari

Studi di archivistica, bibliografia, paleografia 8
e-ISSN 2610-9093 | ISSN 2610-9875
ISBN [ebook] 978-88-6969-919-1 | ISBN [print] 978-88-6969-920-7

Peer review | Open access
Submitted 2023-10-03 | Accepted 2024-01-18 | Published 2025-05-22

DOI 10.30687/978-88-6969-919-1//008

process of transforming complex historical evidence into data is not straightforward and often requires a different methodological approach which can lead to a researcher requiring a new relationship with their sources.[1] The first step in such endeavours centres on the creation of usable datasets, with an increasing emphasis on not just big but smart data, that is, enriched and structured data (Schöch 2013). While the emphasis on history as 'data' can provoke unease, historians have always managed and categorised their information, and it is not such a major conceptual leap to think of these 'datasets' as enriched and open to additional forms of analysis, whether quantitative or digital.[2] Additionally, in an influential article over a decade ago, Johanna Drucker argued that 'capta' is a better word in the Humanities, because 'data' means 'what is given', while 'capta' means 'what is taken', which is a more accurate reflection of the creation of datasets in the humanities (Drucker 2011; Vitali, Pasqual, *supra*).

Terminology aside, however, the biggest challenge when attempting to create viable 'datasets' from humanities sources concerns the reductive nature of the process. The more complex the source – whether life experience, a painting, or a written text – the more complicated the process of data collection, because not all the complexity can be retained (D'Ignazio, Klein 2020, 10). This is not of itself a negative because meaningful data analysis is not possible without some reduction, and decisions about what to retain is part of a heuristic process that feeds back into the analysis. The challenge for the historian is to determine an acceptable level of loss vis-à-vis the potential gains from data analysis, and to ensure the data that is gathered can address the research questions posed. The affordances of a data-driven approach are many: for example, in analysis of social groups, it is possible to quantify relationship data and visualise social networks in whole new ways. In addition, if all individuals are included in the dataset, irrespective of their perceived or otherwise social importance, we can bring to the fore previously less visible or over-looked figures (Hillner, MacCarron, forthcoming).

The two projects that will be discussed here, *Women, Conflict and Peace: Gendered Networks in Early Medieval Narratives* and the *Digital Edgeworth Network*,[3] both used data-led approaches and digital methods, including social network analysis, in analysing our chosen corpora of sources. In what follows, I will outline the data models that we developed and share some of our findings, including our

1 Hitchcock 2013 has lucidly highlighted the challenges historians face from mass digitisation and increasing reliance on searches run by OCR.

2 On quantitative approaches, cf. Lemercier, Zalc 2019.

3 See below for full information including details of funders and composition of project teams.

success in highlighting the presence of individuals or patterns in our datasets that have received little attention in scholarly analysis of these sources. As gender representation was a concern of both projects, though less so in the *Digital Edgeworth Network*, we opted for researcher-led data gathering, rather than computerised methods of data extraction, such as distant reading, which have been shown to be deficient when attending to non-dominant subjects, including women (Klein 2018; Cvetkovich et al. 2010).[4] Although machine learning data gathering processes can ward against potentially subjective data collecting practices, and are not as time-consuming as extracting data through close reading, these methods cannot match the same level of detail as a trained researcher engaged in textual analysis which is what both projects required.

2 *Women, Conflict and Peace: Gendered Networks in Early Medieval Narratives*

Women, Conflict and Peace: Gendered Networks in Early Medieval Narratives (hereafter referred to as *Gendered Networks*), used a data-driven approach to interrogate the role of women in early medieval narrative sources.[5] This was an interdisciplinary project with historians, physicists and a computer scientist as full members of the team.[6] The project's goal was to produce a quantitative large-scale investigation of women's representation in narrative sources, particularly focussing on their roles as connectors of men. This, in part, emerged from the idea that women were often presented as 'peaceweavers' in heroic narratives: political marriages were often arranged in attempts at conflict resolution, whereby a woman would marry the son of a family with which her family were at war and move to her new husband's region in an attempt to weave peace between their

4 Mandel 2016 suggests ways to counteract the gender bias in so many digital projects.

5 This project was funded by the Leverhulme Trust RPG 2018-014, and ran from 1 September 2018 to 31 January 2021; the project website is available here: `https://sites.google.com/sheffield.ac.uk/genderednetworks/home`.

6 The core team consisted of three historians (Julia Hillner, Máirín MacCarron, Ulriika Vihervalli), three physicists (Silvío Dahmen, Ralph Kenna, Sandra Prado) and one computer scientist (Ana Bazzan); we were assisted by another physicist (Pádraig MacCarron) and two historians (Rob Heffron and El Bailey). The historians prepared the datasets using close reading and interpreted results; the scientists wrote bespoke codes to mathematically analyse the datasets, and produced data visualisations including graphs, charts and tables; both groups worked together in analysing and interpreting our findings. See the team page on the project website: `https://sites.google.com/sheffield.ac.uk/genderednetworks/team`.

families.[7] We examined narrative sources, specifically histories and hagiographies (spiritual biographies, usually accounts of saints' lives), from the late Roman world, Merovingian Francia and early medieval Britain, that dated from the fourth to the eighth centuries. We wished to determine if the representation of women and their social role was regionally specific and if it changed during our chosen time period. In applying social network analysis to our datasets, we were specifically interested in character networks and dynamic networks (i.e., assessing temporal influence in a social network).[8]

The process of creating datasets for *Gendered Networks* was complicated because we required a sophisticated data model that could be applied to all our sources that would allow us to conduct social network analysis at a deeper level. To do so, we needed to distinguish between several different types of relationship, and ultimately settled on a data model that accounted for twenty-one different categories of relationships. These were grouped under nine broad headings and expanded into sub-headings for greater clarity: for example, the first heading, 'Family', covers a range of relationships including domestic slaves because we wished to indicate the crucial role such people played in the household **[tab. 1]**. Occasionally, the classification of a relationship is not obvious: for example, if a pagan barbarian king murdered a Christian king from a different polity, should that be recorded as military, political or religious hostility? Ultimately, our decision was determined by the context provided by the source: if our author indicated that the interaction was informed by both political and religious motives, we recorded the connection under both headings; if it was purely political, it was only recorded under that heading and so on. Finally, categories 20 and 21 – post mortem and supernatural – are included because we occasionally encountered interactions between the living and the dead, or with angels and demons. We included such relationships to respect the integrity of the sources by recording all individuals and all connections, but these categories can also be removed from our analyses by focussing on categories 1 to 19 only.

7 Cf. *Beowulf*, line 1942: the Old English word was 'freothuwebbe'. A classic study of this phenomenon is Rosenthal 1966.

8 On applying social network analysis to these sources, cf. Hillner, MacCarron, forthcoming.

Table 1 *Gendered Networks* Data Model

1	Family	Kinship
2		Marriage & betrothal
3		Concubinage
4		Domestic relations (inc. household slaves/ servants)
5		Fosterage
6	Friendship	Friendship
7	Religious	Church meetings/councils
8		Monastic/Religious *familia*
9		Spiritual kinship (e.g., godparents)
10	Patronage	Patronage
11	Politics	Political connections
12		Physical diplomacy including messengers
13	Hostility	Military violence
14		Political hostility
15		Domestic hostility
16		Religious hostility
17		Gender-based violence
18	Transmission of Information	Letter recipients
19		Sources of information, textual or oral
20	Post Mortem	Post mortem (connections between the living and the dead)
21	Supernatural	Supernatural/Divine interventions

In addition to categorising relationships between characters in our sources, we also recorded attribute data for every character. This included their gender, whether they were named or unnamed in the source, if they were a group or individuals, and their ethnicity, social status, religious status, and geographical location when this information was available. Such information ensures our datasets are rich and allow us to address a range of research questions. Our central research question concerned gender, but the process of data gathering led to the development of additional research inquiries: for example, we detected variations in naming practices which led us to add entries to our datasets concerning whether characters were named when they appeared in our sources. On investigating these practices, we discovered that nearly two-thirds of female characters were left nameless by our authors, whose decisions were influenced by pragmatic, rhetorical or socio-cultural reasons. In addition, our attention to naming practices highlighted the important role played by nameless characters in historical sources and revealed that such individuals can become a scaffold on which plots are structured (Hillner, MacCarron, Vihervalli 2022). This research direction underlined the

importance of carefully managing the presence of the anonymous and unnamed in data-driven research projects.

The complexity of our data model with its attention to relationship categories and character attribute data ensured that our data gathering could only be done by close reading of each text, as explained in the introduction. We treated each unit of a text, whether chapters or sections, as discrete, and we recorded a character anew every time they appeared in a new unit of the text. This method enabled us to weight relationships; i.e., in addition to categorising relationships by type, we also measured the number of times that any two characters were connected in a text, leading to greater depth in our analysis of social networks. Secondly, taking account of the units within texts allowed us to trace progression over time in each source. Our particular concern was to follow plot development in the narrative time of the author, rather than in chronological time, though these often mirrored each other; however, if a narrative moves backwards and forwards in chronological time, our method allows us to follow the narrative's internal chronology. We used the time data to analyse dynamic networks within our texts, i.e., assessing influence over time, as well as engaging in more traditional, static social network analysis (Prado et al. 2020).

The efficacy of our approach can be seen in the following case study concerning Osthryth (d. 697), queen of the kingdom of Mercia in early medieval Britain. Osthyrth was born into the royal house of Northumbria in the mid-seventh century, became queen of Mercia during the 670s (precise date is unknown), and was subsequently murdered by her own nobles in the late-seventh century.[9] Her presence in the source record is minimal: she appears three times in Bede's *Historia ecclesiastica gentis anglorum* (completed c. 731) and once in the *Life* of Wilfrid of York (c. 634-709/10) which was written within a few years of Wilfrid's death by Stephen, a member of Wilfrid's community (Bede 1969; Stephen 1927).[10] Osthryth appears in the narrative of Bede's *Historia* due to her role as queen of Mercia, and hers was undoubtedly a politically-motivated marriage intended to bring peace between the warring kingdoms of Mercia and Northumbria (Bede 1969, 3.11, 4.21). Indeed, the marriage of Osthryth and Æthelred was the third such marriage between the royal families of Mercia and Northumbria within a generation, none of which

9 The early medieval kingdom of Mercia roughly corresponds with the modern-day English midlands; the kingdom of Northumbria spanned present-day northern England and southern Scotland. Mercia and Northumbria shared a border, which was often in flux in this period.

10 On Bede, cf. Darby, MacCarron 2023. On Stephen and the *Vita Wilfridi*, cf. Thacker 2013; Stancliffe 2013.

succeeded in bringing about peace.[11] That Osthryth was murdered by her Mercian nobles underlines the difficulties of her role, although the specific reasons for her murder are unknown; Bede recorded her death in a short recapitulation of events included in the final chapter of the *Historia* (5.24) but not in the main narrative.

Our other source for Osthryth, the *Life of Wilfrid*, refers to her only once. Wilfrid of York was a contentious character who was exiled from his bishopric of Northumbria three times, the first two at the hands of King Ecgfrith of Northumbria, Osthryth's brother. On the occasion of his second exile, Stephen related that the bishop fled first to Mercia, then to the kingdom of the West Saxons, further south, and he finally found refuge in the kingdom of the South Saxons. Wilfrid's hagiographer claimed that his hero had been driven from the kingdoms of the Mercians and the West Saxons because of royal kinship networks, noting that the Northumbrian king's sister, i.e., Osthryth, was queen in Mercia and his sister-in-law was queen of the West Saxons. Stephen suggested that these royal couples expelled Wilfrid to ingratiate themselves with Ecgfrith (Stephen 1927, 40).[12] This chapter is Osthryth's only appearance in Stephen's account and she is unnamed, as was the queen of the West Saxons whose name is unknown.

Osthryth is clearly something of a shadowy figure in the narratives of the only two sources that mention her. Consequently, she seems somewhat peripheral in the networks of both sources. However, when we consider her place in these networks using metrics such as degree and betweenness centrality, her significance in the network structure is greater than one might expect. Degree is the number of connections that any character, or node, has in a network: for example, a character who is connected to five other nodes would have a degree of five. Betweenness centrality represents the role that an individual node plays in connecting others within the network: that is, an individual character, A, could have a low degree, but high betweenness if the characters that A is connected to are connected to many others, meaning that A is important in connecting those that A is between. The dataset we created from Bede's *Historia ecclesiastica* contains 594 characters, and just over 12% of those are women: 72 women, 509 men, and 13 non-gender.[13] Of these 594 characters, Osthryth is ranked seventy-fourth for degree, and thirty-first for betweenness. We see a similar pattern in Stephen's *Life of Wilfrid*,

11 Cf. Bede 1969, 3.21, for the earlier marriages of Osthryth's half-sister, Alhflæd, and half-brother, Alchfrith, to Æthelred's brother, Peada, and sister, Cyniburg, respectively. Indeed, many years before Osthryth's murder by her Mercian nobles, Alhflæd was accused of involvement in the murder of her husband, Peada: Bede 1969, 3.24.

12 For discussion of royal networks and Wilfrid, cf. Hillner, MacCarron 2021, 31-40.

13 Non-gender is used for those whose gender is unknown including angels and demons which are of indeterminate gender in our sources.

for which our dataset has 156 characters, and just over 15% are women: 24 women, 140 men, and 3 non-gender. Osthryth is placed sixty-fourth for degree, and fifty-sixth for betweenness.

There are two intriguing elements here; firstly, for a relatively inconsequential character, Osthryth punches above her weight in terms of network metrics in both works. This is especially notable in the *Life of Wilfrid* where she appears only once, and nameless, but is in the top half for both metrics. This result is, no doubt, largely due to her connections to royal houses, but it is nonetheless a significant result. Secondly, in both sources, Osthryth ranked higher for betweenness centrality than for degree. Essentially, she is not distinguished by the number of her direct connections in either source; however, she is of greater importance in both networks for connecting others. This is especially apparent in Bede where, as shown before, her betweenness centrality is surprisingly high, considering the size of the network and that she only appears three times in the text. It also highlights the nature of her role within this society: queens were connectors, of men and women, a reality that Bede's account reflects when examined from the perspective of relationship connections. Indeed, this reality may also underlie Stephen's account, because Osthryth is relatively highly placed for betweenness in the *Life of Wilfrid* despite only appearing once in the text.

We can further develop our understanding of Osthryth's role in these sources, and better understand the sources themselves, when we examine the representation of relationship categories in these networks. It has long been noted that Stephen and Bede treat women very differently in their works, though the reasons for these differences have been disputed. In an influential publication, Stephanie Hollis argued that women have more agency in Stephen's *Life of Wilfrid* than Bede allows for them in his works (Hollis 1992). However, a close analysis of the women in Stephen's account indicates that many are engaged in hostile relationships with Wilfrid **[fig. 1]**.[14] The graph represents hostile relationships in the *Life of Wilfrid* as defined by the *Gendered Networks* data model. The four categories in use are military, political, religious, and domestic hostility. The characters' gender is indicated by the colour and trajectory of the triangles: male nodes are teal-coloured upward-pointing triangles, female nodes are magenta-coloured downward-pointing triangles. The size of the triangles is determined by the number of connections that each character has, i.e., their degree. Wilfrid is clearly at the centre and is the largest triangle, indicating he has the highest degree in this network.

14 The networks in figures 1 and 2 were drawn by Ana Bazzan, Sílvio Dahmen and Sandra Prado using a bespoke programme written for the *Gendered Networks* project. For further discussion of what follows, cf. Hillner, MacCarron 2021, 37-40.

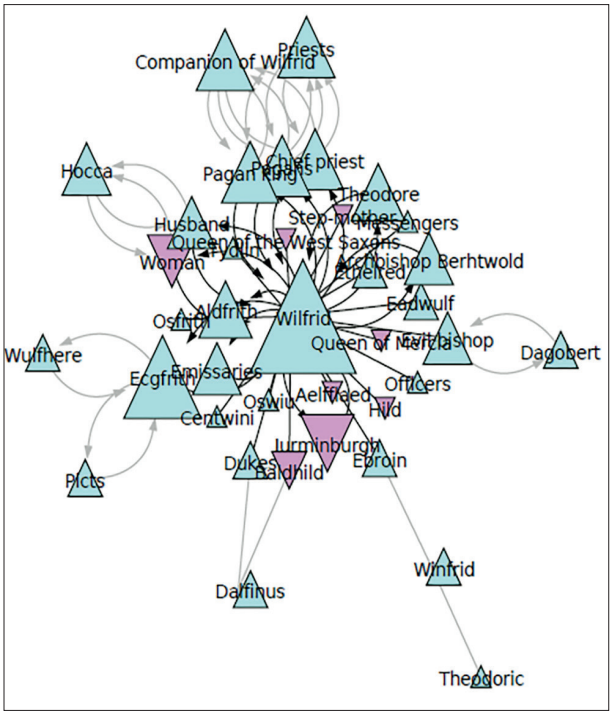


Figure 1
Graph showing hostile relationships (military, political, religious, and domestic hostility) in Stephen's *Life of Wilfrid*

Osthryth, here presented as 'Queen of Mercia' because she is nameless in the text, is to the right of Wilfrid and is one of several women in close proximity to him. The number of nodes in the graph visually represents the high level of hostility that Wilfrid generated during his life, and highlights Stephen's view that women were central to this; the significance of women in this graph becomes more apparent when we compare figure 1 with hostility graphs for Bede's *Historia*.

The *Historia* is a large text presented in five books and covers the history of Britain from the invasions of Julius Caesar in 55 and 54 BC up to the first decades of the eighth century, although most of the work is concerned with events in the seventh century. As the work is so large, it can be difficult to analyse visually, so figure 2 presents the hostility graphs for books 4 and 5 separately **[fig. 2]**. These books cover the periods of Wilfrid's enforced exiles in the late seventh and early eighth centuries. As with the *Life of Wilfrid*, these graphs represent hostile relationships defined by the *Gendered Networks* data model, and include the same four categories of military, political, religious, and domestic hostility. The colour-coding, trajectory and size of triangles are the same as in figure 1.

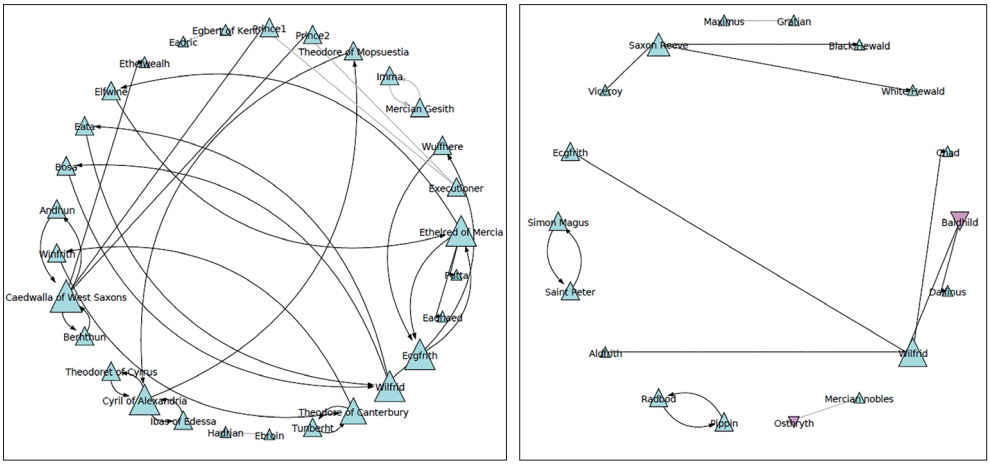


Figure 2 Graph showing hostile relationships (military, political, religious, and domestic hostility) in book 4 and 5 of Bede's *Historia*

It can be difficult to perform detailed visual analysis of these graphs, which is why network metrics (as outlined above) are essential when analysing complex networks. The key visual finding from the graphs in figure 2 is the small number of women who appear, especially when compared with the hostility graph from the *Life of Wilfrid*. Indeed, there are no women in the hostility connections for book 4, despite the fact that there are many women in this book; moreover, book 4 is sometimes referred to as the 'book of abbesses' because Bede devoted so much attention to female religious houses.[15] There are two women in the hostility connections from book 5, but only one is connected to Wilfrid: Baldhild, a Frankish queen, who murdered the archbishop of Lyon, Wilfrid's early patron (Bede 1969, 5.19; Stephen 1927, 6). Osthryth is represented in a hostile relationship with her Mercian nobles, to represent her murder, but she is not connected to Wilfrid. A comparison of these hostility graphs reveals the extent to which Bede omitted the female network hostile to Wilfrid in his account, despite this being a significant feature of Stephen's account, and Stephen was one of Bede's key sources for the *Historia*. These

15 Cf., for example, Abbess Æthelburh and the female community at Barking, Bede 1969, 4.6-10; Abbess Æthelthryth and Ely, Bede 1969, 4.19-20; Abbess Hild and the double-monastery of Whitby, Bede 1969, 4.23-4; Abbess Æbbe and the double-monastery at Coldingham, Bede 1969, 4.25.

findings suggest that Bede may have suppressed evidence of women hostile to Wilfrid, rather than women in general.[16]

This analysis reveals the ways in which our understanding of these sources can be enhanced when taking a data-led approach. A quantitative and digital approach combined with qualitative analysis has added another dimension to our interpretation of the presentation of women in the structure and development of these narratives. Our second project concerns a different type of source material, specifically correspondence collections, but also conducive to a data-driven approach.

3 *Digital Edgeworth Network*

The *Digital Edgeworth Network* (hereafter referred to as *DEN*) was developed to explore and analyse the archive of the celebrated author, Maria Edgeworth (1768-1849) and the Edgeworth family.[17] The Edgeworths were an Anglo-Irish family who owned extensive properties in the midlands of Ireland when Ireland was under British rule. Their archive contains over 40,000 items and the holdings are primarily held in two institutions in different countries: the Bodleian Library in Oxford, UK, and the National Library of Ireland in Dublin.[18] *DEN* was devised as a small-scale proof-of-concept project which would explore the potential of data-driven methods for approaching a complex and divided archive. The project was jointly based in the University of Oxford and University College Cork.[19]

Maria Edgeworth is the primary reason for interest in the Edgeworth archive, but we wished to examine the collection in the broadest possible terms rather than focussing on the family's most famous member. As the archive is extensive, we focussed on the Edgeworth

16 Stephen's pejorative attitude towards queens is further evidenced in calling two of these women Jezebels: the Frankish queen, Baldhild (Stephen 1927, 6); and Queen Iurminburgh of Northumbria, who was Osthryth's sister-in-law (Stephen 1927, 24). Jezebel was Ahab's queen in 1 Kings 16-9, she was accused of killing the prophets and became an archetypal evil queen (Nelson 1986). It is notable that Bede did not resort to this *topos*, despite having Stephen's *Life of Wilfrid* as a source. For further discussion of Bede's nuanced approach to female characters in the *Historia*, cf. MacCarron 2017; Prado et al. 2020.

17 For an introduction to Maria Edgeworth's life and works, cf. Ó Gallchóir 2021.

18 There are also small collections of Edgeworth papers elsewhere, for example, Archives at Yale.

19 *DEN* was jointly funded by the Irish Research Council and the Arts and Humanities Research Council in the UK as part of the UK-Ireland Collaboration in Digital Humanities Networking programme: `https://research.ie/2020/07/27/12-new-uk-ireland-digital-humanities-collaborations-announced/`. The team was Clíona Ó Gallchóir and Máirín MacCarron (Cork), Ros Ballaster and Anna Senkiw (Oxford). The project ran from 1 August 2020 to 30 November 2021.

family correspondence, which encapsulates both the potential and challenges posed by the entire collection. The Edgeworth correspondence extends from the 1760s to the 1850s and comprises about 10% of the holdings in both libraries. It includes letters to and from multiple members of the Edgeworth family including cousins and in-laws. It is generally accessed through two different avenues. The letters in the Bodleian are listed in a typescript document dating from the 1970s which catalogues the correspondence by sender and is known as the 'Colvin Calendar'. It was created by Christina Colvin, a member of the family, and covers 2,754 items.[20] The letters in the National Library of Ireland (NLI) are catalogued in 'NLI Collection list 40'.[21] The NLI holds just over one-third of the total correspondence with 1,409 items. 'Collection list 40' organises the letters by date irrespective of sender. Consequently, the researcher who wishes to access the Edgeworth correspondence is faced with two distinct collections, which are organised differently: one is catalogued by the sender's name, and the other by the date that letters were sent. In addition to the archive being divided, this also means that it is not possible to examine the entire collection either chronologically or by sender, and both catalogues have gaps. A further challenge is that, because comparatively little of the correspondence has been digitised, a researcher who wants to understand the correspondence has to start with these catalogues. Consequently, both the logistical challenge and time commitment are significant.

In *DEN*, we attempted to address these problems by working with both libraries in an effort to harmonise their catalogues, and by engaging in digital analysis of both catalogues together. Our intended purpose was twofold: firstly, we wanted to demonstrate that the collection could be treated as a whole, though working on the catalogues from both libraries rather than the actual holdings; and secondly, we applied social network analysis to the dataset we created to determine the efficacy of applying digital approaches to a divided archive. As this was a correspondence network, the data model was simpler than that developed for *Gendered Networks*. We created a master spreadsheet of the correspondence which contained data on sender, recipient, date, location of sender, location of recipient, and recorded the gender of sender and recipient. The data creation was a laborious process, but we successfully harmonised the data, in so far as was possible, and highlighted gaps in the catalogues: for

20 The 'Colvin Calendar' has been digitised and can be accessed on the Bodleian Libraries website: `https://archives.bodleian.ox.ac.uk/repositories/2/archival_objects/64160`.

21 'NLI Collection list 40' is available as a PDF on the NLI website: `https://www.nli.ie/sites/default/files/2022-12/040_edgeworth.pdf`.

example, information relating to senders or recipients is occasionally missing, and location data was frequently omitted. A further complication was that some letters had two or more recipients, and others had two or more senders; although the existence of multiple recipients and senders is not very surprising when assessing a family's interactions, it posed a further challenge when attempting to reduce the complexities of the correspondence to usable data.

The complete spreadsheet contains 4,544 rows, and we recorded far more data than we have, as yet, been able to use. For example, in addition to the obvious sender, recipient, and location information, we also recorded information on other people named in the letters, additional locations, literary works sent with letters (often the case with Maria Edgeworth), and literary works mentioned in letters (an even more common occurrence). Although the process of data gathering, carried out by Anna Senkiw, was painstaking and time consuming, it demonstrated the feasibility of turning this divided archive into a usable dataset that allowed us to engage with and visualise the Edgeworth correspondence in an entirely new way.

Figure 3 shows everyone in the network who sent or received at least one letter **[fig. 3]**. There are 359 nodes and 545 connections, and the size of each node indicates the number of letters sent or received.[22] The biggest nodes are Maria Edgeworth, and her father, Richard Lovell Edgeworth (both nodes are pink), who dominate the correspondence and is what one would expect as the most prominent members of the family. However, the next most connected members of the correspondence network (green nodes) joined the family at different times: Frances Ann Edgeworth née Beaufort was Richard Lovell's fourth wife, with whom she had six children, and she was a year younger than her stepdaughter, Maria; Charles Sneyd was Richard Lovell's son and twelfth child. The importance of these two members in the network, especially Frances Ann, is somewhat surprising and would have been more difficult to assess using traditional approaches.

Figure 4 shows a subset of the same network, but only represents everyone who appears at least twice in the collection **[fig. 4]**. Consequently, it is more visually accessible. The same four figures dominate: Maria, Richard Lovell, Frances Ann, and Charles Sneyd. Maria and Richard Lovell are green nodes; Frances Ann and Charles Sneyd are grey-blue; the remaining nodes are dark blue.

In the next stage of our analysis, we restricted the correspondence network to that of the immediate Edgeworth family: specifically Richard Lovell, his wives and children (he had four wives and twenty-two children), his siblings and their families, creating a network

22 The graphs in figures 3 and 4 were drawn by Máirín MacCarron using Gephi 0.9.2.

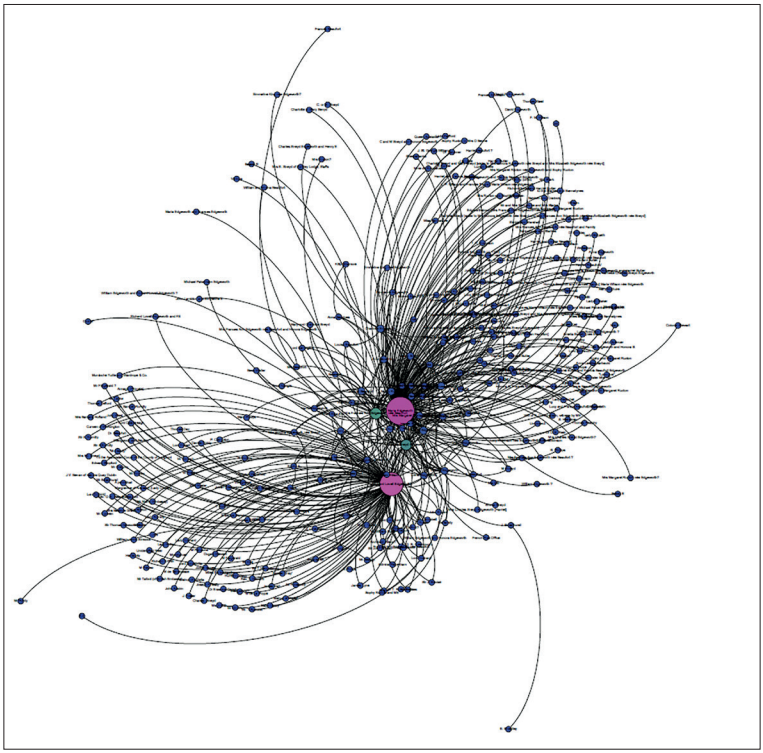


Figure 3 Network of the Edgeworth family correspondence drawn from the 'Colvin Calendar' and 'NLI Collection List 40'

of 38 nodes (Senkiw 2024). This analysis revealed the predominance of women's voices in the network; indeed, women outnumber men by just over two to one (there are 26 women and 12 men). Such an insight can be missed when focussing on the most important figures in a network, rather than a network as a whole. Further analysis also revealed that half of the correspondents in this network were children when they began writing to family members and their letters were preserved alongside those of the adults. Several of these children also appear later in the network as adults.

