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
© 2025 Bloch, Santana, Vasques Filho, Bojanowski | CC BY 4.0
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.

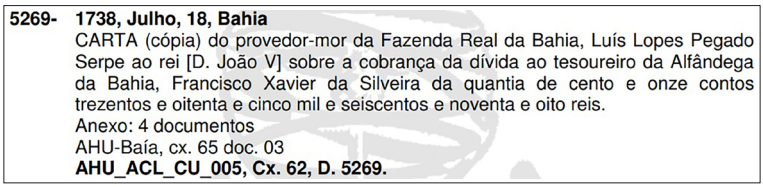
5269- **1738, Julho, 18, Bahia**
CARTA (cópia) do provedor-mor da Fazenda Real da Bahia, Luís Lopes Pegado Serpe ao rei [D. João V] sobre a cobrança da dívida ao tesoureiro da Alfândega da Bahia, Francisco Xavier da Silveira da quantia de cento e onze contos trezentos e oitenta e cinco mil e seiscentos 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]**.