These networks represent the affordances of applying a data-driven approach to a complex archive such as the Edgeworth papers. Our findings, such as the predominance of children, would have been difficult to identify using other methods because of the limitations of the catalogues. Indeed, we did not set out to look for children, but as our analysis developed their presence became apparent. In attempting to put order on what we were finding, we devised new categories

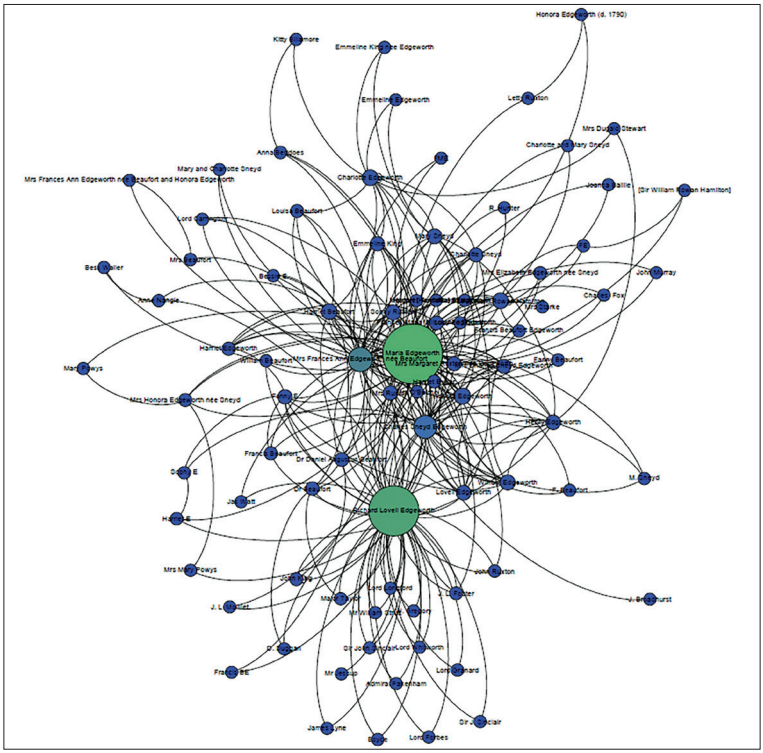


Figure 4 Network of everyone who appears at least twice in the Edgeworth family correspondence as represented in the 'Colvin Calendar' and 'NLI Collection List 40'

to account for the variety in the dataset. We have shared our findings with our library partners, and they are exploring the possibility of including a category for children in subsequent metadata. Our experience indicates, among other things, that when one begins extracting data from historical sources, the very process of extraction reveals types of data that one was not expecting to find. Occasionally, this is data that might previously have been considered invisible.

The *Gendered Networks* and *DEN* projects applied data-driven approaches and digital analysis to material that might not at first seem entirely suitable for such methodologies. Both projects engaged in researcher-led close reading to create their complex and multi-faceted datasets. This was a time-heavy commitment, but the process of data creation allowed us to see aspects of the source material that had hitherto received little attention, which led to the development of new research questions, including the role of nameless characters in narrative texts, and the presence of children in the Edgeworth

correspondence collection. The data-driven approach also ensured that all individuals were included in our data gathering, regardless of their perceived importance, and allowed us to bring to the fore people who have often been overlooked.

Bibliography

Bede (1969). *Bede's Ecclesiastical History of the English People*. Ed. and transl. by B. Colgrave; R.A.B. Mynors. Oxford: Clarendon press. Transl. of: *Historia ecclesiastica gentis anglorum*.

Colgrave, B. (1927). *The Life of Bishop Wilfrid by Eddius Stephanus*. Cambridge: Cambridge University Press.

Cvetkovich, A. et al. (2010). "Women as the Sponsoring Category: A Forum on Academic Feminism and British Women's Writing". *Partial Answers*, 8(2), 235-54. https://muse-jhu-edu.ucc.idm.oclc.org/article/382600

Darby, P.; MacCarron, M. (eds) (2023). *Bede the Scholar*. Manchester: Manchester University Press.

D'Ignazio, C.; Klein, L.F. (2020). *Data Feminism*. Cambridge (MA): MIT Press. https://direct.mit.edu/books/oa-monograph/4660/Data-Feminism

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

Hillner, J.; MacCarron, M. (2021). "Female Networks and Exiled Bishops between Late Antiquity and the Early Middle Ages: The Cases of Liberius of Rome and Wilfrid of York". Bérat, E.O.; Hardie, R.; Dumitrescu, I. (eds), *Relations of Power: Women's Networks in the Middle Ages*. Bonn: Bonn University Press, 19-44.

Hillner, J.; MacCarron, M. (forthcoming). "Social Network Analysis". Flower, R. et al. (eds), *A Companion to Roman Prosopography*. Leiden: Brill.

Hillner, J.; MacCarron, M.; Vihervalli, U. (2022). "The Politics of Female Namelessness between Late Antiquity and the Early Middle Ages, circa 300 to 750". *Journal of Late Antiquity*, 15(2), 367-401. https://doi.org/10.1353/jla.2022.0021

Hitchcock, T. (2013). "Confronting the Digital: Or How Academic History Writing Lost the Plot". *Cultural and Social History*, 10(1), 9-23. https://doi.org/10.2752/147800413X13515292098070

Hollis, S. (1992). *Anglo-Saxon Women and the Church: Sharing a Common Fate*. Woodbridge: Boydell.

Klein, L. (2018). "Distant Reading after Moretti". *Arcade: Literature, the Humanities and the World Blog*, 29 January. https://arcade.stanford.edu/blogs/distant-reading-after-moretti

Lemercier, C.; Zalc, C. (2019). *Quantitative Methods in the Humanities: An Introduction*. Transl. by A. Goldhammer. Charlottesville : University of Virginia Press. Transl. of: *Méthodes quantitatives pour l'historien*. Paris: La Découverte, 2008.

MacCarron, M. (2017). "Royal Marriage and Conversion in Bede's *Historia ecclesiastica gentis anglorum*". *Journal of Theological Studies*, 68(2), 650-70. https://doi.org/10.1093/jts/flx126

Mandel, L. (2016). "Gendering Digital Literary History: What Counts for Digital Humanities". Schreibman, S.; Siemens, R.; Unsworth, J. (eds), *A New Companion to Digital Humanities*. Oxford: Blackwell, 511-23. http://library.ucc.ie/record=b2238281

Nelson, J. (1986). "Queens as Jezebels: the Careers of Brunhild and Baldhild in Merovingian History". Nelson, J. (ed.), *Politics and Ritual in Early Medieval Europe*. London: Hambledon, 1-48.

Ó Gallchóir, C. (2021). *Maria Edgeworth*. Brighton: Edward Everett Root.

Prado, S. et al. (2020). "Gendered Networks and Communicability in Medieval Historical Narratives". *Advances in Complex Systems*, 23(3). https://doi.org/10.1142/S021952592050006X

Rosenthal, J.T. (1966). "Marriage and the Blood Feud in 'Heroic' Europe". *The British Journal of Sociology*, 17(1), 133-44.

Schöch, Ch. (2013). "Big? Smart? Clean? Messy? Data in the Humanities". *Journal of Digital Humanities*, 2(3). https://journalofdigitalhumanities.org/2-3/big-smart-clean-messy-data-in-the-humanities/

Senkiw, A. (2024). "'Much Material of 1821 Not Listed': Troubling Colvin's Calendar of Maria Edgeworth's Correspondence with Digital Analysis". *Digital Studies/Le champ numérique*, 14(1), 1-26. https://doi.org/10.16995/dscn.9865

Stancliffe, C. (2013). "Dating Wilfrid's Death and Stephen's Life". Higham, N.J. (ed.), *Wilfrid: Abbot, Bishop, Saint: Papers from the 1300th Anniversary Conferences*. Donington: Shaun Tyas, 17-26.

Stephen (1927). *The Life of Bishop Wilfrid by Eddius Stephanus*. Ed. and transl. by B. Colgrave. Cambridge: Cambridge University Press. Transl. of: *Vita Wilfridi*.

Thacker, A. (2013). "Wilfrid: His Cult and His Biographer". Higham, N.J. (ed.), *Wilfrid: Abbot, Bishop, Saint: Papers from the 1300th Anniversary Conferences*. Donington: Shaun Tyas, 1-16.

Models of Data Extraction and Architecture in Relational Databases of Early Modern Private Political Archives
edited by Dorit Raines

Extraction, Architecture and Recovery of Family Correspondence Data
The Platform "EpiCAT. Family Letters from Catalonia (Sixteenth-Nineteenth Centuries)"

Javier Antón Pelayo
Universitat Autònoma de Barcelona, España

Abstract The optimal approach for documenting information within a uniform documentary series involves the establishment of a database. To ensure the functionality of datafication, it is recommended to implement a design that encompasses accessibility, interoperability, and efficient data retrieval. An exemplar instance of this is represented by the EpiCAT platform, designed for the curation of family correspondences originating from Catalonia during the span of the sixteenth to nineteenth centuries. While this platform facilitates the association of metadata, its utilization in the context of text mining remains unexplored. This emerging paradigm necessitates novel viewpoints in the formulation of historical discourse.

Keywords Datafication. Data retrieval. EpiCAT platform. Catalonia. Family correspondences. Interoperability.

Summary 1 Introduction. –2 The Databases. – 3 The EpiCAT Platform. – 4 The Horizon of Text Mining. – 5 Conclusions.

Edizioni Ca'Foscari

Studi di archivistica, bibliografia, paleografia 8
e-ISSN 2610-9093 | ISSN 2610-9875
ISBN [ebook] 978-88-6969-919-1 | ISBN [print] 978-88-6969-920-7

Peer review | Open access
Submitted 2023-10-03 | Accepted 2024-01-22 | Published 2025-05-22

DOI 10.30687/978-88-6969-919-1//009

1 Introduction

The most efficient approach to document and record information in a homogeneous documentary series is to establish a comprehensive database. For effective datafication, it is crucial to implement a design that considers accessibility, interoperability, and agile data retrieval.

EpiCAT is a specialized platform designed to manage a vast collection of family letters dating from 1500 to 1850, housed in public archives in Catalonia.[1] Catalonia serves as an extensive yet manageable framework, acting as a laboratory that offers substantial evidence to verify specific historical processes. The primary objective of this project is to demonstrate the utility of this resource when approached systematically and on a large scale. Through this approach, the statements found in official documents can be supplemented, qualified, and at times, amended. Moreover, it allows for an adequate examination of individuals' private and intimate spheres (Borkosky 2002), as well as an in-depth exploration of seemingly mundane cultural practices and historical banalities. Additionally, when exploring these extensive documentary materials, it becomes common to encounter unprecedented epistolary sociability, revealing political, economic, social, cultural, and various other strategic pursuits.

For instance, in the examination of a marriage, official documents such as sacramental marriage books from parishes or notarial marriage chapters certify the final decision with clarity, devoid of any hesitations or consideration of alternative choices. Conversely, letters exchanged between relatives and friends before the wedding unveil the array of potential candidates the family was contemplating, including their strengths, weaknesses, economic status, and the level of compatibility between potential spouses. This broader scope of marriage possibilities allows for an understanding of the relational aspirations of a household, their evaluation of different options, and the underlying family strategy intended for implementation.

To collect and systematize this scattered and diverse information from various sources, a multifunctional database is essential. This is where the expertise of historians intersects with that of information resource professionals. While digital tools can be learned and employed, highly sophisticated computer procedures may necessitate collaboration with specialized technical personnel.

Recently, the Swedish historian Mats Fridlund classified historians into three categories: Historian 1.0 utilizes search engines and personal databases; Historian 2.0

1 https://epicat.uab.cat/.

> systematically use[s] various digital applications and quantitative methodologies for big-data text and data mining, calculations and visualisations, such as topic modelling, network analysis and text and data scraping. Most of these methods necessitate investments in acquiring expertise in or collaborators skilled in coding and database methodologies. (Fridlund 2020, 77)

On the other hand, Historian 1.5 adopts a hybrid methodology that combines quantitative and qualitative, automatic and manual methods, without the explicit use of programming and coding. The question one must ask is, at which level do I find myself? Which level of expertise is required for my research?

2 The Databases

Databases play a fundamental role in historiographical research. In essence, a database system enables the storage, recording, and retrieval of information from documentary sources. Typically, the data recorded in a database is sourced from specific primary sources, organized, and classified according to logical criteria and research objectives. The abundance of quantitative, serial, or numerical data offers numerous advantages, such as providing precise and verifiable evidence to support analyses and arguments, enhancing the consistency of formulated hypotheses, enabling more rigorous simulation models, and ultimately reinforcing the research's conclusions.

Nonetheless, the methodological appropriateness of databases and digital humanities, in general, has been the subject of profound reflections, which have also brought attention to some weaknesses of the digital paradigm. These include concerns about the infallibility of algorithms, the idea of a universal digital archive as a utopian ideal, and the potential lack of contextualisation in the retrieved data (Milligan 2022).

Databases, as excellent tools for historians, can be categorized into two types: relational and non-relational databases. Non-relational databases are structured following a hierarchical approach. On the other hand, relational databases are systems that organize information in tables, with the ability to connect these tables to one another. When dealing with structured and homogeneous data, the use of relational databases yields high-quality results (Gil 2021). A prime example of this would be the *EpiCAT. Family Letters of Catalonia (Sixteenth-Nineteenth Centuries)* database (Antón Pelayo et al. 2023).

3 The EpiCAT Platform

3.1 The Architecture

The development of the EpiCAT digital project was spearheaded by Alicia Calvo Burés. From 2017 to 2021, she served as the technical manager, overseeing the database structure, design, and web application development. Subsequently, she continued to be responsible for updating and maintaining the portal **[fig. 1]**.

Figure 1 The EpiCAT home page

The technical implementation of this project involved the creation of an internal application, comprising several interconnected relational databases, designed for handling all aspects related to the letters (back-end). Additionally, a web application was developed to facilitate the exploration and access of the collected information (front-end). The internal application allows researchers to work with letters stored in different archives in a decentralized and simultaneous manner. On the other hand, the web application primarily caters to users searching for documentary records and navigating between these records and their associated entities. While the tools integrated into the portal are directly aligned with the research objectives, considerations were also given to enabling direct access for citizens and supporting the didactic use of the epistolary materials that have been incorporated. To cater to a broader audience, the consultation is available in three languages: Catalan, Spanish, and English **[fig. 2]**.

The chosen data model for this project is of the Entity-Relationship type, a widely established model for conceptual database design since its introduction by Peter Chen in 1976. Its implementation was carried out using MySQL, which remains one of the most widely used database management systems and is supported by the web hosting service of the Universitat Autònoma de Barcelona.

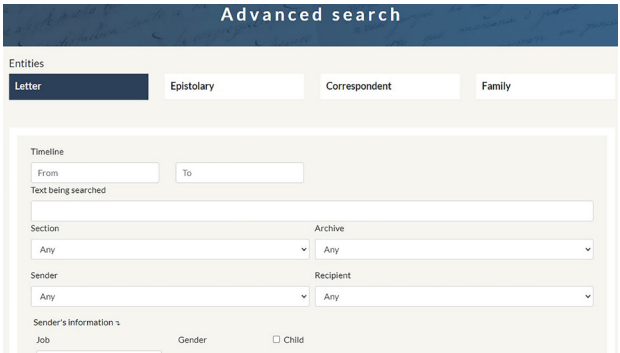


Figure 2
A glimpse of a part of the Advanced research functionality

The primary entities included in the data model are 'Epistolary', 'Letter', 'Correspondent', 'Family', and 'Fund'. The creation of the 'Epistolary' entity was driven by the necessity to consolidate letters from the same family, even if they were located in different archives. For instance, the correspondence of the Burguès family includes letters found in the National Library of Catalonia, as well as the Municipal Archive of Girona and the Archive of the Cathedral of Girona (Antón Pelayo 2005; 2013; 2019a) **[fig. 3]**.

Figure 3 Epistolary of the Burguès family

Epistolary pacts

Significant epistolary pacts.

Show 10 entries Search:

#	Start	End	Overview	Total letters	Correspondent 1	Correspondent 2
8	1811-9-26	1812-5-2	Tres cartes amoroses el 1811, quan eren nuvis, i set cartes el 1812, ja casats, separats per malaltia de Narcís.	10	Burguès i de Guàrdia, Narcís de	Caramany i de Camps, Maria dels Dolors
9	1818-7-14	1818-7-26	Cartes a la seva dona, des de Sant Martí Sescorts a Girona, on li explica les seves activitats i pregunta pels fills i la família.	6	Burguès i de Guàrdia, Narcís de	Caramany i de Camps, Maria dels Dolors
11	1826-8-30	1826-9-13	Narcís explica a la seva dona la gestió del patrimoni que fa a Coromines i pregunta per la família i amics.	4	Burguès i de Guàrdia, Narcís de	Caramany i de Camps, Maria dels Dolors
10	1823-9-27	1824-4-13	Comunicació entre Jaume, presoner a França, i el seu germà Narcís (hereu), a Girona. Descripció de la seva situació i dels seus sorollosos compromisos matrimonials.	6	Burguès i de Guàrdia, Jaume	Burguès i de Guàrdia, Narcís de

Figure 4 A part of the epistolary pacts of the Burguès family

Related letters

Letters included in the epistolary of the family Burguès

Show 10 entries Search:

#	Date	Sender	Origin	Recipient	Destination
Letter [2003]	18 de febrer de 1840	Pombo, José Ignacio de	Barcelona	Burguès i de Caramany, Maria Ignàsia	Girona
Letter [229]	28 de gener de 1840	Burguès i de Caramany, Maria Teresa de	Reus	Burguès i de Caramany, Maria Ignàsia	Girona
Letter [2002]	26 de gener de 1840	Burguès i de Caramany, Maria Ignàsia	Girona	Pombo, José Ignacio de	Barcelona
Letter [2001]	17 de gener de 1840	Claveria y de Haro, Rafaela de	Madrid	Burguès i de Caramany, Maria Ignàsia	Girona
Letter [228]	16 de juliol de 1839	Burguès i de Caramany, Maria Teresa de	Reus	Burguès i de Caramany, Maria Ignàsia	Girona
Letter [227]	28 de maig de 1839	Burguès i de Caramany, Maria Teresa de	Reus	Burguès i de Caramany, Maria Ignàsia	Girona

Figure 5 Part of the related letters of the Burguès family

3.2 The Metadata

Family letters, dating back to antiquity, are typically composed on a simple support, often a sheet of paper, and adhere to a set of elements and formalities that have remained relatively unchanged over time. Despite each letter being a spontaneous and autonomous creation of an individual, this consistency allows for the extraction of abundant and standardized factual information (Antón Pelayo 2019b).

The "factual metadata" (Méndez Rodríguez 2000) that can be derived from a letter includes the names of the sender and recipient, creation date, place of origin, often the intended destination, document measurements, number of sheets in the document, a description of the document (overwritten sections, mail marks, paper

characteristics, etc.), and the language of the text. Additionally, it is beneficial to gather other factual metadata that may not appear in every letter but can help in characterizing the correspondents, such as gender, age, profession, and kinship.

On the other hand, the 'descriptive metadata' of the letters has posed one of the most challenging aspects of the EpiCAT project due to the heterogeneity, uniqueness, and variable nature of the content found in family letters. Initially, controlled descriptors or pre-established vocabularies, designed by project members, were applied. However, the dynamic content of the letters necessitated the inclusion of free descriptors to capture the nuanced aspects of epistolary communication. To ensure a minimal level of control and terminological consistency, the Tesaurus d'Història de Catalunya[2] was adopted as the reference vocabulary (Cuadrado et al. 1994).

EpiCAT incorporates two levels of descriptive metadata: 'topics' and 'subjects'. The 'topic' encompasses terms or expressions that encapsulate the overall subject of the letter, with 45 topics currently available, allowing the application of only one per letter. Labelling the global content of a letter is not always straightforward; however, certain letters may have a specific purpose, making categorization easier. For instance, love letters, travel correspondence, letters of consolation, recommendation, gratitude, and marriage are examples where labelling is more straightforward. Remarkable research has been carried out by Montserrat Jiménez Sureda (2020) on love letters **[fig. 6]**.

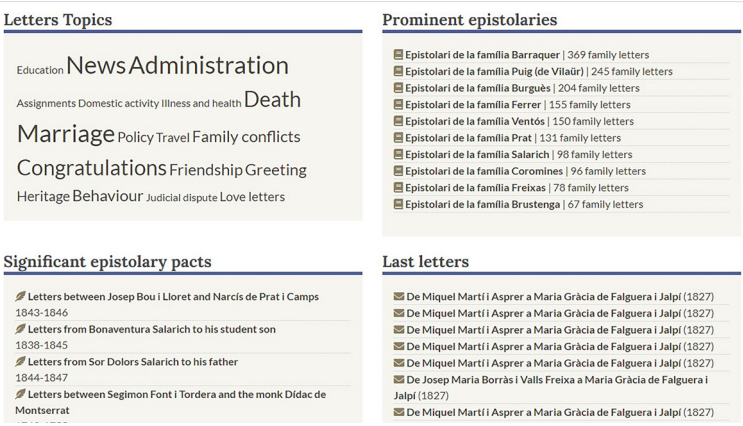


Figure 6 Lists of letters topics, epistolaries and epistolary pacts

2 https://webs.uab.cat/sibhilla/tesaurus-dhistoria-de-catalunya-0/.

The 'subjects' comprise a broader set of labels (currently 183 terms) intended to describe most, or all of the content found in each letter. There is no limit to the number of descriptors that can be applied in this field, which represents the most qualitative and refined content indexing process throughout the registration process. Thanks to this tool, searching and retrieving information on the minutest details of epistolary documentation attains historical significance. Some examples of minimal themes frequently reported in several letters include affection, domestic animals, dances, silkworms, board games, lightning, donkey's milk, lottery, pig slaughter, fashion, private oratories, locust plague, laughter, and sadness. Additionally, there are more prevalent and extensive 'subjects' communicated in the letters, encompassing children's activities, love, Carlism[3], marriage, weather, mail, death, women, education, political events, family news, and health, among others.

3.3 Data Retrieval

The design of the information retrieval system for the EpiCAT platform incorporates three distinct approaches: a simple search engine, an advanced search engine, and faceted navigation. The advanced search engine offers various filters in dropdown format, enabling users to obtain answers to specific questions. For instance, researchers can inquire about letters sent from Barcelona between 1820 and 1835, the number of letters written by women during that period, digitized letters centred around marriage themes, correspondents identified in EpiCAT with a legal profession, and the quantity of letters written by girls in the French language.

The search engine results table allows for flexible modification of the listed elements order criteria, enabling chronological sorting in ascending or descending order, as well as alphabetical sorting based on any of the listed columns. Additionally, a text box is provided for users to apply free-text filters. Moreover, data can be exported in CSV, Excel, or PDF format, if desired.

Faceted navigation, available on the exploration page, facilitates dynamic filtering of records associated with charts, allowing users to narrow down the dataset to highlight the most relevant information. This dynamic filtering functionality is implemented using Vue.js, a versatile and widely used JavaScript library **[fig. 5]**.

3 Spanish political and ideological movement that supported the candidacy of the Infante Carlos against his niece Isabel, in 1833, unleashing a civil war in which the dynastic question was a pretext to try to keep intact the structures of the Old Regime and fight against the expansion of liberalism.

Regarding the files for each entity, careful attention has been given to establishing links between them, enabling independent access to the files of individual letters, families, correspondents, or epistolary relationships.

The epistolary-type records contain various data points, such as transcriptions, subjects, themes, and family relationships between senders and recipients when available. In addition, the database identifies epistolary relationships that link two individuals in specific time and space, revolving around particular themes. These stable epistolary connections have been termed 'relevant epistolary relationships', akin to authentic 'epistolary pacts' **[fig. 4]**.

Most of the letters have undergone complete transcription. The applied transcription criteria aim to ensure the collected material's utility to both historians and philologists. To this end, the criteria seek to facilitate reading while respecting the original text's idiosyncrasies. Thus, essential interventions have been made to serve both purposes. Guiding principles include applying the contemporary punctuation system to the original texts, adopting current conventions for upper- and lower-case usage in the document transcripts, accenting transcribed words following modern regulations, and preserving the phonetic features of the originals as much as possible. Furthermore, square brackets […] have been introduced to indicate added letters that aid readability, while angle brackets <…> denote words, syllables, or letters appearing in the originals that may distort the reading.

The platform also includes some representative scans of charts, constituting approximately 10% of the registered letters. These scans, stored in PDF format, are archived in the UAB (Universitat Autònoma de Barcelona) Digital Document Repository – an open access tool – with policies for document preservation and version control using Git system,[4] ensuring their integrity and long-term accessibility.[5]

4 `https://lore.kernel.org/git/xmqqleh3a3wm.fsf@gitster.g/T/#u`.

5 `https://www.uab.cat/web/our-collections/uab-digital-repository-of-documents-345777080660.html`.

4 The Horizon of Text Mining

A more intensive and detailed approach to annotating the elements of a letter involves using the XML-TEI system, which has been employed in projects like the *Digital Archive of Letters in Flanders*[6] and *P.S. Post Scriptum*.[7] However, it should be noted that the TEI system primarily serves linguistic purposes, and its implementation demands a significant time investment.

While there are automatic marking procedures and experimental methods for the automatic transcription of historical documents, manual review remains essential. In the case of EpiCAT, the TEI procedure was recently considered by our team, and it was concluded that the time involved does not justify the benefits, particularly for historians. The primary objective of analysing letters in this context is to provide insight into historical processes or subjects that are challenging to document (Mora Mellado 2022). These may include parent-child relationships, expressions of emotions, manifestations of friendship, female sociability, educational strategies, family solidarity, and more.

Family letters incorporated into EpiCAT are transformed into structured data upon registration, and the attached textual transcriptions are intended to facilitate reading while preserving the unique writing characteristics of each author from a time when languages lacked standardisation. As such, specific transcription criteria are applied to capture the lexical peculiarities of individual letter writers. Conversely, text mining necessitates text data preprocessing, involving cleansing and transforming the text into a usable format. This process risks discarding valuable information pertinent to historiographic work in favour of identifying patterns and extracting information from vast datasets. It is important to recognize that analysing the texts of a social network, such as Twitter-X, differs significantly from studying private epistolary correspondence from the Modern Age.

6 https://ctb.kantl.be/project/dalf/.

7 http://teitok.clul.ul.pt/postscriptum/index.php?action=home.

5 Conclusions

Digital tools offer immense potential in automating certain processes, granting historians more time to focus on other aspects of their research. These digital tools introduce novel paradigms and innovative approaches to investigating the past. Therefore, engaging in methodological discussions among specialists and disseminating these techniques and procedures to students in History and Humanities is crucial.

When utilizing a database to extract and manage data, careful design, structure, and articulation of research objectives are paramount. Clearly stating the criteria for including information in a documentary set and the assumptions guiding the decision to incorporate or discard specific materials is essential. The results of a search provide isolated data in the form of numbers and percentages, requiring contextualisation to facilitate comprehension and analysis.

Whenever possible, projects should embrace collaboration and interdisciplinarity to address more complex and unforeseen research questions. Such an approach enhances the capacity of collected materials to shed light on previously unimagined inquiries.

Bibliography

Antón Pelayo, J. (2005). *La sociabilitat epistolar de la família Burguès (1799-1803)*. Girona: Quaderns del Cercle.

Antón Pelayo, J. (2013). *La correspondència epistolar de la família Burguès (1750-1850)*. Bellaterra: Universitat Autònoma de Barcelona.

Antón Pelayo, J. (2019a). *La comunicació epistolar de la família Burguès durant l'estada a Coromines (1727-1774)*. Bellaterra: Universitat Autònoma de Barcelona.

Antón Pelayo, J. (2019b). "La teoría de la carta familiar (siglos XVI-XIX)". *Revista de Historia Moderna. Anales de la Universidad de Alicante*, 37, 95-125. https://doi.org/10.14198/RHM2019.37.04

Antón Pelayo, J. (2023). "Ordenando cartas: EpiCAT, el portal para la gestión de cartas familiares en Cataluña". Acosta, F.; Duarte, A.; Lázaro, E.; Ramos Roví, M.J. (eds), *La historia habitada. Sujetos, procesos y retos de la Historia Contemporánea del siglo XXI. Actas del XV Congreso de la Asociación de Historia Contemporánea* (Córdoba, 9-11 septiembre 2021). Córdoba: Universidad de Córdoba, 1575-81.

Antón Pelayo, J. et al. (2023). "EpiCAT. Una plataforma para la gestión de cartas familiares". *Vínculos de Historia*, 12, 358-69. https://doi.org/10.18239/vdh_2023.12.19

Borkosky, M.M. (2002). "Epistolarios: la intimidad expuesta". *Cahiers du GRIAS*, 10, 27-45.

Chen, P.P. (1976). "The Entity-Relationship Model. Toward a Unified View of Data". *ACM Transaction on Database Systems*, 1(1), 9-36. https://doi.org/10.1145/320434.320440

Cuadrado, M. et al. (1994). "Thesaurus d'Història de Catalunya: creació i anàlisi d'un llenguatge documental aplicat a la historia". *Item*, 15, 134-59.

Fridlund, M. (2020). "Digital History 1.5: A Middle Way Between Normal and Paradigmatic Digital Historical Research". Fridlund, M.; Oiva, M.; Paju, P. (eds), *Digital*

Histories. Emergent Approaches Within the New Digital History. Helsinki: Helsinki University Press, 69-87.
https://doi.org/10.33134/HUP-5-4

Gil, T.L. (2021). *How to Make a Database in Historical Studies*. Cham: Springer.

Jiménez Sureda, M. (2020). *Amb el cor al paper. Història i teoria de les cartes d'amor*. Bellaterra: Universitat Autònoma de Barcelona.

Méndez Rodríguez, E.M. (2000). "Metadatos y tesauros: aplicación de XML/RDF a los sistemas de organización del conocimiento en intranets". *La gestión del conocimiento: retos y soluciones de los profesionales de la información*. Bilbao: Universidad del País Vasco, 211-20.

Milligan, I. (2022). *The Transformation of Historical Research in the Digital Age*. Cambridge: Cambridge University Press.
https://doi.org/10.1017/9781009026055

Mora Mellado, S. (2022). *Uso de bases de datos y herramientas de marcación y minería de texto para la preservación y estudio de documentos y hechos históricos* [MA thesis]. Barcelona: Universitat Autònoma de Barcelona.

Challenges: Graziani Archive and Omeka S

Models of Data Extraction and Architecture in Relational Databases of Early Modern Private Political Archives
edited by Dorit Raines

Historical Research and Archival Sciences in a Digital Perspective
Relational Database, Data Architecture and Data Extraction in Graziani Archives Portal

Dorit Raines
Università Ca' Foscari Venezia, Italia

Abstract Historians often struggle with organizing unstructured data, as existing metadata standards and relational databases do not fully capture the complexity of historical narratives. While archival sciences have adapted to digital environments, historians face challenges in structuring data, particularly for fragmented private collections that do not conform to standard models. This article examines key decisions in a Digital Humanities project on historical sources, including platform selection, data structuring, and metadata integration. Using the *Nuncio's Secret Archives* project as a case study, it highlights strategies for standardizing, contextualizing, and interlinking data to enhance searchability and facilitate customized analysis.

Keywords Relational database. Web portal. Data architecture. Data ingestion. Front-end customization.

Summary 1 Introduction. – 2 The Source – The Graziani Archive. – 3 Scope(s) of the Project. – 4 Choice of the Right Platform. – 5 Data Architecture. – 6 Data Extraction, Data Ingestion. – 7 From Backend to Frontend. – 8 Long-Term Preservation.

1 Introduction

Unlike archivists, historians today often struggle with organizing unstructured data. Archival sciences, like library and information sciences, have long established metadata standards for their respective cultural artefacts and have further advanced these standards

Edizioni Ca'Foscari

Studi di archivistica, bibliografia, paleografia 8
e-ISSN 2610-9093 | ISSN 2610-9875
ISBN [ebook] 978-88-6969-919-1 | ISBN [print] 978-88-6969-920-7

Peer review | Open access
Submitted 2025-03-11 | Accepted 2025-04-28 | Published 2025-05-22

DOI 10.30687/978-88-6969-919-1/010

with the advent of digital technology, adapting them to the digital environment. The development of type-specific databases and, subsequently, relational databases that connect digital objects through their properties has significantly advanced archival, library, and cultural studies over the past two decades in unprecedented ways (Kamal, Golub 2025; Aliaga, Bertino, Valtolina 2011; Dallas 2004, 283). Historians, however, have remained somewhat at an impasse. While they have made progress with network analysis, they have faced considerable challenges in representing the complex contents of objects (whether records or other sort of narration) within a binary system.

Historians are primarily concerned with the content of cultural objects and their significance, in addition to the provenance metadata associated with these objects. Existing digital portals that offer tools for constructing customized relational databases are based on conceptual models that describe resources, yet they lack a corresponding conceptual model for content. As a result, even if historians do not attempt to encode the full complexity of historical narratives within a relational database and instead focus on extracting recurring entities to facilitate key access to the text, the task remains complex. Difficult decisions must often be made to ensure conceptual clarity, including standardizing, disambiguating, or even interpreting entities within their historical contexts.

A substantial body of literature currently addresses various aspects of the digitization and indexing of archival sources (for example, Gilliland, McKermmish, Lau 2017; Opgenhaffen 2022; Hawkins 2022). However, while debates within the Digital Humanities community have largely focused on the digitization of entire collections held by cultural institutions or on the development of digital libraries (Agosti, Ferro, Silvello 2011, 17), or on the archive as a research laboratory and its role in the changing digital scenery (Trace 2022) relatively little attention has been given to private archival collections. These collections, often fragmented across multiple institutions and private owners due to historical contingencies, do not always conform to existing conceptual models. Consequently, adjustments are necessary to enable both a diachronic and synchronic virtual reconstruction of such collections.

This article seeks to highlight the decision-making processes that teams of historians navigate when undertaking a Digital Humanities project centered on historical sources, particularly private archival collections.[1] Key considerations include the selection of an appropriate hosting platform, the curation and processing of materials, the chosen data architecture, challenges encountered during data structuring, and the solutions implemented. Most importantly, the project

1 In doing that I follow the suggestion made by Blaney et al. 2021, 27.

aims to incorporate and interlink the maximum possible amount of extracted data – standardized, disambiguated, and contextualized – to facilitate user searches and enable the generation of customized datasets based on specific queries. As a case study, this article examines the three-year, nationally funded project *Nuncio's Secret Archives: Papal Diplomacy and European Multidenominational Societies Before the Thirty Years' War*,[2] illustrating the workings of a historian's digital laboratory.

2 The Source – The Graziani Archive

This article presents a project focused on the Graziani archive, which has been partially explored by scholars in recent years[3] but has never been comprehensively studied or inventoried in its entirety. Three main factors have contributed to this gap in research:

1. the archive is dispersed across multiple institutions in different countries and continents, making comprehensive study challenging;
2. the largest portion of the archive remains in private ownership, held by the Graziani family in Vada, near Livorno. Although the family actively supports scholarly research and facilitates access to the materials, the archive's location is not conducive to systematic and continuous study;
3. the complex history of the archival materials has posed significant obstacles to undertaking the time-intensive and costly process of organization and inventorying.

Before outlining the scope of the project, it is useful to provide a brief historical overview of the Graziani archive and its founder Antonio Maria Graziani.[4] The Graziani family belonged to the nobility and was known as Graziani di Sansepolcro, originating from a small town near Perugia, where they had been established since the

2 Principal Investigator: prof. Elena Bonora, University of Parma. The other three units were: the University of Padua led by prof. Antonella Barzazi, the University of Modena and Reggio Emilia led by prof. Matteo Al-Kalak, and the University Ca' Foscari of Venice led by prof. Dorit Raines, in charge of the Digital Humanities part. For team participants, their roles and credits: `https://grazianiarchives.eu/s/graziani-archives/page/chi-siamo`.

3 Among others: Corsini 2000; Jaitner 2004; Moretti 2012; 2015; 2018; Bonora 2019; Moretti 2021; Jaitner 2021; Moretti 2023; Bonora 2023.

4 The following reconstruction of the history of the archive and his whereabouts is based on research made both in the Graziani archives, by the Soprintendenza archivistica e bibliografica della Toscana (`https://siusa-archivi.cultura.gov.it/cgi-bin/siusa/pagina.pl?TipoPag=comparc&Chiave=209524`), by Marsili 2002; Mariani 2022; Raines 2024.

thirteenth century. This branch of the family descended from the Graziani lineage in Arezzo, Tuscany.
Antonio Maria's father, Giulio Graziani, was a military officer who initially fought with the Venetians in the Holy League in 1509 and later served under the Florentine *condottiero* Francesco Ferrucci in the Battle of Gavinana near Pistoia in 1530. Giulio died in 1543, when Antonio Maria was six years old. Following his father's death, the family ensured that he completed his education before placing him, in 1560, under the protection of Giovanni Francesco Commendone, the future cardinal. At that time, Commendone was tasked with presenting the convocation of the Council of Trent to Emperor Ferdinand I. Graziani remained in his service until the cardinal's death at the end of 1584. As Commendone's secretary, Graziani was responsible for managing his archive. Over time, during his extensive travels in Eastern Europe, he continued to utilize these documents and gradually integrated item with his own materials. His position as one of the secretaries to Pope Sixtus V enabled him to expand his archive further, incorporating additional documents, particularly those related to Poland, when he was appointed in 1587 as the pope's envoy to the region.
In 1592, Pope Clement VIII appointed Graziani as Bishop of Amelia in Umbria. He spent three years in the city before being tasked by the pope with organizing an anti-Turkish league. The success of this mission led to his appointment as nuncio to the Republic of Venice in February 1596. By mid-1598, exhausted and suffering from gout, Graziani was granted an exemption from his duties as nuncio. He subsequently retired to Amelia, where he dedicated himself exclusively to his episcopal responsibilities. He passed away in that city on 16 March 1611.
Following his death, his political archive was supposed to be transferred to the Holy See by his heirs. However, in practice, this extensive collection – containing valuable documents spanning over fifty years (from the 1560s until his death) on sensitive diplomatic relations between the Papacy and various European sovereigns – remained in the possession of the Graziani family, despite repeated efforts by the Holy See to reclaim it.

The archive remained housed in the Graziani residence in Città di Castello, near Perugia. Over the course of the eighteenth century, some manuscripts found their way into the possession of the Spanish-born Jesuit Girolamo Lagomarsini. These manuscripts later became part of the historical archives of the Jesuit *Collegio Romano*, now the *Pontificia Università Gregoriana*. Lagomarsini, who taught Greek at the college until his death in 1773, was deeply attached to its library. His strong opinions about its structure and management are

reflected in manuscript 1487, written in 1758, in which he expressed sharp criticism of the library's organization (Al Kalak 2024, 675-91).[5]

In 1864, a sales catalogue of parchments and manuscripts from the Graziani family archive in Città di Castello was published in Florence. The vendors were identified as "the nobles Niccolò and Teresa Libri", and the catalogue was compiled by Pietro Berti (1864). The Libri family were indirect heirs of the Graziani lineage. In fact, Antonio Maria Graziani's branch became extinct at the beginning of the eighteenth century, after which the archive passed to another branch of the Graziani family. A generation later, the last member of that branch, Porzia Graziani, married Guido Guerra. Four generations later, the last descendant of the Guerra family, Teresa, married Niccolò Libri.

Although the sale did not ultimately take place, some materials were likely dispersed and found their way into various Italian public libraries. Meanwhile, the archive came into the possession of Giovanni Magherini (1852-1924), a local historian with a keen interest in its contents, who had married Melania, the daughter of Niccolò Libri.[6] Magherini expanded the collection by adding documents and manuscripts related to the history of Città di Castello.

By 1904, the archive remained largely intact, as indicated in the description provided by the librarian Giuseppe Mazzatinti in his monumental work *Archivi della storia d'Italia*, published that same year (Mazzatinti 1904). At that time, the collection had been divided into two distinct sections:[7]

1. The *Graziani Archive*, which comprised 143 parchments (dating from 1232 to 1624) and 388 additional items, including manuscripts, files, and bundles of records;
2. The *Magherini Graziani Archive*, which contained 194 archival units.

Subsequently, both collections were transferred to Florence. On 3 January 1941, Niccolò Magherini Graziani formally reported the presence of these two archives in his residence in Palazzo Roffia situated in Borgo Pinti to the Prefect of Florence.

Due to a dispute between Giovanni Andrea Magherini Graziani (1907-1975) and his brother-in-law, Niccolò Mels-Colloredo (1897-1966), the family decided to formally divide their property. By

5 See also https://www.unigre.it/archivioimg/APUG_Topografico/1487_1.jpg and MANUS online, record identification number: CNMD\0000224330.

6 http://www.san.beniculturali.it/web/san/dettaglio-soggetto-produttore?id=10909.

7 The family archive is a separate archive that keeps all the family records from the thirteenth to the twentieth century.

1943, the Graziani collection had already been returned to Città di Castello (albeit without a portion of the collection – mostly Commendone's and Graziani's correspondence – that remained in the hands of Giovanni Andrea Magherini). In 1996, the Ferri Graziani family, heirs of Maria Teresa Magherini Graziani, relocated the Graziani archive to their residence in Vada (Livorno).

The Magherini Graziani collection (along with the Commendone-Graziani correspondence) was transferred from Florence to the Villa di Poggitazzi in San Giovanni Valdarno (near Arezzo). On 26 May 1944, the latter was officially declared to be of significant historical interest. However, that same year, the collection suffered damage when the villa was occupied by German forces. In 1953 and 1954, inspectors from the Superintendency conducted multiple visits to the villa, which was soon to be sold. By that time, little remained of the archive, apart from a few scattered pieces. Further investigations in the 1980s, prompted by reports of documents being offered for sale by the antiquarian Perlini of Arezzo, revealed that between the 1950s and 1960s, the Magherini Graziani collection had been dismembered. Some of its materials resurfaced at the University of Kansas, at least one document was identified in the New York Public Library, and other parts were dispersed among various institutions and private collectors.

The Graziani collection at the University of Kansas is a substantial assemblage of letters and letter-books, consisting largely of correspondence to or from Antonio Maria Graziani and Giovanni Francesco Commendone. It also includes documents, reports, historical texts, and notes, many of which appear to have belonged to Antonio Maria Graziani. The majority of these materials, which primarily concern Commendone's nunciature in Poland, were acquired through the Polish-American bookseller Alexander Janta between 1967 and 1971. Additional items were obtained from the Institute for Canon Law at Boalt Hall, University of California, Berkeley, and from the *Libreria Antiquaria Mediolanum*.[8]

The Graziani initial archive was a family archive that, at a certain point, became integrated with the political archive of one of its members, bishop and nuncio Antonio Maria Graziani. The latter, in turn, incorporated parts of the political archive of Commendone and other high-level prelates transforming this collection into a valuable collection jealously kept by the family. Over the centuries, the archive has been modified and reorganized, partially dismembered

8 I thank Elspeth E. Healey, Special Collections Curator at the Kenneth Spencer Library, University of Kansas, for this information presented in the international conference, "La Chiesa di Roma e l'Europa multiconfessionale nella prima età moderna: attori, politiche, esperienze" (Parma, 17-19 aprile 2024).

and scattered due to family disputes, and ultimately preserved – albeit with unquantified losses – across multiple institutions.

3 Scope(s) of the Project

As previously noted, the three-year, nationally funded project *Nuncio's Secret Archives: Papal Diplomacy and European Multidenominational Societies Before the Thirty Years' War* serves here as a case study for the historian's digital laboratory.

When undertaking a Digital Humanities project that involves data extraction from historical narratives or correspondence, researchers must recognize that multiple approaches exist for processing texts and presenting results. Moreover, existing textual analysis techniques do not fully align with the specific needs of historical research, which prioritizes establishing facts, describing events, analysing complex short- and long-term developments, and integrating findings within the broader framework of scholarly inquiry. Therefore, the first consideration must stem from the project's historically driven research objectives.

In the case of the *Nuncio's Secret Archives*, the stated objectives were twofold:

1. enhance this exceptional and stratified private political archive through the creation of an 'open access' research portal Graziani Archives, which virtually unites the different sections of the original archive (today in Italy, in the US and in Poland), allowing scholars to use them in an integrated and full-scale manner;
2. make an innovative contribution to the understanding of the early stages of papal diplomacy in the aftermath of the Peace of Augsburg (1555) through a complete divulgation of Graziani's 'secret' archives, bringing to light an extraordinary and unexpected range of perceptions, knowledge, and orientations developed by Roman mediators in the face of heresy and the European multi-denominational space.[9]

These two objectives were complementary; however, the project's abstract made it clear that the primary focus would be the creation of the Graziani Archives portal as a research instrument:

9 PRIN: PROGETTI DI RICERCA DI RILEVANTE INTERESSE NAZIONALE – Bando 2017, Prot. 2017JMPYTA, part A, p. 4. The project commenced in March 2020 with an initial duration of three years. However, due to the COVID-19 pandemic, the deadline was extended by an additional twelve months, culminating in March 2024 with the launch of the Graziani Archives portal. The project's concluding conference and the official presentation of the portal were held from 17 to 19 April 2024 at the University of Parma.

> Through a pioneering approach integrating Digital Humanities and traditional research, the project aims to reorganize for the first time an extraordinary private archive, now divided between Italy and Kansas, decisive for reconstructing the history of papal diplomacy in the crucial period between the peace of Augsburg (1555) and the Thirty Years War. The archive, created by two of the most high-ranking diplomats of the late sixteenth century, G.F. Commendone and A.M. Graziani, will be valorized through the portal Graziani Archives, which, in addition to providing a reference model for the interrogation and study of private political archives of modern age, will give access to unpublished documentation, mostly informal and very different from the official and already known one, describing with unprecedented richness and depth the vast networks activated by papal diplomats, their relationship with the regular orders and the clash between papacy and multi-confessional space. The construction of the portal, combined with a rigorous historical-archival investigation, will activate a research laboratory that will give an impulse to a new interpretive perspective, capable of finally including the history of papal diplomacy in the European historiographical debate on multi-confessionality.[10]

The project thus aimed to render the extensive collection of records preserved in the correspondence of Commendone and Graziani accessible through the creation of an online, open-access research portal, Graziani Archives. The research team emphasized in the proposal that this tool would not merely replicate the functions of traditional resources such as paper catalogues, inventories, or thematic bibliographies. Instead, it was designed as an advanced, innovative instrument that would serve as a reference point for the enhancement, interrogation, and study of private political archives from the early modern period.

More specifically, the portal aimed to guide scholars, as well as any interested users, through six distinct yet interoperable search pathways (or sections) accessible from the homepage:

1. *Lettere* – The correspondence of Commendone and Graziani;
2. *Inventari* – A virtually reunified inventory of the Graziani Archives, currently dispersed across multiple institutions in Europe and the United States. This section provides the history and present location of each manuscript, as well as the full inventory of the Graziani family archive (held separately from the nuncio's archive) that shed light on the history of the latter archive;

10 Citation from the project's proposal: PRIN: PROGETTI DI RICERCA DI RILEVANTE INTERESSE NAZIONALE – Bando 2017, Prot. 2017JMPYTA, Part A, p. 3.

3. *Persone* – A database of individuals mentioned in the correspondence and their relationships and titles or occupation;
4. *Luoghi* – A geolocated visualization of the places referenced in the correspondence, displayed on an interactive Google map;
5. *Viaggio di G.F. Commendone* – An interactive map tracing Cardinal Commendone's journey (1560-62) to the bishops and princes of northwestern Europe, undertaken to announce the convening of the Council of Trent. This section includes biographical profiles of individuals he encountered, images of key figures and locations, and textual excerpts drawn from both the diary of his entourage member, the Bolognese Fulvio Ruggieri, and Commendone's own dispatches to Rome;
6. *Biblioteca di A.M. Graziani* – A digital reconstruction of Graziani's personal library.

Figure 1 Homepage of the Graziani Archives portal

When a team initiates a project – meaning when it undertakes the realization of a proposal – it must clearly define the 'who', 'when', and, most importantly, the 'how'. It is useful to consider that "a process is a sequence of tasks [...], through which an output is produced by operating on an asset" (Dunn, Hedges 2012, 21). For instance, if a process involves cataloguing, the 'who' refers to individuals with expertise in cataloguing. The 'when' depends either on the project's established deadline or on the internal timeline set by the team, particularly as other collaborative tasks may be conducted in parallel. However, the most critical aspect is the 'how' – the organization of various processes into complementary tasks within a coherent and feasible timeline.

Given the diversity of institutional holdings, the lack of information regarding the whereabouts of certain materials that were once part of the archive, the time constraints of the funded project, and the financial and human resources available to the team, several critical decisions had to be made by the *Nuncio's Secret Archives* project team:

1. Project Scope – *prioritizing research output or adopting a broader approach to address the research needs of related user communities*.
    - While the production of research results is a fundamental requirement of any funded project, the predetermined timeline often limits the possibility of an online publication of the structured data collected, processed, and utilized during the research, thereby restricting access for other scholars. It seemed to the team that it should take into consideration that the creation of an online portal takes longer than the publication of a book. It was therefore advisable to adopt a realistic approach toward the amount of material and work needed for the realization of the web portal;
2. Digital Tool – *developing a custom-built tool tailored to the team's specific needs or utilizing an existing, more generic solution*.
    - Digital tools for storing structured data range from simple Excel files to complex online platforms designed to establish relationships between different properties, apply markup tools to annotate uploaded digital images, or integrate properties into interactive maps. The selection of an appropriate digital tool depends primarily on financial and time constraints, as well as the desired level of online accessibility and interaction;
3. Online Open-Access Communication – *choosing a website or platform capable of providing a user-friendly interface*.
    - The more a project aims to publish collected data through an intuitive, user-centric interface, the greater the need for a platform that can efficiently generate a seamless frontend from the backend module used for data ingestion and relational structuring;
4. Long-Term Preservation – *ensuring the maintenance and updating of both the digital tool and the data in a suitable repository*.
    - Currently, most projects require a three- to five-year Data Management Plan and accept the online publication of data in formats such as Excel or Comma-separated values (CSV) in an open-source repository. However, this approach often leaves the software or platform used for

storing linked data without further financial support for maintenance and updates beyond the project's duration. As a result, while the data remains searchable at the item level, the relationships between properties become inactive, thereby limiting the potential for advanced research applications.

4 Choice of the Right Platform

The option of building an online database from scratch to meet the project's specific requirements was immediately ruled out by the research team, as it was deemed too costly and time-consuming for a three-year project primarily focused on analytical deliverables.

A publicly funded project must adhere to a set of principles established within a working framework developed by decision-makers at the national and, in this case, European levels. The *European Science Agenda* identifies three distinct levels of data science: data, services, and governance (Ganguly, Budroni, Sánchez Solís 2017, 203-10). These levels are supported by digital infrastructures, which form the foundation for data preservation by ensuring that data are managed and curated effectively. The development of services for data uploading and reuse is based on the principle that infrastructure serves as a cornerstone, aligning with the European Commission's approach to structuring these three levels. Services must be user-friendly and adaptable to various use cases, as defined by data producers. Lastly, governance provides the overarching framework through which functional and publicly articulated policies establish an institutional format for data preservation.

These three levels guide the technical choices of any project, requiring a research team to identify a solution that ensures open access, a user-friendly platform, and a high degree of data preservation. Given the range of available open-source tools capable of combining a content management system with a flexible organization of item-level metadata (Hardesty 2014, 80-4; Tritch Roman 2018), the *Nuncio's Secret Archives* project's team initially explored the possibility of using the PHAIDRA platform.

PHAIDRA (Permanent Hosting, Archiving, and Indexing of Digital Resources and Assets), as stated in its mission, helps organizations manage and preserve the full value of their digital assets: from diverse media files, digitized artifacts, and collections; to research, data, and analysis".[11] Built on Fedora open-source software,

11 https://phaidra.org/. Other similar free and open-source digital preservation systems are DSpace (https://dspace.org/) developed in 2002 by the Massachusetts

which offers a modular architecture,[12] the platform was first developed in 2008 by the ZID – Zentraler Informatikdienst and the Library of the University of Vienna[13] to promote the principles of Open Science, Open Data, and Open Access (Montanaro 2016-17, 12).[14] To enhance interoperability with other platforms, PHAIDRA soon collaborated with Europeana, OpenAIRE, and OpenAIREplus, adopting the FAIR principles[15] and the Data Management Plan (Miksa et al. 2016) as its guidelines. Additionally, it established a network with academic institutions in Austria, Italy, Serbia, Montenegro, and Bosnia-Herzegovina.[16]

The decision to use PHAIDRA as a repository for the project's data was further supported by the fact that both the University of Padua and the University of Venice, which hosted research units involved in the *Nuncio's Secret Archives* project, were also partners of the PHAIDRA platform.[17] Moreover, after reviewing the detailed documentation on PHAIDRA provided online by the Library System of the University of Padua, the research team concluded that the data ingestion process was relatively straightforward and could be mastered with the help of a brief training seminar.

However, upon closer examination of PHAIDRA and after discussions with the University of Padua librarians responsible for managing the platform, several concerns emerged regarding its suitability for the project. PHAIDRA was primarily designed for the reconstruction of digital libraries and archives, ensuring their accessibility and long-term preservation. Consequently, its data architecture was centred primarily around digital objects and their metadata (Cappellatto 2015-16, 55). By contrast, the *Nuncio's Secret Archives* project pursued different objectives. It sought to extract data from a selected body of correspondence, enrich that data, establish relationships with other sources – such as the Graziani inventories – and ultimately use part of the dataset to design an interactive map narrating Cardinal Commendone's journey (1560-62) to the bishops and princes

Institute of Technology (MIT) and Hewlett-Packard Labs (Lewis et al. 2010) or Archivematica (https://www.archivematica.org/en/docs/archivematica-1.17/getting-started/overview/intro/#intro), developed by Artefactual Systems and University of British Columbia Library in 2012 (Sprout, Romkey 2013, 257-68).

12 https://fedorarepository.org/. Fedora is the acronym of Flexible Extensible Digital Object Repository Architecture.

13 https://phaidra.univie.ac.at/.

14 The PHAIDRA project followed the UNESCO Memory of the World report: Bradley, Lei, Blackall (2007, 3), that argued for simplification in the digital and archival preservation system.

15 https://www.go-fair.org/fair-principles/.

16 https://phaidra.org/community/overview/.

17 https://phaidra.cab.unipd.it/.

of northwestern Europe to announce the convening of the Council of Trent.

The exclusion of PHAIDRA as an optimal tool for achieving the project's objectives led to a reassessment of its priorities. It became evident that the chosen solution should rely on an existing, free, open-source platform capable of describing linked properties while adhering to established data standards. Following a brief but targeted search, the research team determined that the most suitable solution was Omeka S, a platform released in 2009 by the Roy Rosenzweig Center for History and New Media, funded by multiple organizations.[18]

This choice was further reinforced by the fact that Omeka S meets the Data Management requirements set by funding institutions. It offers modules for importing resources from repositories such as Zotero, Zenodo, Fedora, DSpace, CKAN, Dataverse, and Invenio and allows data export in multiple formats, including CSV, JSON-LD, JSON-table, ODS, TSV, and TXT (Maron, Feinberg 2018) – ensuring long-term data preservation (Morton 952-3).

Although Omeka S has been criticized by some Library and Information Science (LIS) researchers for not adequately clarifying and supporting the Dublin Core metadata standard on which it relies (Maron, Feinberg 2018), it is important to recognize that Omeka S was not designed solely for the description of properties and values from a LIS perspective. Rather, it was conceived as a "free, open-source platform that, like blogging software, offered an easy-to-use administrative interface, provided syndication for sharing content, and extended the core function of publishing content with a flexible plugin architecture and rich design theme API".[19]

From a historian's perspective, this flexible approach is particularly advantageous for several reasons. First, it enables non-LIS experts to utilize the platform effectively, leveraging their expertise in other fields to develop highly complex projects based on collections that may not conform entirely to standardized classification systems. Second, it allows users to incorporate ambiguous expressions that are typically restricted by the Dublin Core's rigorous approach, thereby avoiding forced disambiguation in cases where insufficient evidence is available. For example, the Dublin Core's 'levels of granularity', which seek a clear, consistent, and standardized format, struggle to provide a satisfactory solution for date values such as *circa 1940*. Should this date be disambiguated on a decadal basis (e.g., 1930-40 vs. 1940-50) or on a yearly basis (e.g., 1939-40 vs.

18 https://omeka.org/s/.

19 https://omeka.org/about/project/.

1940-41) (Maron, Feinberg 2018, 682)? Historians are usually very attentive to these kinds of problems.

Finally, Omeka S also facilitates the extraction and contextualization of complex, incomplete, or even contradictory data from multiple sources while ensuring compliance with Dublin Core standard metadata.

5 Data Architecture

The research team firmly believed that an archive's contents could be effectively represented through dataset models and the creation of essential descriptive categories, facilitating a certain degree of cross-searchability both within and across institutions. This conviction led to the selection of the Omeka S platform. A key challenge was transforming content into a new form of information infrastructure that is user-centred and capable of supporting both content management tasks and those related to communication and collaboration.

As previously explained, the Omeka S platform prioritizes display and utilizes an unqualified Dublin Core metadata standard (Maron, Feinberg 2018). It accommodates multiple sites that draw from a shared pool of resources and offers considerable flexibility, allowing users to design their own projects.

The following discussion provides an overview of our approach to modeling information on the backend. Before determining the appropriate data architecture model, it is essential to conduct a thorough examination of the content to be structured. The process can be better understood through the framework of the Graziani Archives portal.

The first section of the portal concerns a collection of more than 3,000 letters, primarily related to the diplomatic activities of Commendone and Graziani within the Holy Roman Empire and Poland. Another significant portion consists of letters addressed to them by various correspondents. This body of correspondence was perhaps the most crucial focus of the project, as it promised to

> give access to unpublished documentation, mostly informal and very different from the official and already known one, describing with unprecedented richness and depth the vast networks activated by papal diplomats, their relationship with the regular orders and the clash between papacy and multi-confessional space.[20]

20 See note 9, *supra*.

A preliminary review of the letters and dispatches yielded several observations further underlining the complexity of the correspondence content:

1. the majority of the letters follow a recurring pattern, in which the senders report information obtained during their encounters with envoys, prelates, high-ranking officials, or sovereigns, specifying the source of each piece of information. Alternatively, the information is derived from other sources, such as letters or hearsay;
2. the content of the letters pertains to a wide range of individuals, primarily within the political or ecclesiastical spheres, either in Rome or in the territories of the Holy Roman Empire and Poland-Lithuania;
3. some letters contain references to past or unfolding events that, within the context of a single letter, are not always easily identifiable or fully understood;
4. the correspondence mentions various locations, often linking officeholders, prelates, or other individuals to specific places;
5. neither Omeka S nor any other database based on the Dublin Core metadata vocabulary, the RDF data model, or the more recent Records in Contexts-Conceptual Model (RiC-CM)[21] can adequately capture the richness and complexity of the Commendone-Graziani correspondence as a form of historical narration.

These observations, along with the awareness of time and resource constraints, led to the following rationale: given that the research team's primary objective was to provide users with access to this source, and considering that the primary users would be historians, three distinct yet complementary types of access points were proposed.

1. Full Digitization and Metadata: the first access point involved the complete digitization of more than 3,000 selected letters and their online publication, accompanied by appropriate metadata;
2. Data Extraction and Record Cards: the second access point entailed the extraction of data from each letter, focusing on individuals, places, and organizations. A dedicated 'record card' was created for each entity, along with references to the letter(s) in which it was mentioned. This feature was designed to assist users in navigating the full digitized collection and contextualizing the extracted data. However, tagging extracted properties directly onto the digitized material was deemed

21 https://www.ica.org/resource/records-in-contexts-conceptual-model/.

impractical and overly time-consuming. This decision was influenced by the expertise of the researchers recruited for data extraction and ingestion, who were highly trained historians with foundational knowledge of Digital Humanities but limited familiarity with descriptive markup tools;

3. Summaries for Contextualization: given the density and richness of the historical data, the research team determined that providing users with an abstract for each letter would be valuable. Consequently, most letters were accompanied by a 'short summary' (*Regesto veloce*) highlighting key events, individuals, and meetings. However, for a subset of letters containing particularly detailed information, considerable length, or complex interrelations among people and events, a more 'detailed summary' (*Regesto approfondito*) was also provided alongside the brief one. This additional layer of contextualization was considered essential for enabling users to efficiently identify relevant materials.

At this stage, these decisions needed to be translated into a data architecture model based on Omeka S functionalities and on the conviction that the relationships to create are not only of a vertical type: each and every item is to be linked to others following its attributes.[22] The first critical question concerned the nature of the items and their relationships within the architecture: should 'people' function as a central node, serving as the pivot of the structure, or should 'letters' be treated as the primary item, representing the physical objects whose images would be uploaded and linked to each individual letter entry?

The latter approach was ultimately chosen, as it not only represented a tangible object but also needed to be associated with its corresponding collection and hosting institution. The initial proposal for this structure was then visually represented, as shown in figure 2 (where the 'letters' are represented by 'Document').

22 Although, naturally, the letters had to be linked to their current corresponding collection and hosting institution in a 'tree' like architecture. On the complex problem of archival description in the digital hierarchical environment (Michetti 2013, 1002-10).

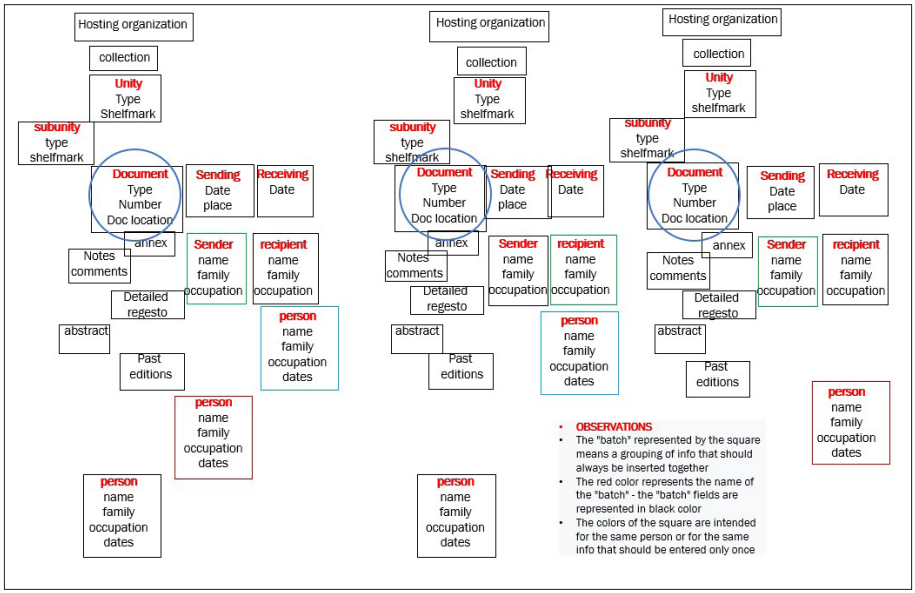


Figure 2 First proposal of data architecture for the *Lettere* section

As observed, the letters were to be linked to their current hosting institutions and collections. Following the prevailing standards of archival description, the data architecture also incorporated additional metadata beyond the hosting institution, collection, and shelf mark, which were essential for each 'letter' item. These included the date and place of sending and receipt, the names of the sender and recipient, and, where applicable, any annexes. All other items were connected either to the letter's contents (such as persons, abstract, or 'detailed summary') or to supplementary information (e.g., past editions of letters, notes, and comments).

An additional step involved the development of resource templates, which were renamed by the research team but corresponded to specific classes based on the Bibliographic Ontology in Omeka S **[fig. 3]**.

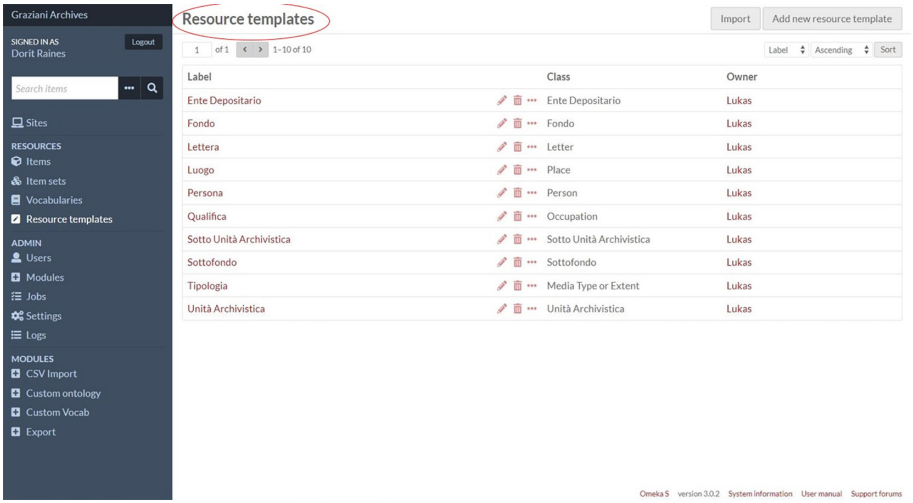


Figure 3 The Resource templates based on the Bibliographic Ontology in *Omeka S*

The creation of resource templates facilitated the final design of fields within each template, drawing on the available modules in Omeka S, including Dublin Core, Bibliographic Ontology, Friend of a Friend (FOAF), and Schema. While some templates were relatively straightforward to create, requiring only a limited number of fields, two proved to be particularly challenging: *Persona* ('Person') and *Lettera* ('Letter'), both of which, as previously noted, were central to the project.[23]

As illustrated in the figure below, the *Persona* template is linked not only to other templates – such as *Lettera*, which establishes familial or marital relationships with other *Persona*, and *Qualifica* ('Occupation') – but also to other sections of the Graziani Archives platform. These include the inventories of the two Graziani archives (Antonio Maria Graziani's archive and the family archive) and the *Itinerario* ('Itinerary') template **[fig. 4]**.

23 One should differentiate between the names given to a portal section and those given to a template: while a section is called *Lettere* or *Persone* (in the plural), the templates were named in the singular form: *Lettera* or *Persona*.

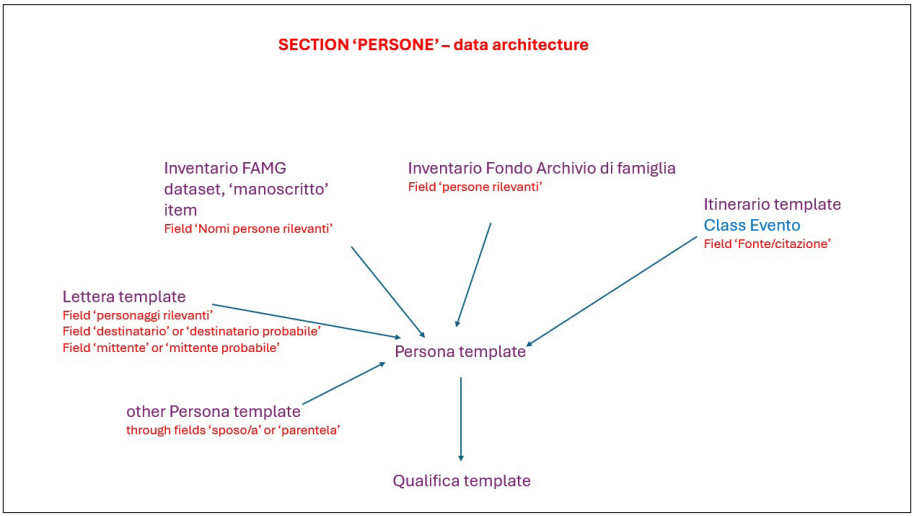


Figure 4 Data architecture for the *Persone* section

The *Lettera* template resulted in an even greater number of connections to other templates. First, it is associated with its current location and shelf mark, including the sub-collection or collection and the archival unit or sub-unit to which it belongs, as well as its present hosting institution. Second, it incorporates metadata specific to each individual letter, such as the names of the sender and recipient, as well as the locations and dates of sending and delivery.

Additionally, the *Lettera* template records all relevant person names and locations mentioned in the letter, along with a specialized vocabulary designed to describe the characteristics of each archival unit. These descriptors include details such as the presence or absence of annexes, whether the document is a copy of the original, whether it is written in cipher, a draft (*minute*), and other relevant classifications **[fig. 5]**.

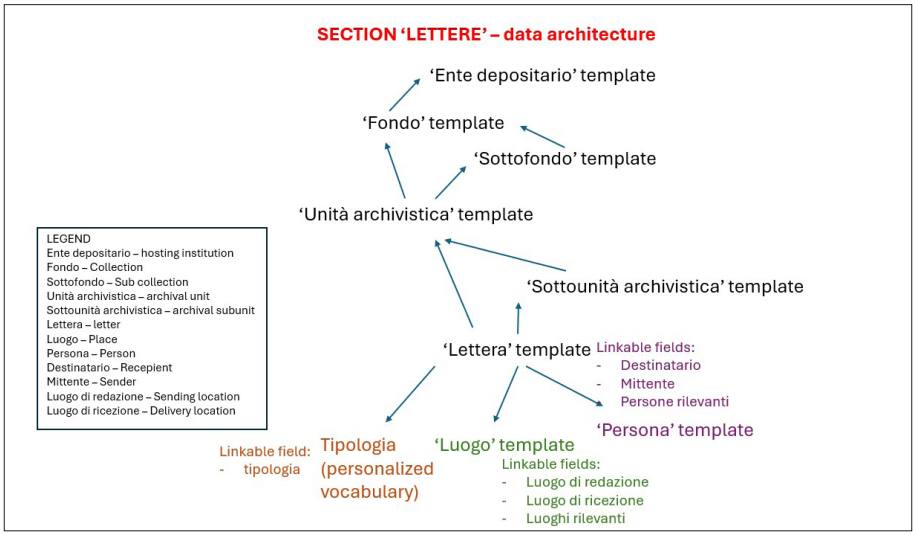


Figure 5 Data architecture for the *Lettere* section

The *Lettere* section on the platform was also linked, through the *Persona* template, to two other sections: *Viaggio di G.F. Commendone* and *Inventari*.

As previously explained, the research team considered the narration of Cardinal Commendone's journey (1560-62) to be an integral part of the project. This journey, undertaken to inform the bishops and princes of northwestern Europe about the convening of the Council of Trent, is reconstructed in an interactive map. The map traces Commendone's travels across Eastern Europe based on a diary attributed to a member of his entourage, as well as on the dispatches he sent to Rome. Each stage of the itinerary is supplemented with biographical profiles of the individuals he encountered, images of people and places visited, and textual excerpts from both the diary and the dispatches **[fig. 6]**.

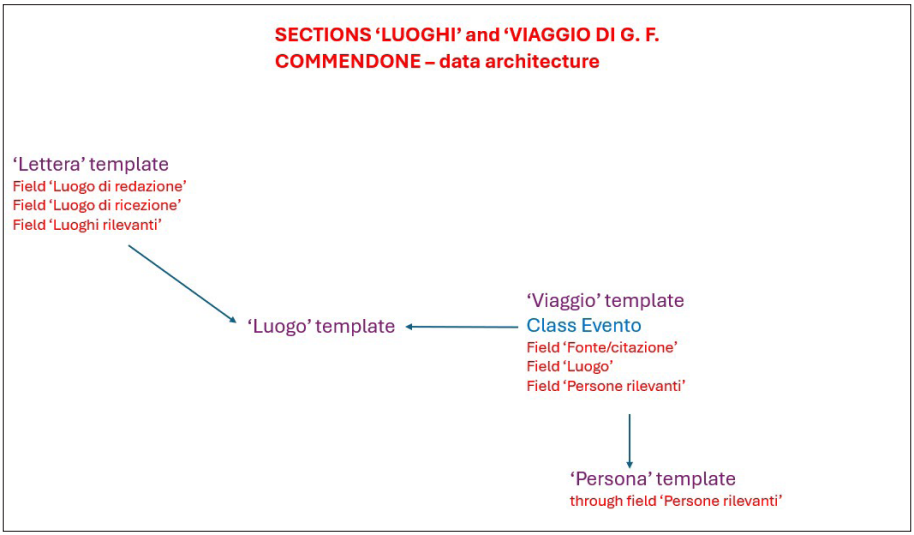


Figure 6 Data architecture for the *Luoghi* and for the *Viaggio di G.F. Commendone* sections

The *Inventari* section presented distinct challenges to the research team. As previously discussed, the current state of the Graziani material is complex, as the collection is housed in three major locations, each associated with a different hosting institution. Additionally, the Vada portion is divided into two separate sections: the first is the Antonio Maria Graziani archive, which is connected to the Graziani manuscripts in Kansas, New York and Poland, while the second comprises the family archive, entirely preserved in Vada.

To provide users with the most comprehensive information possible, the team decided that, in preparing an inventory for the Antonio Maria Graziani archive – one that would virtually encompass all manuscripts historically associated with it – it would be valuable to include links to two key inventory catalogues: the one compiled by Pietro Berti in 1864 and the one prepared by Giuseppe Mazzatinti in 1904. The dual challenge of this archival fragmentation, combined with the objective of virtually reconstructing the Antonio Maria Graziani archive as it existed in the early eighteenth century (prior to the dispersal of its holdings), led to the intricate solution **[fig. 7]**.

The two Graziani sections in Vada were linked through the *Ente depositario* ('Hosting Institution') template, while the Antonio Maria Graziani archive was virtually connected to the Kansas, New York and Polish collections, as well as to the two historical inventory catalogues (Berti and Mazzatinti), via a *Data set* class template named *Inventario integrato* ('Integrated Inventory'). This approach enables

users to determine the past and present locations of each manuscript at any given time.

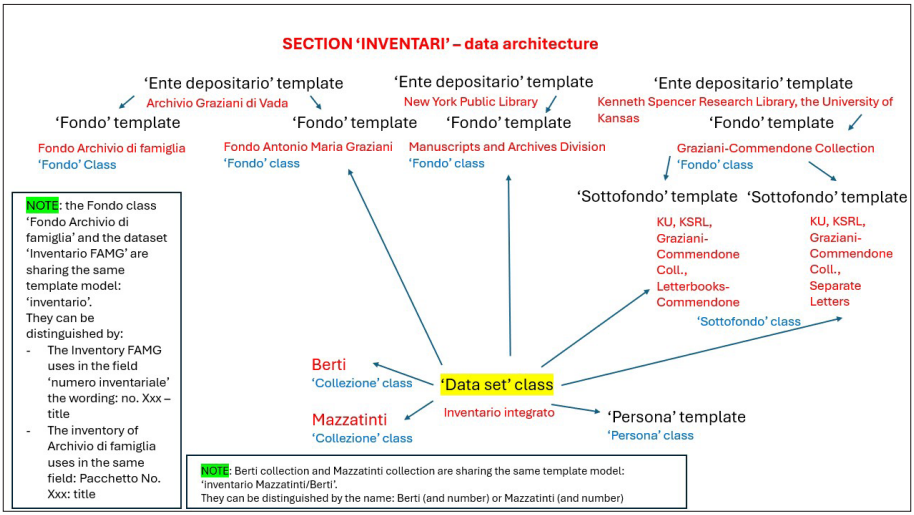
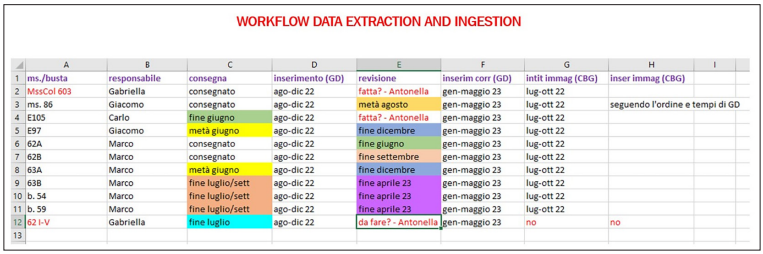


Figure 7 Data architecture for the *Inventari* section

The library section (*Biblioteca di A.M. Graziani*) was addressed separately. The data architecture applied to this section is explained in detail by Luca Iori (*infra*).

6 Data Extraction, Data Ingestion

Parallel to the elaboration of the data architecture for the Graziani Archives portal, the research team set to decide the types of recurrent data to be extracted from the selected Commendone-Graziani letters. As explained beforehand, the decision to provide each letter with a short abstract of its content and in certain cases, to publish also a detailed summary, led the team to focus on the extraction of individuals' names and locations. Four postdoctoral researchers were chosen to read the letters, prepare for each an abstract, and if needed, a detailed summary, locate past (whole or partial) editions of the letter and extract all relevant names and locations. The four were assigned each a number of manuscripts that contained the correspondence and a workflow plan was designed to respect deadlines **[fig. 8]**.



WORKFLOW DATA EXTRACTION AND INGESTION

	A	B	C	D	E	F	G	H	I
1	ms./busta	responsabile	consegna	inserimento (GD)	revisione	inserim corr (GD)	intit immag (CBG)	inser immag (CBG)	
2	MssCol 603	Gabriella	consegnato	ago-dic 22	fatta? - Antonella	gen-maggio 23	lug-ott 22		
3	ms. 86	Giacomo	consegnato	ago-dic 22	metà agosto	gen-maggio 23	lug-ott 22	seguendo l'ordine e tempi di GD	
4	E105	Carlo	fine giugno	ago-dic 22	fatta? - Antonella	gen-maggio 23	lug-ott 22		
5	E97	Giacomo	metà giugno	ago-dic 22	fine dicembre	gen-maggio 23	lug-ott 22		
6	62A	Marco	consegnato	ago-dic 22	fine giugno	gen-maggio 23	lug-ott 22		
7	62B	Marco	consegnato	ago-dic 22	fine settembre	gen-maggio 23	lug-ott 22		
8	63A	Marco	metà giugno	ago-dic 22	fine dicembre	gen-maggio 23	lug-ott 22		
9	63B	Marco	fine luglio/sett	ago-dic 22	fine aprile 23	gen-maggio 23	lug-ott 22		
10	b. 54	Marco	fine luglio/sett	ago-dic 22	fine aprile 23	gen-maggio 23	lug-ott 22		
11	b. 59	Marco	fine luglio/sett	ago-dic 22	fine aprile 23	gen-maggio 23	lug-ott 22		
12	62 I-V	Gabriella	fine luglio	ago-dic 22	da fare? - Antonella	gen-maggio 23	no	no	
13									

Figure 8 Workflow chart for the data extraction and data ingestion phases

In addition to the extraction and ingestion workplan the team had to face another question, currently surfacing when more than one person is working on data extraction: allow each of the four postdoctoral researchers directly ingest the data extracted and the short and detailed summaries into the Omeka S *Lettera* template and eventually correct the verified data in a second moment; or terminate the extraction phase and the verification 'offline' using a shared template and then choose one person for the ingestion phase?

The team preferred the second option for two main reasons:

The short and detailed summaries were first reviewed and corrected, if necessary, by the project's heads of units and subsequently standardized and verified against the extracted data by two editors. Given the time-consuming nature of this process, the team concluded that data ingestion could only be carried out efficiently and accurately once this phase was fully completed;

When a single individual is responsible for data ingestion, they are more likely to identify inconsistencies, missing standardizations, or unresolved ambiguities in the data. To address such issues, the postdoctoral researcher selected for the ingestion part constantly consulted one of the units' heads responsible for standardization to resolve any outstanding questions. Indeed, during the ingestion phase, the designated researcher identified missing or conflicting information which led to the addition of several fields in the template to clarify ambiguous identifications, relationships between individuals, and gender attributions.[24]

To ensure consistency in data extraction, a working template was created using standardized vocabulary and provided to the four postdoctoral researchers. This template was designed to align with the structure of the Omeka S template for letters, thereby facilitating the subsequent ingestion process **[fig. 9]**.

24 For a discussion of this issue, see Gabriella Desideri, *infra*.

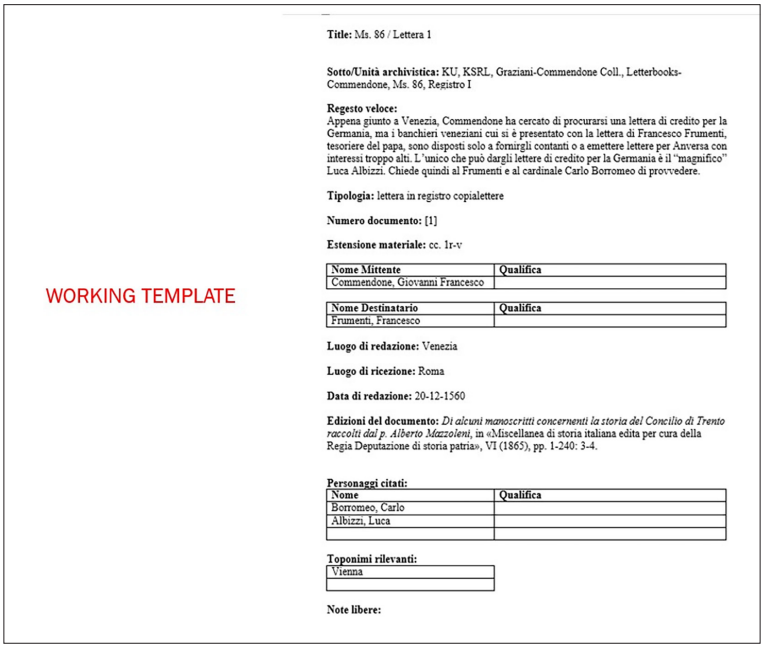


Figure 9 Working template for extracting data from the letters

The team recognized that, in certain cases, disambiguation was necessary for specific individuals or locations. Additionally, ensuring consistency in name standardization was particularly challenging given the involvement of four researchers, necessitating a structured approach to team coordination in order to maintain controlled data usage.

To maximize efficiency and accuracy, a detailed workflow was developed. As explained above, one of the project's unit heads was designated as the standardization editor, and two shared Word files were created in Drive – one for individual names and the other for locations. Each postdoctoral researcher was responsible for entering standardized names in alphabetical order, using an assigned identification colour. The editor would then review the entries, verify them, and either approve the standardized names by marking them in black or flag them for further investigation by highlighting them in yellow **[fig. 10]**. Each researcher was subsequently responsible for updating the working template with the verified standardized names.

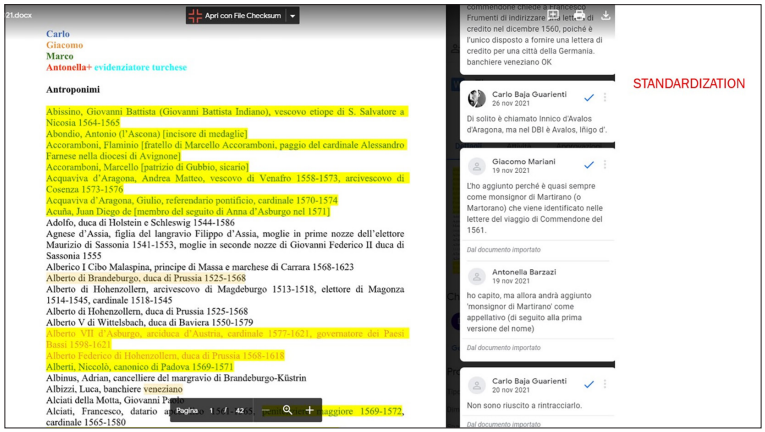


Figure 10 Shared file for individual names' standardization

The outcome of this labour-intensive process was a relatively smooth and efficient ingestion phase, which, in turn, facilitated the rapid uploading of images. Each image had been pre-named according to a standardized format, incorporating the manuscript shelf mark, the letter number, and a sequential number corresponding to its position within the letter unit (e.g., if a letter spanned three pages, the images would be numbered 1, 2, 3, etc.).

The preparation of the *Lettere* section proved to be the most complex and time-consuming phase of the project. In parallel, work began on the *Inventari* section; however, its ingestion had to be postponed until the completion of both the *Lettere* and *Persone* sections, as many individuals mentioned in the inventories were referenced in both sections. The preparatory phase for the inventories ingestion required an initial selection of individuals named in the inventories but not mentioned in the letters, the standardization of their names, research into their titles, occupations, and relationships, and only then the ingestion of this information into the Graziani Archives portal.

7 From Backend to Frontend

A word of caution is necessary for this section. My aim is not to explain the distinction between backend and frontend from an application perspective, nor to provide a technical overview of the programming languages involved. Rather, this is a straightforward humanistic reflection on how humanities scholars may navigate this complex phase and effectively communicate their preferences to the custom web designers responsible for implementation. This account

chronicles the experience of the Graziani Archives team, with the hope that it may assist other teams in transforming their ingested data into a user-centric and user-friendly platform.

Upon completing the data ingestion phase, it became necessary to shift our focus toward the publication phase. As humanists, we are typically well acquainted with the process of print publication; many of us have even learned to format texts using various software programs. However, web publication, and above all – custom web design – presents an entirely different challenge, requiring a distinct approach.

What, then, is the fundamental difference between backend and frontend? In our case, given that we were using Omeka S, the infrastructure was already in place – it merely needed to be adapted to our data architecture based on the predefined sections (as previously discussed). The backend serves two primary functions: first, to enable researchers to ingest data in a structured and retrievable manner; and second, to store this data and facilitate user access. In Omeka S, these functions are performed through a set of modules, each containing multiple fields that the team created according to its needs, along with the possibility to establish relationships between different fields or modules.

From a graphical perspective, each module is designed in the backend in a vertical continuous layout, which can vary in length depending on the number of fields included. Researchers must use the mouse wheel to scroll up and down to access all fields. When establishing a relationship between data elements, a new sidebar is displayed on the right side of the screen, introducing all existing items and allowing for further selection by category to locate the desired item. The team found this design intuitive and worked comfortably with it throughout the ingestion phase.

However, when we were asked to conceptualize a frontend representation of the data we had ingested, we found ourselves at a loss. We were unsure of what was expected of us and unfamiliar with the distinction between backend and frontend. It was at this moment that we realized our primary focus had been on the ingestion process, with little consideration given beforehand to the user experience.

At this stage, we had the support of three professionals: the graphic designer, who had previously designed the project's web pages for the University of Parma (the project's principal investigator), as well as custom web designers from Libnamic Digital Humanities, who implemented the proposed graphic design into a web customization using Figma, a collaborative design tool.[25]

25 https://www.figma.com/.

Before delving into this process, it is useful to take a step back. In many projects, teams first create a webpage or website to communicate the project's objectives, ongoing progress, events, and publications. The design of such a webpage is typically entrusted to a graphic designer and is often influenced by whether it is developed within an existing platform (e.g., a university platform) and whether the platform's administrators have already established fixed layouts, graphic rules, and colour schemes. Fortunately, in our case, the University of Parma permitted the creation of a website while allowing the team full discretion over its layout.

The graphic designer initially proposed a set of colours and fonts for both the project's website and its event brochures. Two years later, when discussing the graphic design of the Graziani Archives portal, she recommended using a different colour scheme to distinguish the portal from the project's main website. We found this rationale compelling and readily accepted the suggestion. Consequently, the team collaborated with her to design the various sections of the portal. The final proposal was then submitted to Libnamic for the web customization phase.

To illustrate the final outcome of the customization phase, I would like to summarize key points that, while seemingly straightforward, may assist other scholars when customizing the frontend of their database. Given that we are not web designers, I wish to highlight several considerations that a humanities team should take into account when approaching the web design and customization phase:

1. Adopting a user-centric approach – It is essential to anticipate the needs of future users, categorizing them – if possible – into groups with distinct objectives and predicting their behaviour. In short, the goal is to effectively communicate what is happening on the screen and to enhance the usability of the portal by minimizing the number of clicks required to access the desired content;
2. Considering screen orientation – Unlike traditional A4 paper formats, where the longer dimension is vertical, computer, laptop, and tablet screens typically have a horizontal orientation. While scrolling allows users to access additional content, the customization process should ensure that the horizontal screen layout immediately conveys the overall content structure of the webpage and enables users to navigate efficiently;
3. Designing the initial portal webpage with distinction – The initial portal webpage may differ slightly from the other pages within the portal. In the case of the Graziani Archives, the goal was to create a visual link between the project's webpage on the University of Parma's website and the portal itself. To achieve this, a distinctive graphic element – a bluish

Figures 11-12
Above the NSA Parma University's project website. Below the Graziani archives initial website page

image of the Graziani manuscripts with the portal's name superimposed in white – was used as a visual bridge between the two websites. This graphic element appears only on the initial portal page, as it occupies significant space that would otherwise be allocated to content **[figs 11-12]**;

4. Signage is essential – Users must always be aware of the portal's name, the menu, and their current location within the portal's sections. This requires the portal header to remain consistent across all pages, displaying the logo and portal name at the top of the screen, with the menu positioned either directly below or on the left or right side. Visual hierarchy is equally important: "Use different font sizes, colours, highlighting effects, etc. to help users distinguish categories of menu items or levels of importance" (Toulson 2021);
5. Menu bar positioning – There are two primary approaches to menu positioning. If the goal is to maximize the space below the logo and dedicate it to the display of content, a left-side menu is preferable. However, if the layout is divided into two or more horizontally aligned columns – as was the case for the Graziani Archives portal – a top menu bar placed directly below the portal logo is a more effective solution. In any case, the header should not occupy more than 25-30% of the initially visible screen space. Users should be able to immediately recognize the presence and nature of the webpage's content;
6. Providing a concise summary of portal content on the homepage – Displaying a brief overview of the portal's content on the initial screen is crucial. Users should be able to grasp the essential information about the portal and its sections at a glance, without needing to navigate between sections to understand their purpose. More detailed descriptions of each section can be provided on their respective landing pages **[fig. 13]**;
7. Dividing the screen into left and right columns – Displaying long passages of text on a horizontally oriented screen may not be the most effective solution, as extended line lengths and high text density can hinder readability. A more user-friendly approach is a split screen: dividing the content into left and right columns. Additionally, if two different types of information need to be presented on the same page, organizing them into columns is preferable to requiring users to scroll down to locate the secondary content. A relevant example is the *Lettere* section of the Graziani Archives portal. In this section, the left-hand column provides a detailed explanation of the section and the materials available, while the right-hand column presents a list of selected manuscripts with two navigation options. Users can either click on the + symbol (located to the left of the manuscript number) to

Figure 13
The Graziani archives initial website page with titling, menu bar, a short explanation of the portal's contents and the short text of the first section

reveal a dropdown displaying the manuscript title and editorial notes, or they can click directly on the manuscript number to access its main page **[fig. 14]**. On this page, the left-hand column displays the list of letters, while the right-hand column contains the full manuscript description. Furthermore, on the manuscript page, at the top of the right-hand column, a navigation widget allows users to browse through the selected manuscripts directly, without the need to click in and out of individual records **[fig. 15]**. This same navigation logic applies to each specific letter's page as well as to all other sections of the Graziani Archives portal, where users can seamlessly move back and forth between the items **[fig. 16]**;

Figure 14 *Lettere* section with a manuscript title and editorial note displayed in the right-side drop-down column

Figure 15 Manuscript page in the *Lettere* section: on the left column the list of letters; on the right column the full manuscript description and on the upper side, the possibility to shift to the next manuscript (in blue)

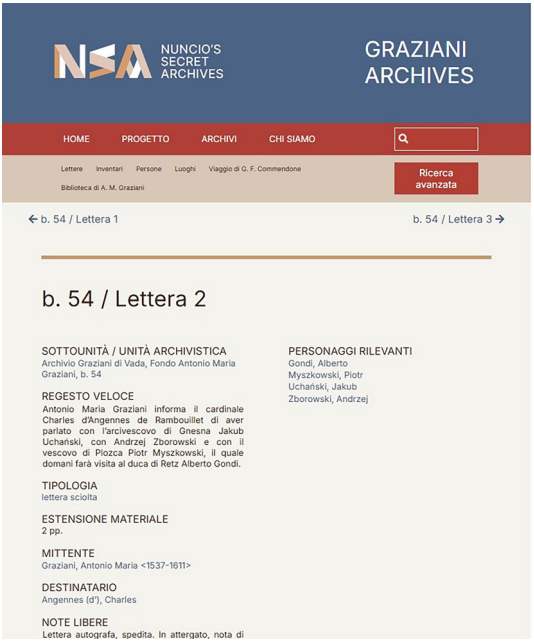


Figure 16 Letter page in the *Lettere* section displaying the possibility to shift to the previous or next letter (top right and left columns – in blue)

8. The Use of Colour – As previously noted, visual hierarchy can significantly enhance user experience. In our case, the graphic design aimed to distinguish the project's website from the Graziani Archives portal by implementing a structured colour scheme. Four distinct colours were used for the Graziani Archives portal: dark electric blue as the background for the logo and portal name; medium carmine red for the top bar, which contains the primary menu presenting the project, its objectives, and the people involved; and pale silver for the secondary menu, which provides access to the content-specific sections of the portal. The fourth colour, a pale grey-yellow (*Isabelline*), was selected as the background for the text. It is crucial to choose a 'neutral' background colour that contrasts well with the text colour, ensuring readability without causing visual strain. Additionally, to help users navigate the relationships between different items, we used dark electric blue to indicate clickable elements – such as persons, places, letter types, hosting institutions, collections, manuscript numbers, and letters – allowing users to access further information about each entity.

Finally, a note on the customization of the interactive map.[26] As previously explained, the objective of this section was to visualize the diplomatic journey undertaken by Cardinal Commendone between 1560 and 1562, during which he travelled across northwestern Europe to inform bishops and princes of the convening of the Council of Trent. This was achieved by geolocating and marking the sites on a Google map he visited and, in selected cases, attaching descriptions of the individuals he met, and, where possible, images of his interlocutors or contemporary maps of the locations.

The primary challenge was to geolocate 229 sites on the map while maintaining a coherent representation of the entire itinerary – covering all of Europe – without requiring a zoomed-in view. Additionally, the map needed to clearly indicate the direction of travel and categorize locations into six distinct geographical areas.

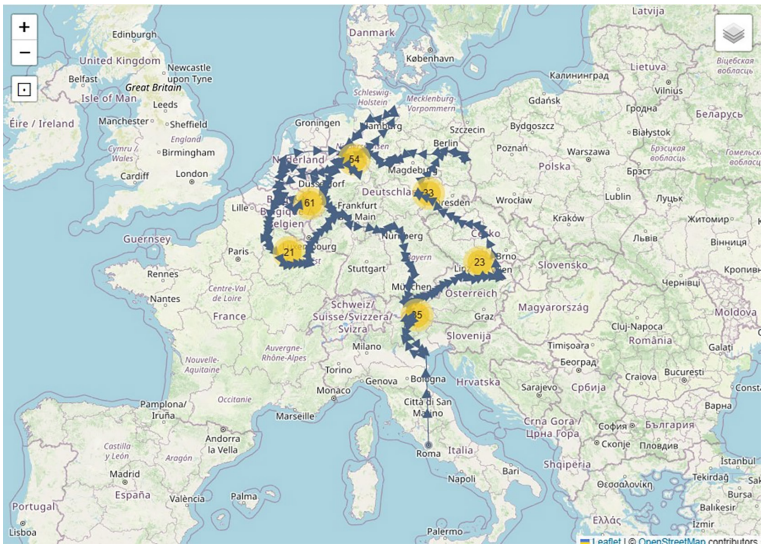


Figure 17 Interactive map with Commendone's diplomatic journey in Europe in 1560-62

The Graziani Archives team requested that when zooming in or using the scrolling function, users would be able to distinguish between locations marked on the map without additional information – represented by small light grey dots – and those containing content, symbolized by blue map markers. When a blue map marker is selected, it turns red, and simultaneously, an expandable sidebar opens on the

26 See Omeka S User Manual: `https://omeka.org/s/docs/user-manual/modules/mapping/`.

left side. It displays either Ruggieri's or Commendone's description of the event, a contemporary image or map of the location, and biographical information about the individuals mentioned in the text. Additionally, the timeline below serves as a chronological search tool, allowing users to locate specific sites along the itinerary.

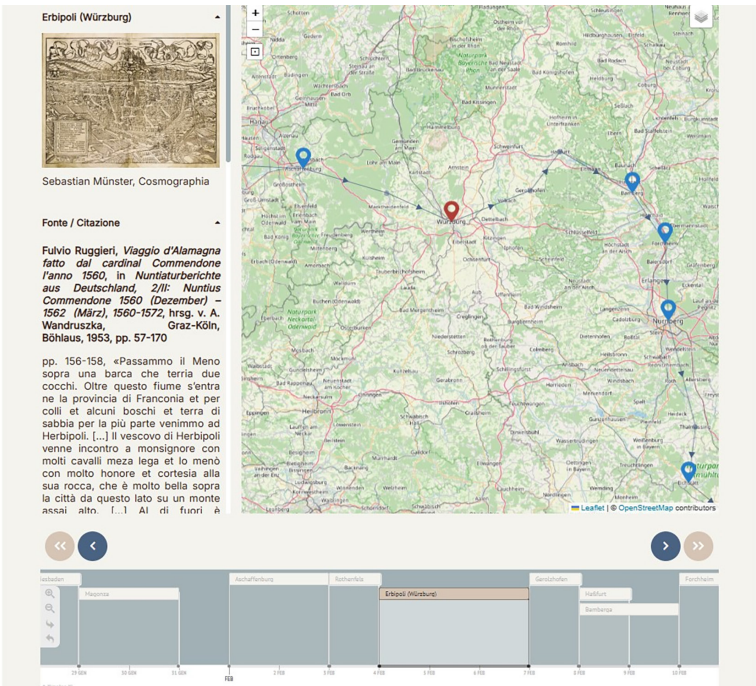


Figure 18 Interactive map displaying locations, further information on the selected location and a timeline

8 Long-Term Preservation

"Double, double toil and trouble," recite the three witches in *Macbeth* (Shakespeare, *Macbeth*, IV, I). Indeed, after the arduous task of constructing a portal comes a double challenge (and trouble): determining where to store the data and ensure long-term preservation. As has long been highlighted in various scientific forums, "the European investments for digital preservation in the last decade have been large and persistent but not able to support, at the moment, an accepted common vision, general services and adequate infrastructures" (Guercio 2013, 467).

In the absence of a national or European solution, the *Nuncio's Secret Archives* project team first had to determine where to host the

Omeka S backend they were about to configure. The cost-free solution was chosen: a virtual machine running the Linux distribution UBUNTU Core,[27] provided by one of the universities involved in the project.

Researchers in the humanities are not typically aware that long-term data preservation involves "policy questions, institutional roles and relationships, legal issues, intellectual property rights, and metadata" (Thibodeau 2002). The primary issue is not merely data preservation, as this does not necessarily require maintaining all of a dataset's digital attributes – data can easily be stored in CSV format on Zenodo, the Internet Archive, or other repositories. Rather, the challenge lies in preserving relational data and the query functionalities that a portal typically provides to users for interrogating the dataset but also links to external web pages, style sheets, graphic images, JavaScript elements, data extracted from databases, and other related components (Thibodeau 2013, 16). Simply storing data in CSV or another tabular format is insufficient, as it does not support user interrogation based on the relationships established between properties.

As Thibodeau noted in 2002, four criteria must be followed when considering long-term preservation:

1. Feasibility – which requires hardware and software capable of implementing the method;
2. Sustainability – which entails avoiding technological obsolescence and ensuring interoperability with other methods, such as those for discovery and delivery;
3. Practicality – which demands reasonable limits in terms of difficulty and cost;
4. Appropriateness – which depends on the types of objects to be preserved and the specific objectives of preservation.

I will begin with the final criterion, which also encompasses one of the core pillars of the FAIR principles: the fair use of data on the web. We obtained permission from the hosting libraries of the selected manuscripts from the Commendone-Graziani correspondence to publish them on the portal for the benefit of users. These users may view or download each image and utilize them for research or publication purposes, but not for commercial use. The images were processed using Mirador,[28] an image viewer optimized for displaying resources compliant with the International Image Interoperability Framework (IIIF, or Triple I-F), a standardized method for describing images on the web.[29] Mirador adheres to the Web

27 https://ubuntu.com/core.

28 https://projectmirador.org/.

29 https://iiif.io/api/.

Content Accessibility Guidelines (WCAG) 2.1 AA, which are legally adopted in the United States, Canada, and Europe.[30]

Additionally, we incorporated for some individuals or locations some images downloaded from Wikipedia, explicitly noting that they are classified as public domain.[31] When working with data that has already been published online or when publishing their own data and images, researchers must remain aware of both their responsibilities toward content created by others and their rights in disseminating their own work. It is crucial to consider in advance the type of licenses they wish to grant to other users.[32]

Among these four criteria, technological obsolescence is particularly problematic: a website can only remain viable if it is dynamic and evolves in response to the "environment of ongoing, open-ended and multidimensional change in which digital information exists" (Thibodeau 2013, 16-17). However, researchers are generally not in a position to anticipate technological changes or determine how best to address them, whether by locating the latest software release, replacing outdated formats, or adapting to shifts in standards.

The current state of academia leaves researchers in a precarious position.[33] They are required to submit a Data Management Plan before even beginning to develop a portal, yet many of them are encountering the online digital landscape for the first time – a sort of 'far west' in which each researcher or team must independently seek out a computer scientist or a storage provider and rely on their expertise, often without fully understanding the technical terminology or the implications of each decision. How can a humanities researcher be certain that the proposed software can support the data architecture they have designed? How should one mitigate the risks of rapid technological obsolescence? As Ramalho et al. observe,

> Databases are complex digital objects that contain heterogeneous information, often accompanied by structural definitions and even documentation. Their complexity makes it difficult to preserve this kind of object whilst maintaining all its significant properties. (2020, 109)

30 https://www.w3.org/WAI/standards-guidelines/wcag/.

31 See an example of the woodcut of Nuremberg in Hartmann Schedel's Nuremberg Chronicle in https://it.wikipedia.org/wiki/Cronache_di_Norimberga#/media/File:Nuremberg_chronicles_-_Nuremberga.png.

32 See the Creative Commons licenses list in https://creativecommons.org/licenses/list.en.

33 Thibodeau (2013, 19) refers to "a plethora of choices for preservation".

Paradoxically, funding programs typically require only a short-term preservation guarantee after a project's conclusion (usually three to five years). Furthermore, no repository exists at the European, national, or even university level that is specifically designed to host digital portals. This issue was raised years ago by several scholars: "the importance of digital preservation research has been growing for the past ten-fifteen years. Many papers and books describing problems, tools and techniques for digital preservation have been written, and standards providing preservation models have been published. However, most of this work focuses on preservation of file-based digital objects like documents, images, and web pages. Much less work has focused on the preservation of databases and scientific data, where there is a recognized need to preserve scientific data" (Stefanova, Risch 2013). No further steps were taken.

As long as we were ingesting data into the Omeka S backend configured by our team, we remained confident that the UBUNTU virtual machine allocated to us by one of the project's partner universities was a suitable solution. However, when the team reached the custom web design stage and began developing the frontend, it became evident that this solution had limitations. Moreover, the university could not guarantee ongoing maintenance or protection against technological obsolescence.

Thus, the project's research team found itself searching for a suitable service provider capable of ensuring the long-term preservation of our Omeka S-based portal. We were unable to locate such a provider in Italy. Fortunately, Libnamic – the company that had already provided us with web design and customization services – also offered long-term preservation solutions for Omeka S-based projects.[34] Initially, we had legal concerns regarding potential litigation or jurisdictional disputes (Duranti 2013, 28), as Libnamic is based in Spain. However, the matter was resolved to the satisfaction of all parties. The migration process took just over an hour, after which the web design and implementation proceeded on Libnamic's servers.

Nevertheless, the broader situation remains uncertain, as was already highlighted by Luciana Duranti (2013, 28):

> what we need is an internationally agreed upon legal framework that will support the development of integrated and consistent local, national and international networks of policies, procedures, regulations, standards and legislation concerning digital records, to ensure public trust grounded on evidence of good governance.

34 Libnamic Hosting. https://hosting.libnamic.com/.

The European Open Science Cloud (EOSC) is designed to

> provide European researchers, innovators, companies and citizens with a federated and open multi-disciplinary environment where they can publish, find and reuse data, tools and services for research, innovation and educational purposes.[35]

It remains far from being able to guarantee long-term preservation for portals developed through funded projects. Such projects frequently involve the creation of relational databases with query functionalities, archival source images, tagging, and other complex digital elements. Only a fully integrated and effective digital research ecosystem at a European level can ensure that research outputs from funded projects remain searchable and usable over the long term.[36]

Bibliography

Agosti, M.; Ferro, N.; Silvello, G. (2011). "Handling Hierarchically Structured Resources. Addressing Interoperability Issues in Digital Libraries". Biba, M.; Xhafa, F. (eds), *Learning Structure and Schemas from Documents*. Berlin; Heidelberg: Springer-Verlag, 17-49.

Aliaga, D.G.; Bertino, E.; Valtolina, S. (2011). "DECHO-A Framework for the Digital Exploration of Cultural Heritage Objects". *Journal on Computing and Cultural Heritage* 3(3), 12, 1-21. https://doi.org/10.1145/1921614.1921619

Al Kalak, M. (2024). "Due secoli e tre continenti. Il carteggio tra Girolamo Lagomarsini e Lodovico Antonio Muratori". Boutier, J.; Forner, F.; Paoli, M.P.; Tinti, P.; Viola, C. (a cura di), *Le stagioni dell'erudizione e le generazioni degli eruditi. Una storia europea (secoli XV-XIX)*. Bologna: CLUEB, 675-91.

Berti, P. (1864). *Catalogo delle pergamene e manoscritti già spettanti alla famiglia Graziani di Città di Castello*. Firenze: M. Cellini.

Blaney, J.; Milligan, S.; Steer, M.; Winters, J. (2021). *Doing Digital History. A Beginner's Guide to Working with Text as Data*. Manchester: Manchester University Press.

Bonora, E. (2019). "Comprendre et décrire un autre monde. Le voyage d'un nonce dans l'Europe des confessions et du pluralisme religieux (1560-62)". Ferrant, J.; Guillabert-Madinier, T. (éds), *Le Langage et la Foi dans l'Europe des Réformes XVIe siècle*. Paris Éditions Classiques Garnier, 215-24.

Bonora, E. (2023). "Gli archivi segreti del nunzio". *Riforma e movimenti religiosi*, 13, 177-84.

35 https://research-and-innovation.ec.europa.eu/strategy/strategy-research-and-innovation/our-digital-future/open-science/european-open-science-cloud-eosc_en and https://open-science-cloud.ec.europa.eu/.

36 These were the conclusions of the EOSC Symposium held in Berlin on 21-23 October 2024. https://open-science-cloud.ec.europa.eu/news/eosc-eu-node-spotlight-highlights-and-milestones-2024-eosc-symposium.

Bradley, K.; Lei, J.; Blackall, C. (2007). Towards an Open Source Repository and Preservation System: Recommendations on the Implementation of an Open Source Digital Archival and Preservation System and on Related Software Development. https://unesdoc.unesco.org/ark:/48223/pf0000154761

Cappellatto, L. (2015-16). *Studio e realizzazione di una piattaforma di archiviazione di contenuti digitali per l'Università degli Studi di Padova* [Tesi di Laurea]. Venezia: Università Ca' Foscari Venezia.

Corsini, M. (2000). "La Biblioteca e l'archivio Graziani di Vada". *Rara volumina* 7, 127-40.

Dallas, C. (2004). "Policies and Strategies for Digital Preservation". Tola, V.; Castellani, C. (eds), *The Future of Digital Memory and Cultural Heritage*. Roma: ICCU, 281-7.

Dunn, S.; Hedges, M. (2012). "Crowd-Sourcing Scoping Study: Engaging the Crowd with Humanities Research". London: Arts and Humanities Research Council. http://crowds.cerch.kcl.ac.uk/

Duranti, L. (2013). "Trust and Conflicting Rights in the Digital Environment". Duranti, Shaffer 2013, 24-30.

Duranti, L.; Shaffer, E. (eds) (2013). *Proceedings of the Memory of the World in the Digital Age: Digitization and Preservation. An International Conference on Permanent Access to Digital Documentary Heritage* (26-28 September 2012). Vancouver: UNESCO.

Ganguly, R.; Budroni, P.; Sánchez Solís, B. (2017). "Living Digital Ecosystems for Data Preservation: An Austrian Use Case Towards the European Open Science Cloud". Chan, L.; Loizides, F. (eds), *Expanding Perspectives on Open Science: Communities, Cultures and Diversity in Concepts and Practices*. Limassol: Cyprus ELPUB, 203-10.

Gilliland, A.J.; McKermmish, S.; Lau, A.J. (2017). *Research in the Archival Multiverse*. Clayton, Victoria: Monash University Publishing.

Guercio, M. (2013). "Digital Preservation in Europe. Strategic Plans, Research Outputs and Future Implementation. The Weak Role of the Archival Institutions". Duranti, Shaffer 2013, 467-81.

Hardesty, J.L. (2014). "Exhibiting Library Collections Online: Omeka in Context". *New Library World*, 115(3-4), 75-86. http://dx.doi.org/10.1108/NLW-01-2014-0013

Hawkins, A. (2022). "Archives, Linked Data and the Digital Humanities: Increasing Access to Digitised and Born-digital Archives Via the Semantic Web". *Archival Science*, 22, 319-44. https://doi.org/10.1007/s10502-021-09381-0

Jaitner, K. (2004). "Miszelle das Archivio Graziani in Vada". *Quellen und Forschungen aus italienischen Archiven und Bibliotheken,* 84, 509-12.

Jaitner, K. (2021). *Instruktionen und Relationen für die Nuntien und Legaten an den europäischen Fürstenhöfen von Sixtus V. bis Innozenz IX. (1585-1591)*. Freiburg; Basel; Wien: Herder.

Kamal, A.M.; Golub, K. (2025). "Subject Matters: Metadata Standards and Subject Access for Library and Museum Catalogues". Hanssen, J.-M.; Furuseth, S. (eds), *The Hermeneutics of Bibliographic Data and Cultural Metadata*. Oslo: National Library of Norway, 204-39.

Lewis, S.; Hayes, L.; Stangeland, E.; Shepherd, K.; Jones, R.; Roos, M. (2010). "DSpace Under the Hood: How DSpace Works". *The 5th International Conference on Open Repositories (OR2010)* (Madrid, Spain, 6-9 July 2010). https://doi.org/10.2390/biecoll-OR2010-52

Mariani, G. (2022). "Sisto V e Antonio Maria Graziani. Documenti del pontificato sistino. Dall'archivio privato della famiglia Graziani". *Sisto v e la comunità camerte = Conference proceedings* (Camerino, 12 December 2022).

Maron, D.; Feinberg, M. (2018). "What Does It Mean to Adopt a Metadata Standard? A Case Study of Omeka and the Dublin Core". *Journal of Documentation* 74(4), 674-91. https://doi.org/10.1108/JD-06-2017-0095

Marsili, M. (2002). s.v. "Graziani, Antonio Maria". *Dizionario biografico degli Italiani*, vol. 58, 801-4.

Mazzatinti, G. (1904). *Archivi della storia d'Italia*, IV. Rocca S. Casciano: Cappelli.

Michetti, G. (2013). "Archives Are Not Trees. Hierarchical Representations in Digital Environment". Duranti, Shaffer 2013, 1002-10.

Miksa, T.; Sánchez Solís, B.; Schrauf, C. (2016). "Data Management Plans. Fortbildungsseminar für Forschungsdaten und e-Infrastrukturen". http://phaidra.univie.ac.at/o:441308

Montanaro, D. (2016-17). *Digital Ecosystems for Open Science. Il caso Università di Vienna, l'esempio di Phaidra* [Tesi di laurea]. Venezia: Università Ca' Foscari. https://unitesi.unive.it/retrieve/fc08629a-5ddd-4fc3-a0be-10c31b239a40/855470-1213774.pdf

Moretti, M. (2012). *Lettere di Pieter de Witte. Pietro Candido nei carteggi di Antonio Maria Graziani, edizione critica*. Roma: De Luca.

Moretti, M. (2015). "Committenti, intermediari e pittori tra Roma e Venezia attorno al 1600. I ritratti di Domenico Tintoretto per il nunzio Graziani e una perduta 'pentecoste' di Palma il Giovane per Fabio Biondi". *Storia dell'Arte,* 41, 21-43.

Moretti, M. (2018). "Antonio Maria Graziani e le fatiche della carriera. L'altare di famiglia a Sansepolcro e la commissione dell''Assunta' a Palma il Giovane". *Storia dell'Arte*, 15(2), 18-67.

Moretti, M. (2021). "L'altare Graziani da Raffaello a Palma il Giovane. Una copia della 'Madonna' Canossa e una 'sentenza' sfavorevole a Giovanni De' Vecchi". *Storia dell'Arte* 155-6(1-2), 61-87.

Moretti, M. (2023). "L'ozio religioso in villa. Un ritiro 'alla Cappuccina' per Carlo Graziani, maestro di casa del cardinale Francesco Barberini". Porzio, G.; Primarosa, Y. (a cura di), *Orazio Gentileschi e l'immagine di san Francesco. La nascita del caravaggismo a Roma*. Milano: Officina Libraria, 112-23.

Morton, A. (2011). "Digital Tools: Zotero and Omeka". *The Journal of America History,* 98(3), 952-3. https://doi.org/10.1093/jahist/jar520

Opgenhaffen, L. (2022). "Archives in Action. The Impact of Digital Technology on Archaeological Recording Strategies and Ensuing Open Research Archives". *Digital Applications in Archaeology and Cultural Heritage*, 27. https://doi.org/10.1016/j.daach.2022.e00231

Raines, D. (2024). "Un archivio diffuso. La storia dell'archivio di Antonio Maria Graziani". *La Chiesa di Roma e l'Europa multiconfessionale nella prima età moderna: attori, politiche, esperienze* (Parma, 17-19 aprile 2024).

Ramalho, J.C.; Ferreira, B.; Faria, L.; Ferreira, M. (2020). "Beyond Relational Databases: Preserving the Data". *New Review of Information Networking* 25(2), 107-18.

Sprout, B.; Romkey, S. (2013). "A Persistent Digital Collections Strategy for UBC Library". Duranti, Shaffer 2013, 257-68.

Stefanova, S.; Risch, T. (2013). "Scalable Long-Term Preservation of Relational Data through SPARQL queries". *Semantic Web journal*. https://www.semantic-web-journal.net/content/scalable-long-term-preservation-relational-data-through-sparql-queries-1

Thibodeau, K. (2002). "Overview of Technological Approaches to Digital Preservation and Challenges in Coming Years". *The State of Digital Preservation: An International Perspective*. Washington D.C.: Council on Library and Information Resources. https://www.clir.org/pubs/reports/pub107/thibodeau/

Thibodeau, K. (2013). "Wrestling with Shape-Shifters. Perspectives on Preserving Memory in the Digital Age". Duranti, Shaffer 2013, 15-23.

Toulson, S. (2021). "Four Header Styles That Will Optimize Your Service Portal Navigation and Look Good Doing It". *Portalorem*, 7 April 2021. https://portalorem.com/blog/five-header-styles-that-will-optimize-your-service-portal-navigation)

Trace, C.B. (2022). "Archives, Information Infrastructure, and Maintenance Work". *Digital Humanities Quarterly,* 16(1). https://www.digitalhumanities.org/dhq/vol/16/1/000603/000603.html

Tritch Roman, G. (2018). "Scalar and Omeka". *Journal of the Society of Architectural Historians*, 77(1), 122-3.

Models of Data Extraction and Architecture in Relational Databases of Early Modern Private Political Archives
edited by Dorit Raines

Reconciling Complex Historical Records with Omeka S Relational Database
The Case of the Graziani Archive

Gabriella Desideri
Università Ca' Foscari Venezia, Italia

Abstract The paper focuses on extracting data from correspondence related to diplomatic missions by Giovanni Francesco Commendone and Antonio Maria Graziani in sixteenth-century Poland. This data is integrated into Omeka S, an open-source platform allowing organisation and entity-relationship mapping. The approach virtually reunites dispersed historical documents on the Graziani Archives portal. It discusses balancing documentation complexity and Omeka S-imposed standardisation's impact on historical research. Concrete examples highlight scholars' ability to interrogate this vast structured data via Omeka S.

Keywords Geolocation. Document summary. Standardisation. Omeka S. Data extraction.

Summary 1 Introduction. – 2 Navigating Historical Names and Places in the *Nuncio's Secret Archives* Project. – 3 Striking a Balance between Standardisation and Historical Complexity. – 4 Analysing Commendone's Letters to Rome Through Detailed Document Summaries. – 5 Exploring Relational Dynamics through a Database of Historical Documents.

1 Introduction

The increasing accessibility of documentary materials online raises epistemological inquiries concerning their suitability as historical sources for scholarly utilisation (Vitali 2004). Consequently, historians must diligently contemplate their distinct characteristics in

Edizioni Ca'Foscari

Studi di archivistica, bibliografia, paleografia 8
e-ISSN 2610-9093 | ISSN 2610-9875
ISBN [ebook] 978-88-6969-919-1 | ISBN [print] 978-88-6969-920-7

Peer review | Open access
Submitted 2023-10-03 | Accepted 2024-01-22 | Published 2025-05-22

DOI 10.30687/978-88-6969-919-1//011

comparison to their physical paper counterparts, encompassing the criteria employed for digitising these documents. This research aims to reflect upon the choices made to strike a delicate balance among the complexity of documentation in the data extraction process, specifically focusing on the letters related to Giovanni Francesco Commendone's diplomatic mission in Poland during the latter half of the sixteenth century.[1] Additionally, this study examines the standardisation imposed on the entities by Omeka S, a web publication system employed for generating, editing, and managing online data.[2] These historical documents constitute a component of the Graziani Archives, a privately held collection currently dispersed among several hosting institutions, including the Graziani Archives of Vada (Tuscany, Italy), the Kenneth Spencer Research Library (University of Kansas, Lawrence, Kansas, USA), and the New York Public Library (USA).[3]

Through the provision of concrete examples, this paper will demonstrate the potential for scholars to query the extensive volume of structured data derived from historical records, alongside exploring the possibilities offered to users by the relational database.

2 Navigating Historical Names and Places in the *Nuncio's Secret Archives* Project

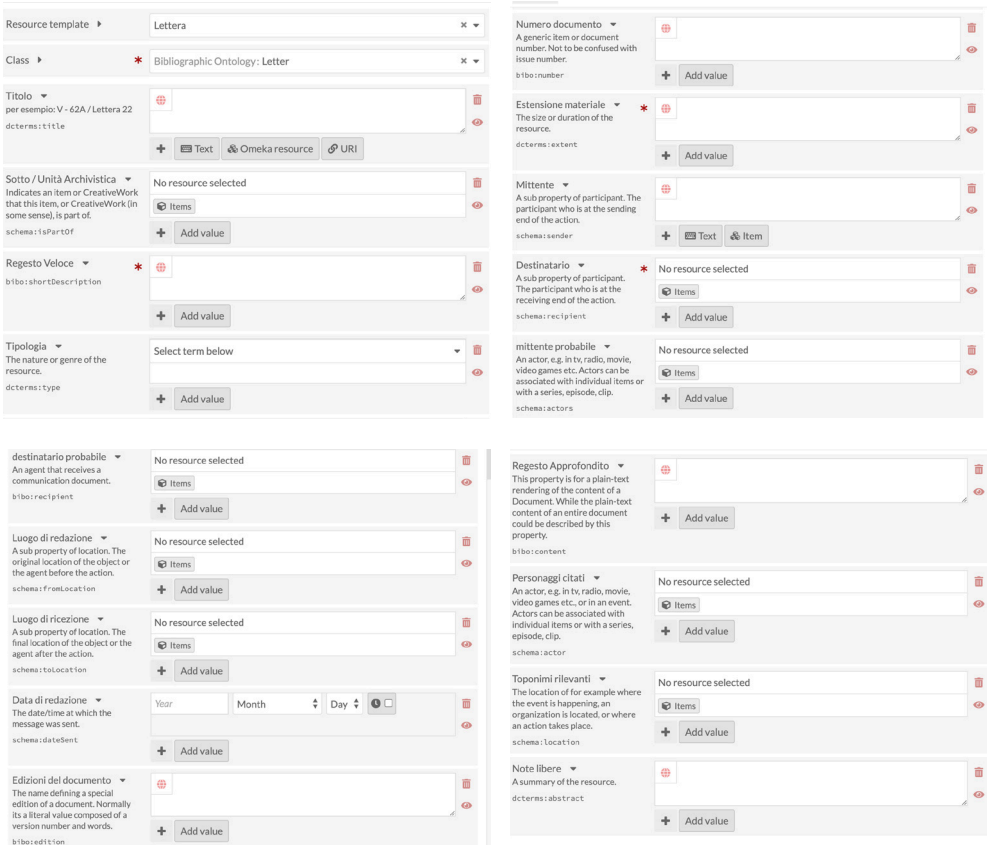
Before delving into specific aspects of this subject, it is necessary to provide a methodological introduction. The project *Nuncio's Secret Archives*, funded under the PRIN 2017 program, has the objective of virtually reassembling the Graziani Archives. Scholars are granted access to the new version of the Archives' inventory and summaries of selected documents (*regesti*), mostly letters, through the homonymous web portal. The research team engaged in processing these letters utilised a resource template, which was created through a data modelling process under the guidance of the project's scientific leaders. This template facilitates the selection and organisation of information for ingestion into Omeka S.

The template has nineteen fields, encompassing essential information for document identification (archival reference, topical and chronological date, sender, and addressee), bibliographic references for published sources, as well as the names of individuals and places mentioned that are considered most relevant **[fig. 1]**. Additionally, it incorporates *Note libere* 'free notes', focusing on the material aspects

1 On Commendone's first mission to Poland, cf. Pastor 1928, 369-74.

2 https://omeka.org/s/.

3 https://nsa.unipr.it/.



Figures 1a-d The back-end template for the Letters section used by researchers for the data entry

of the document. Users also have access to the image of each document from the same template.

As observed in the template, the project's researchers had to extract the main names of individuals and places mentioned (the fields *Personaggi citati* 'cited persons' and *Toponimi rilevanti* 'relevant places'). Direct engagement with the documents was revealed vital for the elaboration of metadata. Initially, sharing the extracted data related to individuals and place names between the different researchers working on the documents was not contemplated. However, as our work progressed, it became evident that we needed to share the collected names of people and places, leading to the creation of two shared lists – one for individuals and another for places – where all researchers inserted the extracted data. These lists provided a platform for problematising cases and discussing choices. Sharing these

lists of names also allowed for standardisation optimisation and facilitated the reviewing process.

Following this experience, guidelines were established to structure the items of people and places, as well as for their selection, allowing for some discretion, particularly concerning the name/place given as title, which serves as a key access point to each item. For instance, it was decided to title the items with the version of the name found in the sources, both for people and places. This practical approach was largely influenced by frequently encountering such versions in transcriptions reported in document summaries. This naming convention was consistently applied in all the regesti, facilitating the seamless transition between items of individuals and places, document summaries, and the digitalisation of each letter. For example, the significant Polish location, Piotrków Trybunalski,[4] where an important Sejm took place in 1565, was titled as 'Petricovia', the version found in the documents. This form was also used in contemporaneous letters of Andres Dudith, presenting the Italian and Latin versions of the place name. Alternative variants of the name, such as 'Peterkaw', 'Petherkaw', 'Petterkaw', 'Petricovium', and 'Piotrków', were mentioned by Dudith only in the places' index (Dudithius 1992, *ad indicem*).

Various historians' works demonstrate the different versions in which place names can be found, underscoring the significance of interpretative choices made by researchers and the value of sharing outcomes among them. In his study on the relationship between the Jesuit order and the Polish nobility during the eighteenth century, Andrea Mariani utilised the contemporary version of the place name, 'Piotrków Trybunalski', and included its Latin translation solely in the index (Mariani 2014, *ad indicem*). On the other hand, Robert Frost (2015, *ad indicem*) employed the current version of the name in his work, *The Oxford History of Poland-Lithuania*, but presented a more concise rendition ('Piotrków'). In contrast, Ludwig von Pastor (1928, 360, 369, 372) opted for the German version, 'Petrikau', in his book, *Storia dei Papi*. These examples serve merely as illustrations and do not aim to encompass the entirety of the place name's usage in historiography. Nevertheless, they show the importance of results sharing in order to avoid further confusion.

While the chosen form aids Italian-speaking users in consulting the items, the project maintains an international orientation, leaving the user the option of choosing names found in the sources or in common use. For example, the name 'Kiev' is widely used in

4 Until 1569, the Sejm convened in Piotrków. Subsequently, in 1578, the Crown Tribunal for the lands of Greater Poland was established there. This is probably the reason why the present location name is Piotrków Trybunalski (Słonia 2021, 1591).

historiography, so titling the item with the version found in the documents ('Chiovia') might raise more questions than it answers. In some other cases, questions were raised regarding the use of the contemporary version of a person's name, such as 'Teodoro' for Teodore I, son of the czar Ivan IV. Our first orientation was to choose the Russian version of the name, Fëdor I Ivanovič. However, after conducting further research, we opted for the Italian version of the name as it was found in reputable sources like the *Enciclopedia Treccani* (Epstein 1937) and Nicholas V. Riasanovsky's important handbook, *Storia della Russia* (2010, *ad indicem*), largely used in Italian historiography on the subject.

3 Striking a Balance between Standardisation and Historical Complexity

Let us now explore the decision-making process behind each item. In the case of both place and person items, when there are different translations of the same name, the researcher had the discretion to retain one or more variants if they were considered relevant. These additional variants could be included in the *Altre varianti del nome* 'other name variants' field. An illustrative example of this practice can be seen in the case of Pawel z Bzezin Grzegorz, referred to as *ministro et predicatore de trinitarii* 'Minister and Preacher of the Anti-Trinitarian heretical group' in the documents.[5] In Commendone's letter of 6 July 1565, he is mentioned as 'Gregorio Paulo'. As a result, we made the decision to title the item using the Latin version of the name, 'Paulus, Gregorius', while adding 'Brzezinensis' and 'z Brzezin' as nicknames. In the *Altri nomi* 'other names' field, we included the Polish version of the name: 'Pawel Grzegorz'. Similarly, for the item of Théodore de Bèze, we chose to use the Italian version of the name, 'Teodoro Beza' and to insert in the Altri nomi field: 'Bèze (de), Théodore'.

For place items, the decision-making process typically involves recording the version of the name found in the sources, along with other relevant language versions. Additionally, the current version of the name is included in order to geolocate it. This approach allows users to visualise the place name on a map, offering a clear and interactive representation.

5 New York Public Library, Manuscript and Archive Division, MssCol 603, *Giovanni Francesco Commendone, Diplomatic Correspondence, 1563-5* (NYPL, MssCol 603), second Register, Letter 44, Commendone to Borromeo, 6 July 1564, `https://grazianiarchives.eu/s/graziani-archives/item/3840`. See note 15 for an explanation of the difference between the use of the word 'Trinitarii' and its meaning in this context.

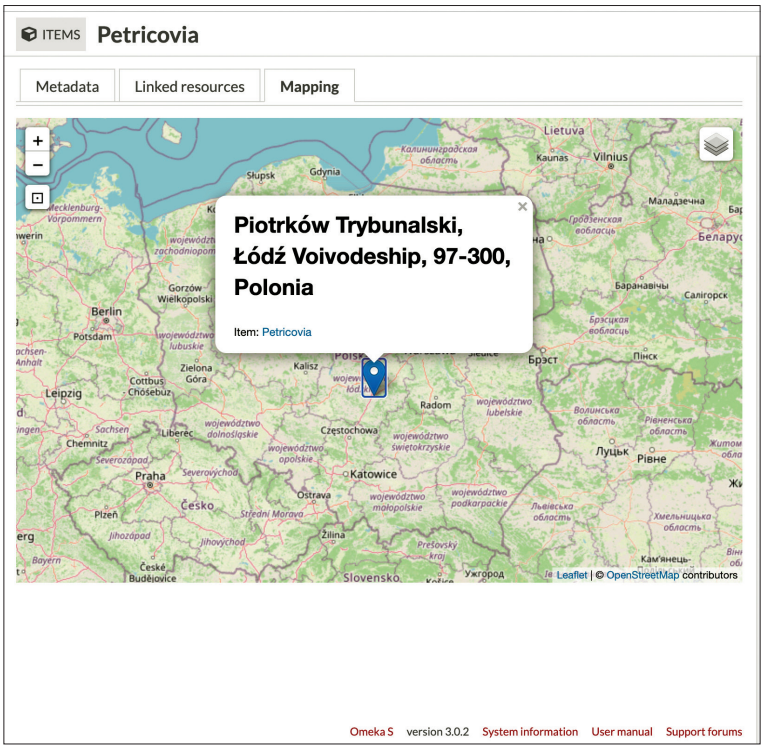


Figure 2 The current version of the place item 'Petricovia' geolocated on Google Maps

As demonstrated in the case of Petricovia, the version of the place name is recorded as it appears in historical documents. The inclusion of the current version of the name, Piotrków Trybunalski, enables users to relate it to modern geographical references, fostering a meaningful connection between the historical context and contemporary perspectives **[fig. 3]**. This dynamic presentation enriches the user experience and facilitates a more comprehensive analysis of the historical sources.

Geolocating historical place names can lead to potential anachronisms (Drucker 2016, 241-2), as the present-day boundaries and configurations often differ from those during the time period under study. To address this concern, initially, we decided against recording the names of historical regions, recognising that they might not align with present-day boundaries. However, as our work progressed, we recognised the significance of historical regions in understanding the context of important events or phenomena discussed in the correspondence. For instance, we included the place name 'Germania' when Commendone mentioned imperial questions related to the

laity's reception of the cup (Pastor 1928, 346-61), and decided to record it without geolocation. Yet, at the end of the process, we decided against including these place names because in most cases they became meaningless in respect to the geographical space they were supposed to represent due to constant changes of the reigning authority and sometimes even its borders.

The geolocation process, while valuable, does not imply a direct and absolute correspondence between historical place names and their present locations, given the evident historical complexities. Instead, it serves as a guiding reference for users, particularly for lesser-known villages and suburbs. For cities, we encountered similar challenges due to the differences between their current configuration and historical counterparts during the time period of the correspondence under study. To mitigate these challenges, we adopted a pragmatic approach. In cases where it was not feasible to geolocate the historical place name, we titled the item with the name of the suburb, followed by the name of the encompassing city in brackets **[fig. 3]**. As a result, we could still retain the reference to the suburb while geolocating the associated city. This approach allows users to establish a visual connection on the map, even in instances where direct geolocation of the historical place name is not possible.

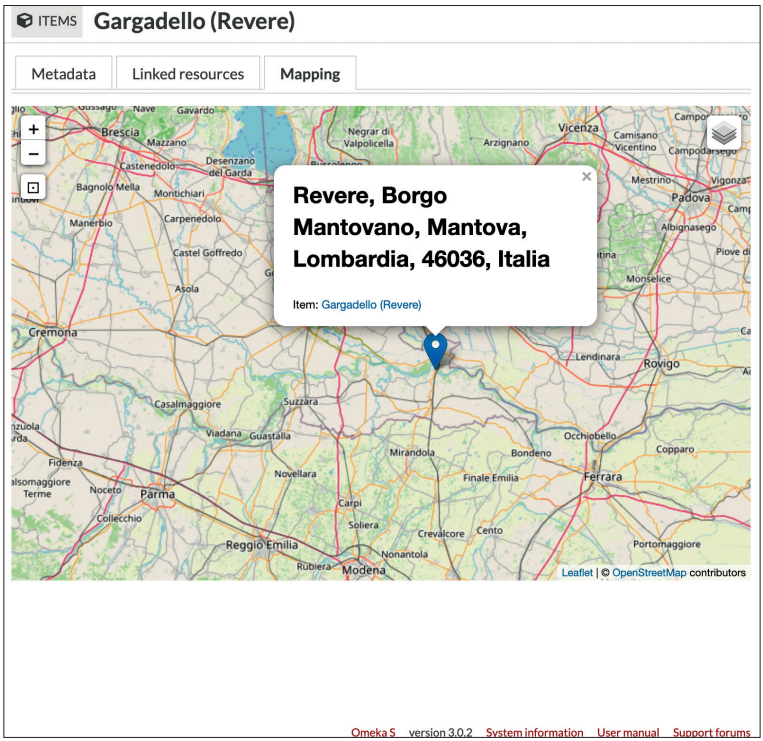


Figure 3 Geolocating a suburb within a city on Google Maps

Ultimately, our methodology aims to strike a balance between providing accurate geographic information and acknowledging the historical complexities surrounding place names and their present-day associations. Hence, geolocating historical place names should be approached with a critical perspective and an awareness of potential inaccuracies.

In summary, while geolocating historical place names can be a valuable tool, it requires careful consideration and transparency about the inherent limitations and potential anachronisms. By combining this approach with informative metadata, we strive to offer a nuanced and accurate representation of the locations referenced in the historical correspondence.

Regarding the person item, in addition to recording the usual data of name and surname, we have incorporated one or more occupations or titles along with their related chronological time frame. These details were obtained through thorough research on national and international repertories, aiming to identify the most significant occupation and/or title held by the person during their lifetime. However, when an individual had multiple relevant occupations, we

focused on those held during the year mentioned in the letter where the person is referenced. This approach facilitates the immediate identification of the person, enhancing the comprehension of the document summary. Consequently, the user gains a primary access point for exploring the research on the document and can decide whether to read it in its digital version.

In selecting occupations, we first decided not to include indefinite or ambiguous terms such as 'intellectual' or 'noble'. For individuals with multifaceted roles like intellectuals, preachers, or 'heretics' (such as Martin Eisengrein, Bernardino Ochino, or Giovanni Paolo Alciati della Motta), we preferred not to assign a specific occupation due to the complexities of their identities. These individuals are usually well-known in both Italian and foreign historiography (Tüchle 1959; Gotor 2013; Sella 1960). Nevertheless, some exceptions to this approach exist. Initially, we avoided using the title 'doctor' (*medico*) as it was considered too vague. However, later on, we decided to include it in certain cases where doctors played a significant role in the documents' summaries. Furthermore, when the inventory section was added sometimes rendering the identification of individuals complicated, the same was applied in case the only information we had on an individual was his being 'scholar' or 'noble'. We preferred then to use these ambiguous terms as a distinctive characterisation of the individual.

These instances demonstrate that the application of design principles from a top-down perspective is not always feasible. Instead, we must carefully assess and adapt them to meet the specific requirements and complexities of the documentation, aiming to represent its intricacies to the fullest extent possible.

Another challenge we encountered pertained to letters that lacked a clear sender and/or addressee, or where accurate identification was not possible, such as in some letters from manuscript E105 today in Kansas (Baja Guarienti, *infra*). This issue particularly affected letters that posed difficulties in attributing them either to Commendone or Graziani. Initially, we attempted to address this by creating a new item labelled 'Commendone o Graziani', inserting it into the sender field. However, after several attempts, we realised that this solution was ineffective for structured research. The new item was distinct from those of Commendone and Graziani, thus preventing the linkage of letters associated with the 'Commendone o Graziani' item to the respective items representing Commendone and Graziani as distinct individuals. In short, the new item led to the loss of any connection between the letter and its potential senders.

To overcome this issue, we introduced a new field called *Mittente probabile* 'probable sender'. In the sender field, we inserted the item *Incerto* 'uncertain', and in the new field, we entered the item representing the probable sender. This solution allowed for more effective

and accurate data structuring as it preserved the link between the letter and its potential senders in the linked sources section, ensuring a clear connection between the letter and the items of its probable senders, as exemplified in the following image for Commendone's item **[fig. 4]**.


ITEMS **Commendone, Giovanni Francesco**

Metadata | Linked resources

The following resources link to this item:

Filter by property: All

28 of 30 676–700 of 742

actors

Title	Alternate label	Class
Ms. E105 / Lettera 222	mittente probabile	Letter
Ms. E105 / Lettera 223	mittente probabile	Letter
Ms. E105 / Lettera 229	mittente probabile	Letter
Ms. E105 / Lettera 233	mittente probabile	Letter

sender

Title	Alternate label	Class
Ms. E105 / Lettera 224	Mittente	Letter
Ms. E105 / Lettera 225	Mittente	Letter
Ms. E105 / Lettera 226	Mittente	Letter
Ms. E105 / Lettera 230	Mittente	Letter
Ms. E105 / Lettera 231	Mittente	Letter
Ms. E105 / Lettera 232	Mittente	Letter

Figure 4 Commendone's registration both as confirmed and probable sender (back-end)

Subsequently, we extended the same principles to all probable senders and addressees, introducing the field *Destinatario probabile* 'probable recipient' as well. This comprehensive approach allows users to access all letters with an uncertain sender and/or addressee through the linked resource of the 'uncertain' item.

In cases where the researcher was unable to make any identification due to the absence of internal elements, we recorded the reference provided in the document. Similarly, when a person could not be identified, we titled the person item as referenced by the document and within quotation marks. Additionally, we created a new field *Identificazione possibile* 'possibly same as', inserting the item that corresponded to our hypothesis of identification (in case we had

one). This strategy enabled us to maintain transparency and provided relevant contextual information, thereby acknowledging the limitations and uncertainties present in historical research.

These decisions illustrate our commitment to striking a balance between the characteristics of the documentation, historiography, and standardisation. By registering the 'variants' of names, we preserve the complexity of the sources, offering multiple access points to the data. The solutions implemented for probable senders and addressees achieve a harmonious blend between the need for disambiguation imposed by the database and the nature of historical research and historical sources, which may not always provide definitive and incontrovertible data.

It is essential to emphasise that all the choices made in this project for the Letters section, including the selection of the most relevant names of persons and places, occupations, and informations in the document summaries, are firmly grounded in their relevance to the primary focus of our project – the Polish diplomatic experience of Antonio Maria Graziani and Giovanni Francesco Commendone. Maintaining this focus ensures that the database remains a valuable resource for scholars, providing comprehensive insights into the historical context and the interactions between these key figures.

In conclusion, our methodology reflects a thoughtful balance between data standardisation and historical complexity, while constantly keeping in mind the overarching objectives of the project. Through meticulous research and adaptive decision-making, we aim to offer an accessible and informative platform that serves the needs of scholars and researchers in their exploration of the Graziani Archives and its historical significance.

4 Analysing Commendone's Letters to Rome Through Detailed Document Summaries

As previously stated, the researchers were engaged in the endeavor of generating document summaries (*regesti*) for a selection of documents. In this section, I will focus on the regesti of Commendone's letters to Rome, which were authored between 1563 and 1565 and transcribed into a letter copybook currently housed at the New York Public Library. These documents hold significant importance for the project, as they provide firsthand accounts of Commendone's first diplomatic mission in Poland following the conclusion of the Council of Trent. Furthermore, they serve as crucial sources for comprehending the subsequent strategies employed by both Commendone and Graziani in Poland, and for gaining a deeper insight into the evolution of Papal diplomacy in Central-Eastern Europe. Due to their significance, meticulous summaries of these documents have been crafted, encompassing a diverse array of themes and issues addressed within them. These comprehensive summaries are documented in the *Regesto approfondito* 'detailed summary' section. Moreover, for each letter, a concise summary has been produced, documented in the *Regesto veloce* 'short summary' section.

In these regesti, a conscious effort has been made to maintain a close correlation with the original documents, adhering closely to their structure. Specifically, the summaries mirror the division between encrypted and non-encrypted sections present in the documents. Concerning Commendone's letters to Rome, the text has been decoded and transcribed by a copyist from the sixteenth century into the letter book. The extracted information has been incorporated into the 'Free notes' section (*Note*), dedicated to the material aspects of the document. It has been noted, in line with observations made by Henry Biaudet,[6] that sixteenth-century letters to Rome often contain an unencrypted section featuring the date and sender's signature, followed by a fully encrypted supplement devoid of any date or signature. Retaining this division aids in enhancing the comprehension of stylistic and thematic variances between the two segments of the letter. This approach facilitates easier reference and provides additional insights into the diplomatic activities of Commendone and Graziani in Poland.

During the creation of these summaries, two primary criteria were rigorously adhered to: first, to ensure an unambiguous interpretation

6 "A Rome, par exemple, la chiffre constituait toujours un supplément à un dépêche au clair, supplément écrit sur feuille à part, ne contenant aucune parole non chiffrée, aucune formule initiale ni finale de politesse, et surtout aucune date ni signature" (Biaudet 1910, 3-4).

of the information, and second, to meticulously capture the linguistic subtleties employed by the writer by transcribing specific expressions from the original document. Initially, a strategy of minimal transcription was chosen to prevent undue densification of the text. However, as progress was made, it became evident that retaining characteristic expressions and segments of the document was imperative. This includes excerpts of Commendone's descriptions of the confessional situation in the visited locales,[7] mentions of heretical sects and practices, and the propagation of publications throughout the Kingdom of Poland.[8] Passages were transcribed in which Commendone highlighted issues related to the coexistence of diverse religious denominations, such as the allocation of public space between rival confessions in Danzig subsequent to the promulgation of an edict condemning those who persecuted Catholics to death.[9] Additionally, significant political occurrences and definitions were transcribed, such as the passage in which Commendone recounts his address before the Polish Senate, emphasising the importance of the Council of Trent and of its decrees.[10] The transcriptions provide readers with insights into primary sources pertaining to the religious and political landscape in Poland, which stood as a confessional boundary, in the immediate aftermath of the Council of Trent.

In this manner, users gain direct access to the distinct vocabulary utilised in these documents, furnishing essential insights into the writer's ideas, imagination, and references, whether political or scriptural. Certain words utilised by Commendone, laden with significance, would lose their potency if diluted into a bland synthesis. Similarly, specific expressions from Commendone's lexicon, such as describing the spread of a heretical confession as an infection, a

7 NYPL, MssCol 603, first Register, Letter 6, Commendone to Borromeo, 24 November 1563, https://grazianiarchives.eu/s/graziani-archives/item/3615; NYPL, MssCol 603, second Register, Letter 38, 7 October 1564, https://grazianiarchives.eu/s/graziani-archives/item/3833; NYPL, MssCol 603, sixth Register, Letter 169, 1 October 1565, https://grazianiarchives.eu/s/graziani-archives/item/3974; NYPL, MssCol 603, seventh Register, Letter 174, 4 November 1565, https://grazianiarchives.eu/s/graziani-archives/item/3952.

8 NYPL, MssCol 603, second Register, Letter 41, Commendone to Borromeo, 8 June 1564, https://grazianiarchives.eu/s/graziani-archives/item/3837; NYPL, MssCol 603, second Register, Letter 44, 20 June 1564 https://grazianiarchives.eu/s/graziani-archives/item/3840; NYPL, MssCol 603, second Register, Letter 53, 9 September 1564, https://grazianiarchives.eu/s/graziani-archives/item/3854.

9 NYPL, MssCol 603, second Register, Letter 43, Commendone to Borromeo, 20 June 1564, https://grazianiarchives.eu/s/graziani-archives/item/3839.

10 NYPL, MssCol 603, second Register, Letter 47, Commendone to Borromeo, 8 August 1564, https://grazianiarchives.eu/s/graziani-archives/item/3843.

'disease' ("*postema*") to be eradicated, have been retained.[11] Passages describing the interactions between various figures within the Polish context, such as Andrzej Frycz Modrzewski, characterised as a 'man of letters who writes extensively, having been a heretic for a long time and closely associated with the archbishop' ("*huomo di lettere che scrive assai, già gran tempo heretico et molto intrinseco dell'arcivescovo*"), have also been transcribed.[12] Such information aids in reconstructing networks among individuals mentioned in the documents. Moreover, references to primary sources have been maintained for individuals, places, or events that could not be readily identified. For instance, in the letter dated 26 February 1565, Commendone referred to an 'elderly and highly esteemed canon' ("*canonico vecchio et di molta stima*") who favoured the divorce between Sigmund II August and Catherine of Ausburg, without providing further details for identification.

Additionally, passages have been preserved in which Commendone contemplates the most effective *modus operandi* to implement in Poland, as in the letter of 6 July 1564, where he ponders the optimal approach to 'conquer or at least weaken' ("*espugnare o almeno raffreddare*") the heretics.[13] Expressions of this nature embody practical knowledge and expertise acquired by diplomats in the field, offering insight into the practices and strategies employed in the diplomatic endeavours of Commendone and Graziani. This facet holds increasing value in recent historical research pertaining to diplomatic history, which places greater emphasis on the individuals involved in diplomacy, prioritising the examination of individual diplomats, monarchs, and personal and informational networks (Plebani, Valeri,

11 Commendone stated: "*quanto più quel paese era infetto, tanto più conveniva provederle di pastore et tosto et bene et che, per desperata infirmità che habbiano figlioli, non devono i padri lasciar la cura loro né abbandonarli, che di simili infirmità d'heresia non more se non chi vuole et però non conviene di haver mai la cura per disperata*" (The more infected that country was, the more it was necessary to provide them with a shepherd promptly and well. Even though the children may have desperate infirmities, the fathers must not forsake their care or abandon them, only those who wish to die of such heretical infirmities. Hence, it is never fitting to relinquish care as if it were hopeless). NYPL, MssCol 603, first Register, Letter 18, Commendone to Borromeo, 27 February 1564, https://grazianiarchives.eu/s/graziani-archives/item/3808; cf. also NYPL, MssCol 603, Commendone to Borromeo, second Register, Letter 48, 14 August 1564, https://grazianiarchives.eu/s/graziani-archives/item/3845; NYPL, MssCol 603, fifth Register, Letter 119, 2 April 1565, https://grazianiarchives.eu/s/graziani-archives/item/3922; NYPL, MssCol 603, seventh Register, Letter 174, 4 November 1565, https://grazianiarchives.eu/s/graziani-archives/item/3952.

12 NYPL, MssCol 603, fourth Register, Letter 99, Commendone to Borromeo, 2 March 1565, https://grazianiarchives.eu/s/graziani-archives/item/3901; NYPL, MssCol 603, fourth Register, Letter 106, Commendone to Borromeo, 20 March 2023, https://grazianiarchives.eu/s/graziani-archives/item/3908.

13 NYPL, MssCol 603, second Register, Letter 44, Commendone to Borromeo, 6 July 1564, https://grazianiarchives.eu/s/graziani-archives/item/3840.

Volpini 2017; Sabbatini, Volpini 2011; Sowerby 2016). Of particular significance is Commendone's contemplation on the role of pontifical nuncios in challenging contexts like Poland, as he suggests to the Pope the allocation of greater authority to pontifical diplomats.[14]

Moreover, terms used in the document that denote religious experiences or denominations whose lexicon was only developed in subsequent centuries, such as *trinitarii* 'Trinitarians', referring to those associated with the Anti-Trinitarian movement, have also been retained. Specifically, Commendone's letter provides information about the movement's activities and prevalence in Poland and Lithuania.[15]

The transcription meticulously preserves the original forms, including frequent errors, present in the document. Nonetheless, certain adjustments have been introduced to enhance readability for users. Specifically, articulated prepositions have been modernised to adhere to contemporary usage norms. Furthermore, various linguistic anomalies have been rectified, such as the transformation of 'ij' at the end of words into 'ii', alongside the rectification of evident material or linguistic inaccuracies. Along similar lines, the practice of utilising festivities as a means of dating, as observed in Commendone's writings, has been retained. In instances involving lesser-known or movable celebrations, additional contextual information has been provided to render the summaries more accessible to users.

In a broader context, much thought was given to the user's experience in the configuration of the database, both in the process of generating summaries and in the extraction of metadata.[16] Specifically, a profile has been delineated for a prototypical database user, encompassing historians or individuals possessing a foundational understanding of the subjects and queries addressed within the letters. This characterisation has significantly informed our methodology, directing us to abstain from burdening the summaries with exhaustive discussions of topics and inquiries that have already received extensive coverage within historiography.

14 NYPL, *Ms. Div.*, MssCol 603, third Register, Letter 73, Commendone to Borromeo, 8 January 1565, https://grazianiarchives.eu/s/graziani-archives/item/3875.

15 NYPL, *Ms. Div.*, MssCol 603, first Register, Letter 12, Commendone to Borromeo, 16 January 1564, https://grazianiarchives.eu/s/graziani-archives/item/3802; cf. also NYPL, *Ms. Div.*, MssCol 603, first Register, Letter 25, Commendone to Leonardo Contarini, 25 March 1564, https://grazianiarchives.eu/s/graziani-archives/item/3818; NYPL, *Ms. Div.*, MssCol 603, third Register, Letter 70, Commendone to Borromeo, 26 and 27 December 1564, https://grazianiarchives.eu/s/graziani-archives/item/3872; NYPL, *Ms. Div.*, MssCol 603, third Register, Letter 71, https://grazianiarchives.eu/s/graziani-archives/item/3873; NYPL, *Ms. Div.*, MssCol 603, sixth Register, Letter 144, 13 June 1565, https://grazianiarchives.eu/s/graziani-archives/item/3947.

16 Starting from the output of user's queries. On this subject, cf. Drucker 2021, 76.

5 Exploring Relational Dynamics through a Database of Historical Documents

One of the database's most intriguing attributes is its relational structure. Within each regesto record, one can locate pertinent names of individuals and locations mentioned. Clicking on these items grants access to comprehensive records. Enhanced structured data amplifies visualisation options. In particular, heightened standardisation of individuals' occupations or titles fosters more effective linking of these entries. For instance, we opted to present the title 'cardinal' without further specifications **[fig. 5]**. This approach facilitates broader associations of this entry with a diverse array of individual items, thereby augmenting interconnections among items.

ITEMS **cardinale**

The following resources link to this item:

Filter by property All

1 of 6 1–25 of 134

hasOccupation

Title	Alternate label	Class
Acquaviva d'Aragona, Giulio	Qualifica	Person
Alberto di Hohenzollern	Qualifica	Person
Alberto VII d'Asburgo	Qualifica	Person
Alciati, Francesco	Qualifica	Person
Aldobrandini, Giovanni	Qualifica	Person
Altemps, Jacopo Annibale	Qualifica	Person
Babou de la Bourdaisière, Philibert	Qualifica	Person
Bandini, Ottavio	Qualifica	Person
Bianchetti, Lorenzo	Qualifica	Person
Bobba, Marcantonio	Qualifica	Person
Bolognetti, Alberto	Qualifica	Person
Bonelli, Michele	Qualifica	Person
Borromeo, Carlo	Qualifica	Person
Angennes (d'), Charles	Qualifica	Person

Figure 5 Results of the query 'cardinal' in the occupation item

Another example may show how a lack of precise information can sometimes favour an access to a distinct group of persons: by selecting the occupation *Membro del seguito di Commendone* 'member of Commendone's entourage', users can efficiently access all individuals

part of this group cited in the letters and inventory **[fig. 6]**. Had we categorised each individual according to their specific role within Commendone's entourage, it would have been impossible to provide the user with an overarching view of the group surrounding Commendone. This consolidated presentation serves as a compelling foundation for conducting social network analyses.

Title	Alternate label	Class
Arnolfo, Mauro	Qualifica	Person
Assalti, Marcantonio	Qualifica	Person
Boldieri, Benedetto	Qualifica	Person
Carga, Giovanni	Qualifica	Person
Leoni, Giovanni Battista	Qualifica	Person
Nauclerio, Prospero	Qualifica	Person
Pagliarini, Giovanni	Qualifica	Person
Pendasio, Federico	Qualifica	Person
Ruggieri, Fulvio	Qualifica	Person
Scappi, Giovanni Battista	Qualifica	Person
Schöneich (von), Kaspar	Qualifica	Person
Sirti, Vincenzo	Qualifica	Person
Sirti, Vincenzo	Qualifica	Person
Toledo (de), Francisco	Qualifica	Person
Tridapali, Giulio Cesare	Qualifica	Person
Viale, Francesco	Qualifica	Person

Figure 6 Results of the query 'Member of Commendone's entourage' (back-end)

Additionally, person entries of royals were linked to those of their spouses when mentioned. Consequently, users gain access to all correspondences sent and/or received by the individual, as well as entries pertaining to all their spouses, as observed in the case of Zigmund II August **[fig. 7]**.

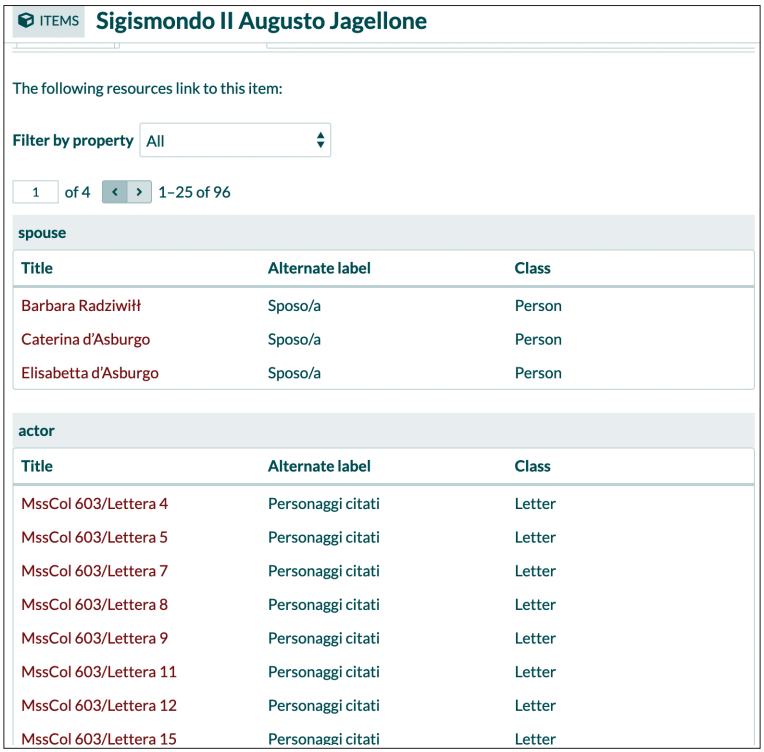


Figure 7 Results of items linked to a single individual: Sigismondo II Augusto Jagellone (back-end)

The database also includes documentation of familial relationships among individuals mentioned in the documents. This encompasses kin of Commendone and Graziani, along with members of prominent families like, for example, the Savorgnan. Users can also encounter this data within the linked resource, as exemplified in the case of Graziani, Antonio Maria **[fig. 8]**.

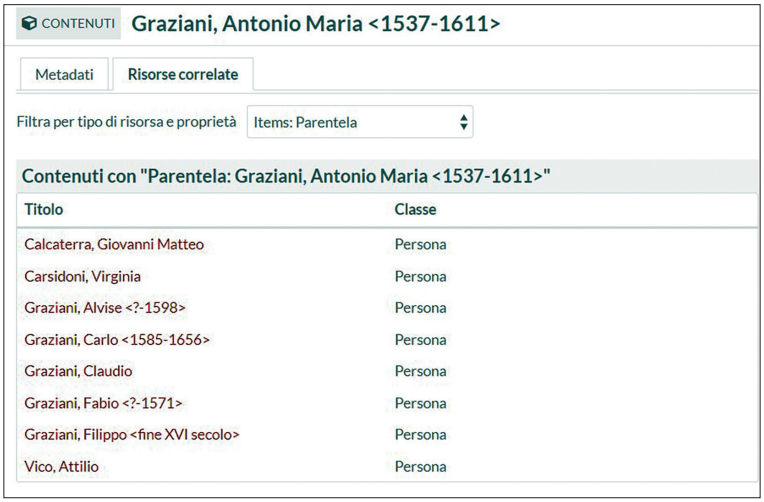


Figure 8 Results of the query for family relations of Antonio Maria Graziani (back-end)

The relational capabilities of the database are further evident in the interplay between letters, locations, individual entries, and the new version of the Graziani Archives' inventory (Raines, *supra*). Without delving into the specific details of the inventory, it is important to emphasise that within it a dedicated section details the most frequent senders and/or recipients of letters, as well as individuals whose names feature in the titles of works integrated into manuscripts or archival containers. This approach establishes a link between an individual's name, letters, and the other archival units of the Graziani Archives. This ongoing effort is poised to furnish users with the ability to observe, when seeking an individual, not only the summaries of documents wherein the person is mentioned or functions as the sender or recipient, but also whether they appear in other archival units within the Graziani Archives that are not summarised. Users can accomplish this by navigating to the linked resource associated with the entry.

The study of meticulously chosen documentation, inventory records, and research endeavours was vital to the establishment of the online portal. Our hope is that this portal will emerge as an invaluable research tool, enhancing prospects for data visualisation and facilitating novel modes of interpretation (Graham, Milligan, Weingart 2016). Rather than aiming to supplant physical document consultation, a key objective of this project is to furnish an initial gateway to the treasury of this archive's richness.

Bibliography

Biaudet, H. (1910). "Un chiffre diplomatique. Étude sur le code Nunz Polonia '27A' des archives secrètes du Saint Siège". *Annales Academiae Scientiarum Fennicae. Ser. B*, 2(4), 1-16.

Plebani, E.; Valeri, E.; Volpini, P. (a cura di) (2017). *Diplomazie. Linguaggi, negoziati e ambasciatori fra XV e XVI*. Milano: FrancoAngeli.

Drucker, J. (2016). "Graphical Approaches to the Digital Humanities". Schreibman, S.; Siemens, R.; Unsworth, J. (eds), *A New Companion to Digital Humanities*. Chichester: John Wiley & Sons, 238-50.

Drucker, J. (2021). *The Digital Humanities Coursebook. An introduction to Digital Methods for Research and Scholarship*. London; New York: Routledge.

Dudithius, A. (1992).*Epistulae. Pars I (1554-1567)*. Ed. by T. Szepessy, S. Kovács. Budapest: Akadémiai kiadó.

Epstein, F. (1937). s.v. "Teodoro I zar di Russia". *Enciclopedia Italiana*, vol. 33. https://www.treccani.it/enciclopedia/teodoro-i-zar-di-russia_%28Encicloped ia-Italiana%29/

Frost, R. (2015). *The Oxford History of Poland-Lithuania*. Vol. 1, *The Making of the Polish-Lithuanian Union (1385-1569)*. Oxford: Oxford University Press.

Gotor, M. (2013). s.v. "Ochino, Bernardino". *Dizionario biografico degli Italiani*, vol. 79, 90-7. https://www.treccani.it/enciclopedia/bernardino-ochino_%28Dizionario-B iografico%29/

Graham, S.; Milligan, I.; Weingart, S. (2016). *Exploring Big Historical Data: The Historian's Macroscope*. London: Imperial College Press.

Mariani, A. (2014). *I gesuiti e la nobiltà polacco-lituana nel tardo periodo sassone (1724-1763). Cultura e istruzione fra tradizione e innovazione*. Poznań: Instytut Historii UAM.

von Pastor, L. (1950). *Storia dei papi dalla fine del Medio Evo*. Vol. 7, *Pio IV (1559-1565)*. Transl. by A. Mercati. Roma: Desclée & C. Transl. of *Geschichte der Päpste seit dem Ausgang des Mittelalters*. Bd. VII, *Geschichte der Päpste im Zeitalter der katholischen Reformation und Restauration. Pius 4. (1559-1565)*. Freiburg im Breisgau: Herdersche Verlagshandlung, 1920.

Riasanovsky, N.V. (2010). *Storia della Russia. Dalle origini ai giorni nostri*. A cura di S. Romano. Bologna: Bompiani.

Sabbatini, R.; Volpini, P. (a cura di) (2011). *Sulla diplomazia in età moderna. Politica, economia, religione*. Milano: FrancoAngeli.

Sella, D. (1960). s.v. "Alciati della Motta, Giovanni Paolo". *Dizionario biografico degli Italiani*, vol. 2, 68-9. https://www.treccani.it/enciclopedia/alciati-della-motta-giovanni-paolo _%28Dizionario-Biografico%29/

Słonia, M. (ed.) (2021). *Atlas historyczny polski*. Vol. 2, *Ziemie polskie korony w drugiej połowie XVI wieku*. Warszawa: Instytut historii Pan.

Sowerby, T.A. (2016). "Early Modern Diplomatic History". *History Compass*, 14(9), 441-56. https://doi.org/10.1111/hic3.12329

Tüchle, H. (1959). s.v. "Eisengrein, Martin". *Neue Deutsche Biographie*, Bd. 4, 412-3. https://www.deutsche-biographie.de/pnd118681826.html#ndbcontent

Vitali, S. (2004). *Passato digitale. Le fonti dello storico nell'era del computer*. Milano: Mondadori.

Models of Data Extraction and Architecture in Relational Databases of Early Modern Private Political Archives
edited by Dorit Raines

A Puzzle with Missing Pieces
Extracting, Deciphering, and Digitally Rearranging Data in Antonio Maria Graziani Private Archives

Carlo Baja Guarienti
Università Ca' Foscari Venezia, Italia

Abstract The Graziani Archives web portal utilises the Omeka S digital platform to digitally reconstruct a substantial portion of the private archives of the apostolic nuncio Antonio Maria Graziani. The Graziani Archives project facilitates the digital rearrangement and analysis of data, addressing the complexities inherent in this private archives. This collection, which also encompasses the correspondence of Graziani's predecessor, Cardinal Giovanni Francesco Commendone, holds historical significance within the context of papal diplomacy during and after the Council of Trent. However, unlike the majority of the letters, the manuscript E105, comprising 1,157 letters authored by both diplomats, presents challenges due to a lack of explicit attribution, information regarding addressees, and dates or locations of composition. To differentiate between the letters authored by both nuncios, the present study employs simulations to extract metadata, resulting in quicker and more precise identification compared to manual examination.

Keywords Metadata collection. Data ingestion. Omeka S. Graziani Archives. Antonio Maria Graziani. Giovanni Francesco Commendone.

The research project, titled *Nuncio's Secret Archives: Papal Diplomacy and European Multi-denominational Societies before the Thirty Years War,* has the primary objective of establishing the Graziani Archives web portal through the utilisation of the Omeka S digital platform. This web portal endeavours to virtually reconstruct a

Edizioni Ca'Foscari

Studi di archivistica, bibliografia, paleografia 8
e-ISSN 2610-9093 | ISSN 2610-9875
ISBN [ebook] 978-88-6969-919-1 | ISBN [print] 978-88-6969-920-7

Peer review | Open access
Submitted 2023-10-03 | Accepted 2024-01-22 | Published 2025-05-22

DOI 10.30687/978-88-6969-919-1//012

substantial segment of the private archives of Giovanni Francesco Commendone and Antonio Maria Graziani, both distinguished diplomats from the late sixteenth and early seventeenth centuries. Their manuscripts were in time scattered across various hosting institutions in both Italy and the USA.[1]

The significance of this archives lies in both its completeness and continuity, as well as the importance of the two diplomats within the context of papal diplomacy during and after the Council of Trent. Giovanni Francesco Commendone (1524-1584), a Venetian, became a part of the Roman court under Pope Julius III and undertook crucial diplomatic assignments, such as presenting the bull of convocation of the Council in Trent to Emperor Ferdinand I in 1560. In subsequent years, he was sent to meet with the emperor and the king of Poland to address sensitive issues such as implementing the decrees of the Council, combating heresy, and dealing with the annulment of the marriage between Sigismund II Augustus Jagiellon and Catherine of Habsburg. In 1565, Commendone was appointed cardinal, and in 1566, he served as a legate to the imperial Diet. In 1570-71, he travelled to Vienna and Krakow once more to advocate for a Catholic league against the Turks and address the matter of Poland's throne succession (Caccamo 1982).

Antonio Maria Graziani (1537-1611) served as Commendone's secretary from 1560 until his death and accompanied him on all missions. Graziani also held the role of vice-legate in Poland in 1573, during the delicate phase following the appointment of Henry of Valois as king. After Commendone's demise, Graziani became one of the most trusted advisors of Cardinal Montalto, Alessandro Peretti, and was subsequently appointed as Bishop of Amelia in 1592 and as papal nuncio to Venice in 1596.

Giving the importance of the two nuncios, the design of the platform and its functionalities, particularly its capacity to render archival series preserved in various and distant institutions interoperable, prompt us to consider the challenges posed by a specific portion of the documentation. A notable example is the manuscript E105,[2] held in the Kenneth Spencer Research Library in Lawrence (Kansas), which contains 1,157 documents, primarily consisting of letters written by both Commendone and Graziani. The manuscript's miscellaneous nature is acknowledged at the beginning of the volume itself: *"Lettere parte per il card[ina]le Commendone e | parte per Ant[oni]o M[ari]a Graziani | dal 1566 sino al 1581"* (Letters from both Cardinal

1 https://nsa.unipr.it/. For the history of the archives, cf. briefly in this volume Raines, *supra*.

2 Lawrence, Kansas, Kenneth Spencer Research Library (KSRL), Graziani-Commendone Collection, Letterbooks-Commendone, MS E105.

Commendone and Antonio Maria Graziani, from 1566 to 1581).[3] However, the manuscript does not provide a clear distinction between the letters authored by the two diplomats. Consequently, to discern between the letters attributed to the cardinal and those attributed to his secretary, internal criteria must be applied. To historians, this distinction is essential due to the possibility that these letters may address similar topics but present different perspectives or focus on distinct subjects.

Besides the absence of explicit attribution and frequent omission of information on the addressee, many letters in this manuscript also lack a date or place of writing. However, the manuscript's somewhat imperfect chronological order does allow for the determination of at least the year, and sometimes the month, of most of these letters **[fig. 1]**.

Figure 1 Lawrence, Kansas, Kenneth Spencer Research Library, Graziani-Commendone Collection, MS E105, fol. n.n.

Furthermore, this manuscript presents additional challenges. Firstly, it is primarily written by a single hand, with only minor corrections made by a second hand (possibly that of Antonio Maria Graziani) **[figs 2-3]**.

3 KSRL, MS E105, fol. n.n. [2]*v*.

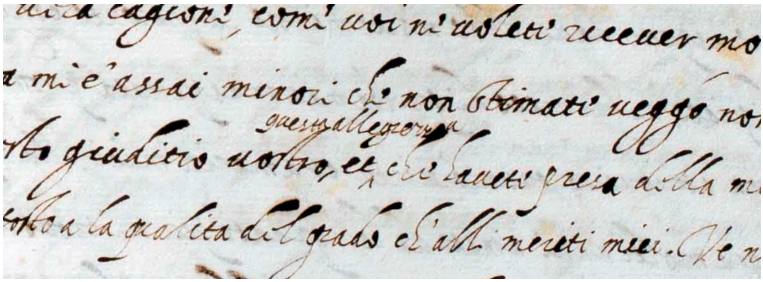


Figure 2 Lawrence, Kansas, Kenneth Spencer Research Library, Graziani-Commendone Collection, MS E105, fol. 40*r*

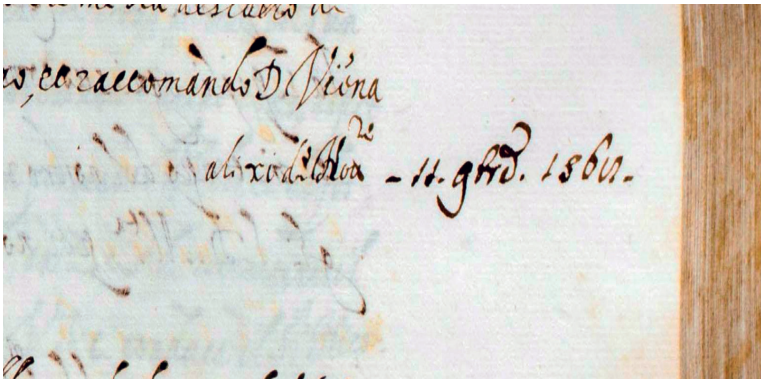


Figure 3 Lawrence, Kansas, Kenneth Spencer Research Library, Graziani-Commendone Collection, MS E105, fol. 12*r*

Additionally, there are some notes written by a third hand and, at a later stage, contributions by the Jesuit scholar Girolamo Lagomarsini (1698-1773), who extensively studied the archives in the second half of the eighteenth century and published some of Graziani's works (1745-76) **[fig. 4]**.[4]

4 For the life of Lagomarsini, cf. Arato 2004. After the publication of the *De scriptis invita Minerva*, Lagomarsini began an interesting dialogue with Lodovico Antonio Muratori: cf. Muratori 2023.

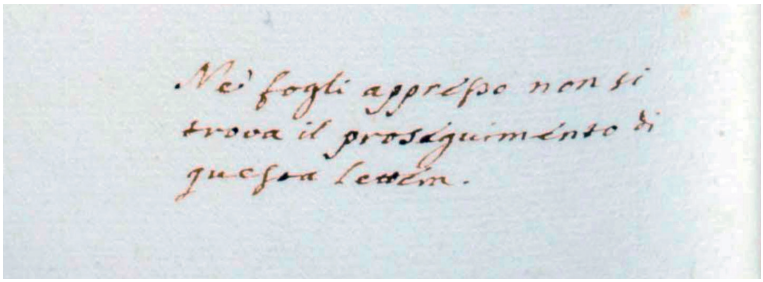


Figure 4 Lawrence, Kansas, Kenneth Spencer Research Library, Graziani-Commendone Collection, MS E105, fol. 226*v*

Despite its apparently unified origin, the manuscript seems to have been assembled by combining three distinct groups of sheets: the first group (fols 1a-52a) comprises undated letters written in the spring of 1565, shortly after Commendone's elevation to cardinalate; the second group (fols 53a-79a), starting after a few blank pages and introduced by the note "*del 1566*" (of 1566), contains letters written between November 1566 and March 1567; the third group (which adopts a different numeration: fols 1-379) contains letters written between September 1568 and March 1581. This compositional division is evident in the two series of cartulation: the sheets in the first two sections are numbered from 1a to 79a, while those in the third section are numbered from 1 to 379 **[figs 5-6]**.

Figure 5 Lawrence, Kansas, Kenneth Spencer Research Library, Graziani-Commendone Collection, MS E105, fol. 79a*r*

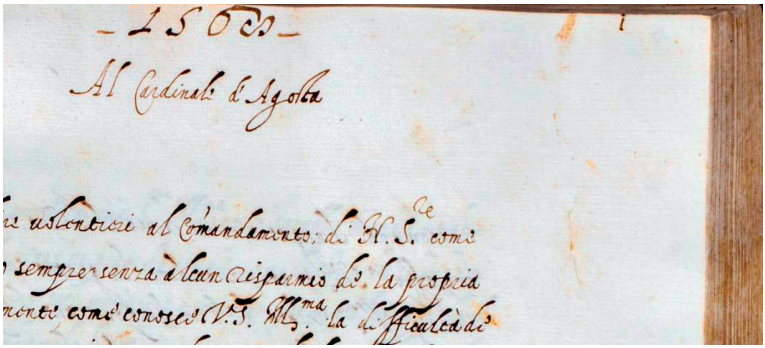


Figure 6 Lawrence, Kansas, Kenneth Spencer Research Library, Graziani-Commendone Collection, MS E105, fol. 1*r*

One possible explanation for the unusual structure of the manuscript is that it was assembled from three originally independent sections, which were later bound together, possibly when the third section had already been numbered. The fact that the sheets in the third section have been trimmed suggests that this section may have been created as a separate collection, either before or after being combined with the other two. The copyists of the manuscript were somewhat incompetent or careless, as they committed errors regarding commonplace areas, names, and Latin expressions, which would not be expected from individuals serving an important figure such as Commendone.[5] Unfortunately, without access to the originali, it is not possible to compare the copy with the correct original to identify the mistakes.

The most significant detail that remains elusive is the purpose behind the compilation of letters from various authors, encompassing diverse arguments and spanning a period of fifteen years, subsequently copied and bound together within the manuscript.[6] Although the manuscript presents several challenges of a diverse nature, our focus will be on the lack of crucial information in many letters, such as reliable attribution, the addressee, or the date, in an attempt to unravel at least part of its logic. By conducting an in-depth metadata harvesting and virtual reconstruction of the archives through the Graziani Archives web portal (still under construction at the time of writing this essay and online as of April 2024), the hope is to move closer to a solution.

5 Cf., for example, the "*Foggia/Foglia*" oscillation in Letters 825, 826 and 827 and the Latin grammatical errors in Letter 186.

6 The fact that some letters contain substantial corrections seems to exclude that we are dealing with a simple letter book, but rather points towards the hypothesis of a selection of letters destined for reworking or maybe publication.

Title: MS E105 / Lettera 1006

Sotto/Unità archivistica: KU, KSRL, Graziani-Commendone Coll., Letterbooks-Commendone, Ms. E105

Regesto veloce: Commendone ringrazia il cardinale di Perugia, Fulvio Della Cornia, per il suo impegno nel tentativo di riconciliare la famiglia Ranieri e i nipoti di Girolamo Mannelli, vescovo di Nocera. Commendone è stato informato di un'adunanza di banditi nella zona di Schifanoia, quindi consiglia al cardinale di far concludere la pace alle parti entro il raccolto per evitare ulteriori violenze. Si scusa se non potrà fare visita al cardinale durante il prossimo viaggio verso la Toscana e trasmette i saluti di Gabriele Del Monte, vescovo di Jesi.

Tipologia: lettera in registro copialettere

Numero documento: 1006

Estensione materiale: cc. 316r-v

Nome Mittente	Qualifica
Commendone, Giovanni Francesco	Cardinale

Nome Destinatario	Qualifica
Della Cornia, Fulvio	Cardinale

Luogo di redazione: Serra de' Conti

Luogo di ricezione:

Data di redazione: 22-06-1579

Edizioni del documento:

Personaggi citati:

Nome	Qualifica
Mannelli, Girolamo	Vescovo di Nocera
Del Monte, Gabriele	Vescovo di Jesi
Ranieri, famiglia	

Toponimi rilevanti:

Schifanoia

Note libere:

Figure 7 *Nuncio's Secret Archives Project*: Letter template

The simulations conducted on templates that served as a preliminary step to data ingestion in Omeka S, incorporating collected metadata, yielded results more rapidly than manual examination of the manuscripts **[fig. 7]**. However, they were found to be less precise. Although these simulations required more time compared to the corresponding operations performed through Omeka S, the advantages of using a relational database extend beyond mere speed of information retrieval. The database's utility lies in its capacity to address the complexity and depth of the investigation, particularly regarding qualitative data.

As an example, Letter 166 from the E105 manuscript, an important communication to Diego de Àvila regarding the challenging relationship between Commendone and Cardinal Zaccaria Dolfin, lacks topical and chronic date information. Conducting research to gather complete information from the corpus of letters, including the names of the addressee and mentioned persons (in this case, Pope Pius IV and Cardinals Carlo Borromeo, Zaccaria Dolfin, and Stanisław Hozjusz), yields an immediate response. Letter 155 from the MssCol 603 of the New York Public Library is indeed another copy of the same letter that contains the missing information. It reveals that the letter was written on 16 July 1565, in Heilsperg.[7]

Naturally this example provides us with date and place of the letter, an information that can become useful in understanding the whole mechanism of selecting the letters for the manuscript. Some inquiries though are more intricate. On 16 January 1572, an unidentified person (Commendone or Graziani) writes to a member of the Puccini family ("*signor Puccini*") explaining that due to uncertainties in the mission to Poland, the author cannot guarantee the safe return of the addressee's son to Italy. The Graziani archives include several letters sent to brothers Giovanni Battista and Agostino Puccini.By searching for Giovanni Battista OR[8] Agostino Puccini in the 'Addressee' field and combining it with the date in the 'Date' field, another copy of the same letter is found in the E97 manuscript of the Kenneth Spencer Research Library. This additional copy attributes the E105 letter to Commendone and confirms Giovanni Battista Puccini as the addressee.[9] Using Giovanni Battista Puccini's son name, Marcantonio, two other letters, this time found in the 62A archival

7 New York Public Library, Manuscript and Archive Division, MssCol 603, *Giovanni Francesco Commendone, Diplomatic Correspondence, 1563-65*, sixth Register, Letter 155, https://grazianiarchives.eu/s/graziani-archives/item/3959.

8 From now on, I will use the words 'OR' and 'AND' in capital letters to refer to the logical operators.

9 KSRL MS E105, fol. 227*r-v*, Letter 773, https://grazianiarchives.eu/s/graziani-archives/item/5759; KSRL, MS E97, fol. 124*v*, Letter 131, https://grazianiarchives.eu/s/graziani-archives/item/6801.

container, reveal that already on 30 November 1571 Puccini asked Antonio Maria Graziani to help him accompany his son, at that time novice at the Jesuit College in Bamberg, to Italy together with Vincenzo Sirti, member of Commendone's household. The second letter from Puccini to Graziani is dated 13 March 1572, where the former asked to conduct his son, at that time in Warsaw under the care of a merchant, Sebastiano Montelupi, back to Italy.[10] The structured and relational data in the Graziani Archives database helps in this case not only to discover the addressee of the letter, but to get acquainted with further details, included in other letters, through the relationships created between persons and letters.

The presence of multiple copies of the same letters in different collections raises further research questions: why were these copies created? How does their presence in different selections affect their significance for research and their role in the study? What connections and interactions should be established between them using Omeka S tools?

Unfortunately, many letters do not have a second copy, so filling in information gaps relies on replies or on similar/related letters, such as those exchanged with the same person in a nearby date. For instance, in Letter 758 of the E105 manuscript,[11] Graziani reports to Francesco Tranquillo Andreis that an "office" requested by him was performed by Commendone, with no further explanation. By searching for Andreis' name AND the month of October 1571, a letter from Andreis to Graziani in the Graziani archives in Vada clarifies the nature of the request: Andreis, who holds an ecclesiastic benefit in Ternavia, needs a letter of recommendation addressed to "Telegdino", the parish priest Miklós Telegdi, who is hostile to him for personal reasons.[12]

Another example can be found in Letter 682,[13] dated 13 June (the year is not expressed), where the sender writes to Giovanni Giacomo Diedo about various topics and mentions a seal he asked Diedo to procure. By searching for Giovanni Giacomo Diedo in the Graziani Archives database, a group of letters sent by Diedo to Graziani in

10 Archivio Graziani di Vada, Fondo Antonio Maria Graziani, b. 62A, Letter 46 dated 13 March 1572, https://grazianiarchives.eu/s/graziani-archives/item/4230; Letter 164, dated 30 November 1571, https://grazianiarchives.eu/s/graziani-archives/item/4374.

11 KSRL MS E105, fol. 220*v*, Letter 758, https://grazianiarchives.eu/s/graziani-archives/item/5744.

12 Archivio Graziani di Vada, Fondo Antonio Maria Graziani (Fondo AMG), b. 62A, Letter 105, https://grazianiarchives.eu/s/graziani-archives/item/4304.

13 KSRL MS E105, fol. 191*r-v*, Letter 682, https://grazianiarchives.eu/s/graziani-archives/item/5661.

June 1571 is found in the Graziani archives in Vada, which mentions the same topics and specifically refers to the seal.[14]

The letter numbered 742 is a complex case. The sender, unspecified, writes to the governor of Todi mentioning a letter he had recently written to Cardinal Girolamo Rusticucci regarding a license requested by the recipient. The date, 20 June of an unspecified year (likely 1571), is corrected by a different hand to July. By searching for Ludovico Cattaneo OR Evangelista Sbroiavacca in the 'Addressee' field AND for Girolamo Rusticucci in the 'People mentioned' field, a letter dated 17 July 1571, in the manuscript E97 of the Kenneth Spencer Research Library is found, where Commendone asks Cardinal Rusticucci for a license for Ludovico Cattaneo. This provides confirmation of the sender, recipient, and date of Letter 742.[15]

Various other operations can be performed with Omeka S on the Graziani Archives' documents to help understand single letters or groups. For instance, creating a direct link from a letter to its copy or reply, or to another related document, aggregating all items (letters, persons, places) referring to a particular subject matter (e.g., 'Commendone's family') or event (e.g., 'Commendone's mission to the Emperor, October 1568–April 1569'), and creating trans-archival chronological series of documents.

Furthermore, the possibility of continuously editing and increasing the data, inserting new information, and creating new relationships between items allows for modifying variables and aggregations, opening new study paths within the documentation, and hopefully filling in the gaps.

However, the quest for completeness does not overlook the role of uncertainty in constructing this (or any) digital archives. All questions, in the absence of an answer beyond a reasonable doubt, must be acknowledged in the digitisation process to provide users with a faithful representation of a varied and complex documentation.[16]

14 Fondo AMG, b. 62A, Letter 41, https://grazianiarchives.eu/s/graziani-archives/item/4225; b. 62A, Letter 70, https://grazianiarchives.eu/s/graziani-archives/item/4261; b. 62B, Letter 86, https://grazianiarchives.eu/s/graziani-archives/item/6341; b. 62B, Letter 93, https://grazianiarchives.eu/s/graziani-archives/item/6349.

15 KSRL, MS E97, fols 233*v*-234*r*, Letter 161, https://grazianiarchives.eu/s/graziani-archives/item/6838.

16 On the vast topic of uncertainty in digital humanities, cf., among others, Martin-Rodilla, Gonzalez-Perez 2018; Therón Sánchez et al. 2019; cf. also the growing collection dedicated to the topic by "Informatics": https://www.mdpi.com/journal/informatics/topical_collections/UDH.

Bibliography

Arato, F. (2004). s.v. "Lagomarsini, Girolamo". *Dizionario biografico degli Italiani*, vol. 63, 70-3.
https://www.treccani.it/enciclopedia/girolamo-lagomarsini_%28Dizionario-Biografico%29/

Caccamo, D. (1982). s.v. "Commendone, Giovanni Francesco". *Dizionario biografico degli Italiani*, vol. 27, 606-13.
https://www.treccani.it/enciclopedia/giovanni-francesco-commendone_(Dizionario-Biografico)/

Graziani, A.M. (1745-46). *De scriptis invita Minerva ad Aloysium fratrem libri XX*. 2 voll. Florentiae: Ex Typographio ad Insigne Apollinis in Platea Magni Ducis.

Marsili, M. (2002). s.v. "Graziani, Antonio Maria". *Dizionario biografico degli Italiani*, vol. 58, 801-4.
https://www.treccani.it/enciclopedia/antonio-maria-graziani_(Dizionario-Biografico)/

Martin-Rodilla, P.; Gonzalez-Perez, C. (2018). "Representing Imprecise and Uncertain Knowledge in Digital Humanities: A Theoretical Framework and ConML Implementation with a Real Case Study". García-Peñalvo, F.J. (ed.), *Proceedings of the 6th International Conference on Technological Ecosystems for Enhancing Multiculturality (TEEM 2018)* (Salamanca, 24-26 October 2018). New York: Association for Computing Machinery, 863-71.
https://doi.org/10.1145/3284179.3284318

Muratori, L.A. (2023). *Contro l'Inquisizione*. A cura di M. Al Kalak. Roma: Donzelli.

Therón Sánchez, R. et al. (2019). "Towards an Uncertainty-Aware Visualization in the Digital Humanities". *Informatics*, 6(3), 31.
https://doi.org/10.3390/informatics6030031

Models of Data Extraction and Architecture in Relational Databases of Early Modern Private Political Archives
edited by Dorit Raines

How to Digitally Reconstruct the History of an Early Modern Private Library?
Antonio Maria Graziani (1537-1611) and the Vicissitudes of His Books

Luca Iori
Università degli Studi di Parma, Italia

Abstract The paper illustrates the historical problems and the technical solutions adopted to create the Graziani Archives portal section dedicated to Antonio Maria Graziani's private library. The purpose of this section is to offer an entire catalogue of the volumes belonged to Graziani, trying to reconstruct the general outline and the internal articulation of Graziani's collection as well as the historical sources that allow us to virtually reassemble his own library.

Keywords Graziani Archives. Antonio Maria Graziani. Early modern private libraries. Digital catalogue. Library. Papal diplomat.

Summary 1 Introduction. – 2 Cataloguing Challenges and Solutions: Reconstructing Antonio Maria Graziani's Library. –3 Navigating Complexity: Designing a Comprehensive Digital Catalogue for Antonio Maria Graziani's Library.

1 Introduction

In this paper, I intend to elucidate the theoretical intricacies alongside the technical solutions that have been adopted in the development of a distinct section within the Graziani Archives online portal.[1]

1 The Graziani Archives portal (www.grazianiarchives.eu) was realised within the project *Nuncio's Secret Archives. Papal diplomacy and European multi-denominational*

Edizioni Ca'Foscari

Studi di archivistica, bibliografia, paleografia 8
e-ISSN 2610-9093 | ISSN 2610-9875
ISBN [ebook] 978-88-6969-919-1 | ISBN [print] 978-88-6969-920-7

Peer review | Open access
Submitted 2023-10-03 | Accepted 2024-01-22 | Published 2025-05-22

DOI 10.30687/978-88-6969-919-1//013

This dedicated section is centred upon the library amassed by Antonio Maria Graziani (1537-1611). The designated area, denoted as *La biblioteca di A.M. Graziani* (The Library of A.M. Graziani), aims to provide a comprehensive catalogue encompassing the literary collection that once belonged to this eminent personality within the Roman Church. Notably, Antonio Maria Graziani, a cultivated humanist, held esteemed positions as a Papal diplomat of elevated rank and as bishop of Amelia (Marsili 2002). My present discourse does not primarily underscore the significance of this compilation in terms of unravelling Graziani's multifaceted intellectual disposition (on this topic, Iori 2025). Instead, the emphasis is directed towards the rationale behind and the process by which we have conceptualised a specialised digital instrument to address specific historiographical quandaries.

2 Cataloguing Challenges and Solutions: Reconstructing Antonio Maria Graziani's Library

In designing our digital catalogue, we encountered two major challenges: the initial hurdle involved establishing the true scope and breadth of Graziani's library, given that his books were dispersed among larger collections. The primary and most consequential of these challenges emanates from the Graziani family library located in Vada, containing approximately 2,000 volumes, predominantly from the sixteenth and seventeenth centuries.[2] While this library underwent recent cataloguing, numerous volumes attributed to Graziani remained unidentified prior to our undertaking.[3] Another segment of Graziani's book collection found its way to the Kenneth Spencer Research Library at the University of Kansas (Lawrence, KS).[4] However, these latter volumes are not isolated as a distinct compilation and

societies before the Thirty Years War (PRIN2017 JMPYTA). Raines (*supra*) offers a detailed description of the structure and content of the portal. For a brief presentation of the *Nuncio's Secret Archives* (NSA) project, cf. Bonora 2023 and the project website `https://nsa.unipr.it/`.

2 For a brief description of the Graziani family library in Vada, cf. Corsini 2000, 132-9 and Corsini 2021, 104-5. A comprehensive catalogue of the complete library is absent; instead, only catalogues pertaining to specific categories of volumes are available: Corsini 1998 (incunables and sixteenth-century *quarto* editions); Fiorini 2002 (sixteenth-century *folio* editions); Corsini 2004 (sixteenth-century *ottavo* editions printed outside Italy); Garfagnini 2023 (seventeenth-century editions in *folio* and *quarto*); Corsini, Garfagnini 2025 (sixteenth-century editions in *ottavo* and smaller formats).

3 Incomplete lists of Graziani's volumes are published in Corsini 2021, 110-6.

4 The dispersion of the Graziani family archives and library will be illustrated in Raines, forthcoming.

have been catalogued by the library without specific attribution.[5] Additionally, a third group of books eluded our grasp due to historical vicissitudes, such as the spoliation of the Graziani family residences over the centuries that caused some dispersion,[6] as well as sales conducted by some descendants of Antonio Maria during the last century in a bid to generate income (Raines, *supra*).

Given these foundational considerations, our initial and foremost endeavour entailed the meticulous reconstruction of the boundaries of Antonio Maria's library, facilitated by the fusion of three distinct categories of sources. The first category comprises seventeenth-century book inventories presently archived in Vada. Of particular significance is what we term *Inventario 1* (Inventory 1), a list of Graziani's books that likely dates back to 1608 – three years preceding Graziani's demise.[7] *Inventario 1* is accompanied by an additional catalogue, referred to herein as *Elenco tematico* (Thematic List), subsequently compiled by a distinct scribe. This catalogue categorises the same set of books as presented in *Inventario 1*, classifying them into sixteen thematic groupings: *S. Scriptura et Sancti Presbiteri* (Sacred Scriptures and Saintly Presbyters); *Theologi Morales* (Moral Theology); *Libri Spirituales* (Spiritual Literature); *Grammatici* (Grammar Texts); *Poetici* (Poetic Works); *Epistularii* (Epistolary Works); *Historici* (Historical Texts); *Retorici* (Rhetorical Works); *Philosophici* (Philosophical Texts); *Mathematici* (Mathematical Works); *Oeconomici* (Economic Texts); *Politici* (Political Treatises); *Iuridici Canonici* (Canonical Legal Texts); *Civiles* (Civil Law); *Medici* (Medical Writings); *Varij* (Miscellaneous Works). A subsequent enumeration, denoted as *Inventario 2* (Inventory 2),[8] was drafted by the representatives of the Apostolic Chamber on 22 March 1611, a few days after Graziani's decease. The envoys of Pope Paul V, however, did not fully scrutinise Graziani's library, but managed to catalogue only a few volumes they found inside a crate stored in the room where Graziani passed away. Thus, *Inventario 2* enumerates no more than 19 titles: of these, 10 align with those

5 The Kenneth Spencer Research Library has reunited manuscripts from the Graziani family archives in the Graziani-Commendone Collection but did not create a specific collection for the books alienated from the Graziani library.

6 Cf., for example, a letter dated 8 September 1643 (Archivio Graziani di Vada (Rosignano Marittimo, Livorno) (AGV), *Fondo Antonio Maria Graziani* (FAMG), unit 299, "M.M." to Giovan Battista Graziani), describing the great damage suffered by the library after the looting of Villa Graziani in Celalba (near Perugia), occurred during the summer of 1643 in the context of the first War of Castro (1641-44).

7 AGV, *Fondo archivio di famiglia* (FAFG), unit 210/4, "Libri Illustrissimi et Reverendissimi Domini Episcopi Amerini". A first transcription and presentation of the document is provided in Corsini 2021, 106-10. For the dating of *Inventario 1*, cf. Corsini 2021, 114-5.

8 AGV, FAFG, unit 91/1, fol. 9*v*-10*r*.

present in *Inventario 1*, while 9 are conspicuously absent.

The second category of sources encompasses the physical volumes retained in both Vada and Kansas (of special relevance are those bearing inscriptions or dedications signifying ownership by Graziani), while the third one involves insights derived from the extremely rich correspondences of Graziani.[9] Through the synergistic analysis of these sources, we deduced that Antonio Maria's private library held no fewer than 173 titles. Among these, 163 were featured in seventeenth-century book inventories, with the remaining 10 absent from the inventories but supported by tangible evidence like ownership inscriptions and dedications, thus validating their inclusion in Graziani's library (more details in Iori 2025).

Our secondary challenge revolved around identifying the editions and the physical copies owned by Graziani. This task entailed a systematic comparison between inventory-listed items and volumes present in both Vada and Kansas. The outcome revealed that 36 volumes (32 in Vada; 4 in Kansas) contained ownership inscriptions or dedications linking them definitively to Graziani. An additional 44 volumes stored in Vada lacked such inscriptions or dedications but could reasonably be associated with Graziani's collection, aligning seamlessly with bibliographic details found in seventeenth-century inventories. Regrettably, the remainder of the works listed in these same inventories – totalling 93 items – failed to correspond to any existing physical copies and must consequently be categorised as historical copies (Iori 2025). In summary, our current knowledge allows us to confidently identify 46% of the physical copies formerly possessed by Graziani (80 works out of 173). For the portion that remains without attribution, the inventories at least provide insight into titles, authors, and, in some instances, specific editions.

9 Most of these correspondences, which are largely unpublished, are preserved at the Graziani family archives in Vada and at the Kenneth Spencer Research Library. The NSA project members have systematically examined this massive documentation (cf. Raines, forthcoming), to retrieve relevant information about the history of Graziani's library (cf. Iori 2025).

3 Navigating Complexity: Designing a Comprehensive Digital Catalogue for Antonio Maria Graziani's Library

Upon amassing the requisite data for our digital portal, we initiated an exploration into the optimal digital tool for organising and presenting this information in a functional and historically pertinent manner. Initially, we contemplated the construction of a dual-purpose instrument, aiming to delineate distinct treatment for the physical copies owned by Graziani and the historical copies – those books that lack attribution. In regard to the former, our intention was to formulate individualised catalogue entries for each volume, encompassing comprehensive metadata, including the diverse material attributes that differentiate each extant copy, such as provenance notes, dedications, and apostils. Conversely, the historical copies would have been relegated to a separate listing, consolidated within a static page. Our design prioritised the thorough description of the extant physical copies, with the intent to digitally reconstruct and enable virtual engagement with the surviving segment of Graziani's collection.

However, we soon realised that such an approach held the potential to distort the essence of Graziani's library by omitting pertinent data. This partitioning of the library into two divergent and disparately treated sections – the surviving physical copies on one hand and the historical copies on the other – would obscure the comprehensive view of Graziani's library. This unintended consequence would relegate the historical copies to a peripheral status within the portal, despite their essential role in reconstructing Graziani's intellectual background (Iori 2025). Additionally, this configuration would have obfuscated the differentiation between inventoried and non-inventoried volumes, necessitating users to laboriously inspect each of the 80 individual catalogue entries dedicated to physical copies in isolation, thus negating a holistic overview. Furthermore, users would have been denied insight into the sequential arrangement of the books as listed in the book inventories, a structural element that assumes significance, especially in the context of *Inventario 1*. Indeed, this deliberate order suggests thematic subcategories within the collection, potentially reflecting Graziani's conceptualisation of their thematic relevance to his scholarly pursuits – for instance, the discernable division between Latin and vernacular books,[10] or the sequential arrangement of thematically coherent works, such as theological books, historical works, and others (Iori 2025).

Given these considerations, we opted for an alternative digital tool – one that can effectively capture the overarching outline

10 Cf. e.g. the headings "Testus [sic!] Canonici" and "Libri Volgari" in AGV, FAFG, unit 210/4, "Libri Illustrissimi et Reverendissimi Domini Episcopi Amerini", fol. 1*r*, 3*r*.

and internal structure of Graziani's library, alongside the historical sources facilitating its reconstruction. The idea was to integrate this tool into a single web page, commencing with a succinct introduction detailing the collection's history and elucidating the principles underpinning our cataloguing methodology **[fig. 1]**.

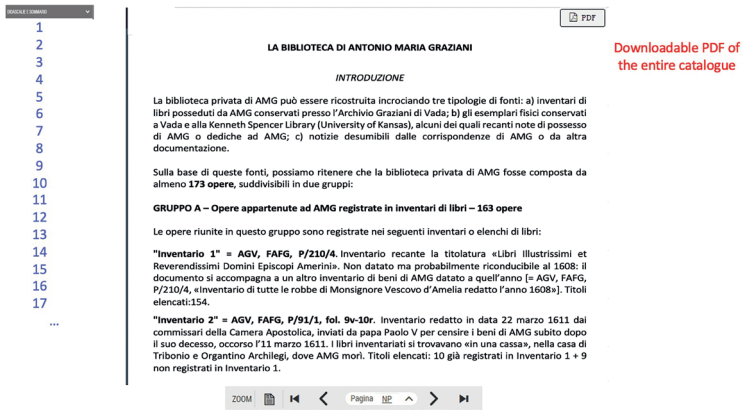


Figure 1 Initial mock-up of the Graziani's library section

Subsequently, the catalogue itself would have followed, incorporating the 173 titles constituting Graziani's private library. We further hypothesised that these titles would be sequentially numbered and grouped into two primary categories. The first group would encompass the items catalogued within seventeenth-century book inventories (163 items), presented in the sequence dictated by these inventories **[fig. 2]**.

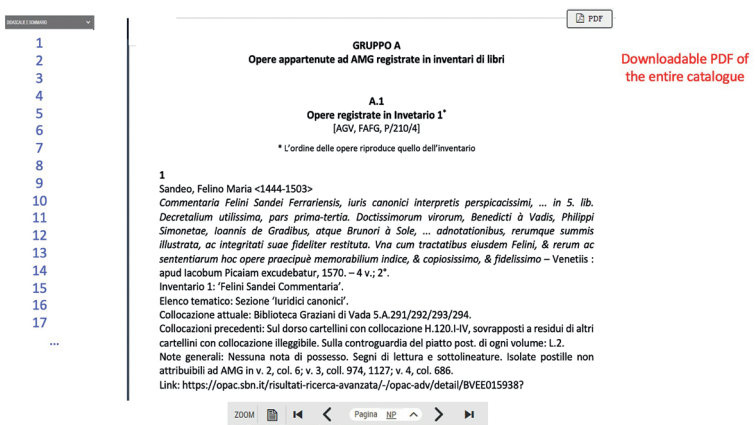


Figure 2 The books listed in *Inventario 1* as presented in the initial mock-up

The second group would then comprise a mere 10 items, numbered 164 to 173, cataloguing physical copies not documented in the inventories but marked by ownership inscriptions or dedications attributed to Graziani.[11]

The mock-up envisaged that each item within both groups would be conveniently accessible via the left-hand column, facilitating direct access by clicking on the corresponding number, thus obviating the need for exhaustive scrolling **[fig. 3]**.

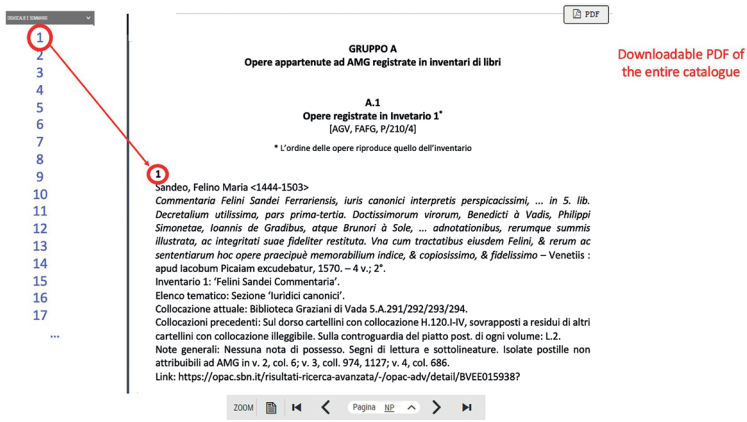


Figure 3 Direct access to each volume from a left-hand column

Moreover, a uniform descriptive structure was applied to each item, adhering to standardised templates. These templates were devised to impart the highest attainable levels of historical and bibliographic information in a uniform format. It was decided that divergent templates would be employed based on whether an item is an extant physical copy or a historical one. To illustrate this, we will consider some examples, moving from maximal to minimal information disclosure.

One highly informative entry revolves around an edition of Livy presently located in Kansas, gifted to Graziani by his pupil and young friend Mikołaj Tomicki.[12] This entry is assigned number 85 within our digital catalogue. The entry was thus structured across five hierarchical levels: the first encompassed bibliographical data (author, title, edition); the second pertained to the seventeenth-century inventories; the third level concerned current physical location (Vada or Kansas) and shelf mark; the fourth level documented material

11 Possible reasons for their exclusion from *Inventario 1* and *Inventario 2* will be discussed in Iori 2025.

12 On the relationship between Graziani and Tomicki, cf. Barzazi 2025.

attributes distinguishing the physical copy (ownership inscriptions, dedications, apostils by Graziani); and the fifth and final level provided a hyperlink to an entry of another online catalogue furnishing comprehensive metadata for the edition.

85
Livius, Titus
T. Livii Patavini Romanae Historiae Principis Decades Tres, Cvm Dimidia, Partim Caelii Secvndi Cvrionis industria, partim collatione meliorum codicum iterum diligenter emendatae. Eiusdem Caelii Praefatio, summam continens de mensuris, ponderibus, réq[ue] nummaria Romanorum & Graecorum… Simonis Grynaei de utilitate legendae historiae. Bartholomaei Marliani de origine urbis Romae Compendium… L. Flori Epitomae,… Doctorum uirorum… Annotationes, Glareani annotationibus… insertae – Basileae: per Ioannes Heruagios, Anno M.D.L.V. Mense Septembri); 2°.
Inventario 1: "Titi Livij Patauini Romanae Historiae. Cum reliquis libris".
Elenco tematico: Sezione "Historici".
Collocazione attuale: KSRL, Summerfield G191.
Collocazioni precedenti: Nessuna traccia.
Dedica: Sul frontespizio "Nicolaus Thomicius suo Antonio Mariae Gratiano". Sul verso della *praefatio* invocazione al libro di Antonio Maria Graziani: "Furem fac moneas, liber, rapaces / ad se contineat manus, crucemque, / si te surripias, statim mirare / surreptum metuens crucis reponet. / Ant. Mar. Grat". (= Giovanni Pietro Artemisio, *carmen* 95; con leggere modifiche).
Note generali: Postille di Antonio Maria Graziani, con segni di lettura e sottolineature, su quasi ogni pagina del volume.
Link: `https://catalog.lib.ku.edu/cgi-bin/Pwebrecon.cgi?BBID=3868964`.

Other items described historical copies solely known through the textual descriptions found within seventeenth-century inventories. In the luckiest cases, the detailed inventorial descriptions not only facilitate the identification of the title and author but also discern the specific edition owned by Graziani. Such an instance is exemplified by the commentary on the Psalms authored by Stanislaw Reszka and published in Krakow in 1591,[13] denoted as item 16 within our catalogue. In this context, the descriptive structure encompassed three levels: bibliographical information (author, title, edition, with the latter two enclosed in square brackets to indicate their historical nature); references to seventeenth-century inventories; and a hyperlink to the entry of another catalogue providing extended metadata. Notably, levels dedicated to the location and material attributes were understandably absent, given that the physical copy is no longer extant.

13 Cf. *Inventario 1*: "Paraphrasis in septem psalmos penitentiales Rescij Craccoviae in Architypo" (AGV, FAFG, unit 210/4, "Libri Illustrissimi et Reverendissimi Domini Episcopi Amerini", fol. 1*r*).

16
Reszka, Stanisław <1544-1600>
[*Paraphrasis in septem psalmos, quos vocant Poenitentiales S. Rescii. Accessit eiusdem Panegyricus* – Cracoviae: in Architypo. regia et ecclesiastica Lazari, 1591; 12°].
Inventario 1: "Paraphrasis in septem psalmos penitentiales Rescij Craccoviae in Architypo".
Elenco tematico: Sezione "S. Scripturae et Sancti Presbiteri".
Note generali: Unica edizione dell'opera pienamente riconducibile alla dicitura di Inventario 1.
Link:`https://opac.sbn.it/risultati-ricerca-avanzata/-/opac-adv/index/14/ITICCUNAPE017364?fieldstruct%5B1%5D=ricerca.parole_tutte%3A4%3D6&fieldvalue%5B1%5D=Reszka%2C+Stanis%C5%82aw+&fieldaccess%5B1%5D=Any%3A1016%3Anocheck&struct%3A1001=ricerca.parole_almeno_una%3A%40or%40&sort_access=Data_ascendente%3A+min+31%2C+min+3086%2C+min+5003`.

On certain occasions, however, the book descriptions present in seventeenth-century inventories only provide information regarding authors and titles. Such circumstances are exemplified by Francesco Sansovino's *Cronologia del mondo*,[14] featured as item 113 within our catalogue. In this instance, our catalogue entry continued to comprise three descriptive levels, albeit configured in a distinct manner. The initial level was dedicated to the 'author and title'. The second level, pertaining to the inventory details, remained unaltered. Meanwhile, the third level explicitly acknowledged our inability to ascertain the edition owned by Graziani. However, when pertinent, this level also denoted the presence of the *editio princeps* (first edition) of the work.

113
Sansovino, Francesco <1521-1583>
Cronologia del mondo
Inventario 1: "Cronologia del Mondo di messer Francesco Sansovino".
Elenco tematico: Sezione "Historici".
Note generali: Edizione non identificata. Editio princeps: *Cronologia del mondo di m. Francesco Sansouino diuisa in tre libri. Nel primo de' quali s'abbraccia, tutto quello ch'è auuenuto cosi in tempo di pace come di guerra fino all'anno presente. Nel secondo, si contiene vn catalogo de regni, & delle signorie, che sono state & che sono, con le discendenze & con le cose fatte da loro di tempo in tempo. Nel terzo, si tratta l'origine di cinquanta case illustri d'Italia, co soccessi de gli huomini eccellenti di quelle, & con le dipendenze & parentele fra loro. Con tre tauole* – In Venetia: nella stamperia della Luna, 1580; 4°.

Finally, the most regrettable scenario arises when the data furnished by book inventories prove inadequate to even discern the specific

14 Cf. *Inventario 1*: "Cronologia del Mondo di messer Francesco Sansovino" (AGV, FAFG, unit 210/4, "Libri Illustrissimi et Reverendissimi Domini Episcopi Amerini", fol. 3*r*).

work in question. Such an instance is exemplified by the enigmatic title *Auisi della Cina*,[15] as represented by item 154 within our catalogue. In such cases, the descriptive structure further diminished across the three levels. In the initial level, we presented the book description from *Inventario 1* enclosed within square brackets. The second level faithfully documented the seventeenth-century inventories, maintaining the conventional format. The third and final level endeavoured, to the extent possible, to offer identification suggestions in situations where the available information remains sparse.

154
["Auisi della Cina"]
Inventario 1: "Auisi della Cina".
Elenco tematico: Sezione "Historici".
Note generali: Opera non identificata. Possono essere le seguenti:
1. *Auuisi della Cina et Giapone del fine dell'anno 1586* – In Milano: per Pacifico Pontio, 1588; 8°.
2. *Auuisi della Cina et Giapone del fine dell'anno 1586. Con l'arriuo delli signori Giaponesi nell'India. Cauati dalle lettere della Compagnia di Giesù. Riceuute il mese d'ottobre 1588* – In Roma: appresso Francesco Zannetti, 1588; 8°.
3. *Auuisi della Cina, et Giapone del fine dell'anno 1587. Con l'arriuo de' signori giaponesi nell'India. Cauati dalle lettere della Compagnia di Giesù, riceuute il mese d'ottobre 1588* – In Venetia: appresso i Gioliti, 1588; 8°.
4. *Avvisi della Cina et Giapone, del fine dell'anno 1586, con l'arriuo delli signori giaponesi nell'India, cauati dalle lettre della Compagnia di Giesù* – In Anuersa: appresso di Christophoro Plantino architypographo regio, 1588; 8°.

Naturally, these examples do not encompass the entirety of the descriptive frameworks we formulated. Nevertheless, they may prove adequate to elucidate the configuration and the substance of the portal section dedicated to Graziani's library as originally conceived. The approach taken was intended to benefit users, primarily historians and scholars specialising in the history of books, aiming to provide them with an immediate overview of the contents of Graziani's library, without necessitating an extensive amount of clicking through numerous items. Nevertheless, the blueprint of the tool we delineated appeared to possess certain deficiencies. Admittedly, it could be construed as somewhat conventional and too similar to catalogues in paper form; above all, the user's experience could potentially have been restricted by the extensive nature of the text, which at that juncture encompassed not only the customary identifiers of the edition (such as author, title, place of publication, printer, date, and format), but also an array of supplementary particulars concerning the individual copy in question. As a result, users would have been compelled to

15 Cf. *Inventario 1*: "Auisi della Cina" (AGV, FAFG, unit 210/4, "Libri Illustrissimi et Reverendissimi Domini Episcopi Amerini", fol. 4*r*).

scroll through a lengthy list of entries, overloaded with information, which, due to the excess of data, ran the risk of obscuring the overall character of Graziani's book collection.

Thus, we set out to exploit the expansive capabilities afforded by contemporary digital humanities and, above all, the extensive features offered by Omeka S, to develop a catalogue structure more aligned with the standards of interchangeability and accessibility of information that characterise contemporary relational databases. In this perspective, we remained steadfast in our conviction that the presentation of Antonio Maria Graziani's collection, encompassing all associated books, within a comprehensive catalogue organised in accordance with the previously outlined two groups (after a concise introduction to the catalogue itself), was well-justified **[fig. 4]**. However, we have opted to condense this presentation by only preserving information regarding the edition, or at the very least, the author's name and title in instances where the edition remains unidentified **[fig. 5]**.

Figure 4 Introduction to Graziani library's catalogue as presented in the Graziani Archives portal

In addition, each catalogue entry has been hyperlinked to a corresponding Omeka S item providing interested users with the specifications of each entry. In fact, intricate and pertinent information concerning citations within the three inventories, potential identifications, particulars about copies (inclusive of ownership annotations, dedications, and apostils by Graziani himself), along with a hyperlink directing to an Online Public Access Catalogue for more extensive information, have all been incorporated into dedicated Omeka S items

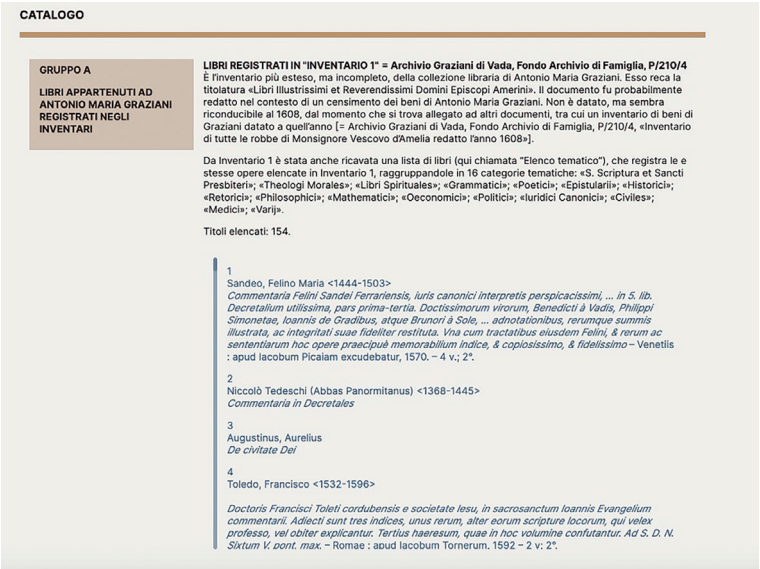


Figure 5 Catalogue of Graziani's library (identification of title and edition) as presented in the Graziani Archives portal

for each individual title. And these items are structured according to the standardised templates previously defined. These same items also provide, for the benefit of the user, all the information regarding the edition, or the author's name and title in instances where the edition remains unidentified. Furthermore, like every other item in our portal, the library items host a specific field that defines their document class – in this case: *Libro* (Book) – facilitating users who wish to search only for this type of document within the database. Finally, another field assigns a unique inventory number to each item, thereby preventing any confusion with other items stored in the database **[figs 6-7]**.

We believe the solution adopted of this specific tool effectively furnishes a portrayal of Antonio Maria Graziani's collection that is suitably intricate and historically trustworthy. At the same time, we feel that such a structured catalogue offers a balanced compromise between the needs of researchers and the capabilities/functionality provided by current digital tools. At least, this is how we have tried to combine the pursuit of innovative digital tools and the expectations of historical research. These kinds of efforts, perhaps among the most challenging in digital history, are likely to increasingly confront historians and data scientists in the years to come.

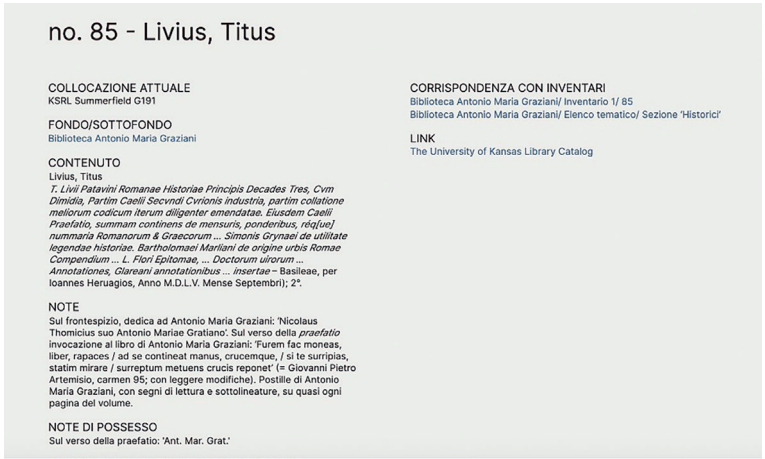


Figure 6 Livy's edition (Biblioteca Antonio Maria Graziani / 85) as described by a Book item template in Omeka S in the Graziani Archives portal

Figure 7 "*Auisi della Cina*" (Biblioteca Antonio Maria Graziani / 154) as described by a Book item template in Omeka S in the Graziani Archives portal

Bibliography

Barzazi, A. (2025). "Educating the Catholic Nobleman. Projects and Models in Padua and Poland". *Rocznik Filozoficzny Ignatianum*, 31(2), no. monogr., *Związki polskich elit z kulturą Republiki Weneckiej*.

Bonora, E. (2023). "Gli archivi segreti del nunzio". *Riforma e movimenti religiosi*, 13, 177-84.

Corsini, M. (1998). *Gli incunaboli e le cinquecentine in 4° della biblioteca di Villa Graziani* [tesi di laurea]. Pisa: Università di Pisa.

Corsini, M. (2000). "La biblioteca e l'archivio Graziani di Vada". *Rara volumina*, 1(2), 127-40.

Corsini, M. (2004). *Le edizioni del XVI secolo straniere in 8° della biblioteca di Villa Graziani: Descrizione e analisi bibliologica* [tesi di laurea]. Pisa: Università di Pisa.

Corsini, M. (2021). "La biblioteca del vescovo Antonio Maria Graziani: storia e caratteristiche della biblioteca di un nunzio apostolico all'epoca della Controriforma". Del Grazia, C.; Fiasconi, L. (a cura di), *La Biblioteca: crocevia e connessione di mondi*. Pisa: Edizioni ETS, 101-17.

Corsini, M.; Garfagnini, E. (2025). *"Nova et Vetera". La Biblioteca Graziani di Vada tra inventari e cataloghi. Catalogo delle edizioni del XVI secolo in 8° e formati minori*. Viareggio: Edizioni La Villa.

Fiorini, B. (2002). *Le cinquecentine in folio (2°) della Biblioteca di Villa Graziani.* Typewritten catalogue held at the Graziani family library in Vada.

Garfagnini, E. (2023). *Una biblioteca privata: i Graziani e i loro libri. Analisi delle edizioni del XVII secolo in folio e in 4° della biblioteca di Villa Graziani di Vada* [tesi di laurea]. Firenze: Università di Firenze.

Iori, L. (2025). "The Library of Antonio Maria Graziani. Books and Papal Diplomacy Between Italy and Poland in the Late Sixteenth Century". Bonora, E. (ed.), *The Church of Rome and Multi-confessional Europe in the Early Modern Period*. Rome: Viella.

Marsili, M. (2002). s.v. "Graziani, Antonio Maria". *Dizionario biografico degli Italiani*, 58, 801-4.
https://www.treccani.it/enciclopedia/antonio-maria-graziani_(Dizionario-Biografico)/

Raines, D. (a cura di) (forthcoming). *Archivio al confine. Eredità e potere nell'Archivio Graziani (XIV-XX secolo)*.

Studi di archivistica, bibliografia, paleografia

1. Raines, Dorit (a cura di) (2012). *Biblioteche effimere. Biblioteche circolanti a Venezia (XIX-XX secolo)*.

2. Minuzzi, Sabrina (a cura di) (2013). *Inventario di bottega di Antonio Bosio veneziano (1646-1694)*.

3. Pistellato, Antonio (a cura di) (2015). *Memoria poetica e poesia della memoria. La versificazione epigrafica dall'antichità all'umanesimo*.

4. Zanetti, Melania (a cura di) (2018). *Dalla tutela al restauro del patrimonio librario e archivistico. Storia, esperienze, interdisciplinarietà*.

5. Brunello, Mauro; De Martino, Valentina; Speranza Storace, Maria (a cura di) (2020). *Oltre le mostre*.

6. De Rubeis, Flavia; Rapetti, Anna (a cura di) (2023). *«Con licenza de' Superiori». Studi in onore di Mario Infelise*.

7. Zanetti, Melania (a cura di) (2024). *La legatura dei libri antichi. Storia e conservazione*.

Historical research involves particular challenges when it comes to organizing and analysing data, especially due to the unstructured nature of historical narratives and the complexity of archival contexts unfit for a binary environment. This complexity is further compounded when dealing with private collections that lack standard metadata models. The essays included in this publication are penned by scholars renowned for their expertise in digital humanities and historical research, providing multidimensional insights into the evolving landscape of historiography. Through meticulous examination, they illustrate the transformative power of digital tools in reshaping the methodologies of historical inquiry, augmenting traditional practices with innovative approaches. By addressing these issues, scholars can better navigate the intricacies of historical narratives and contribute to a deeper understanding of the past.

Università
Ca'Foscari
Venezia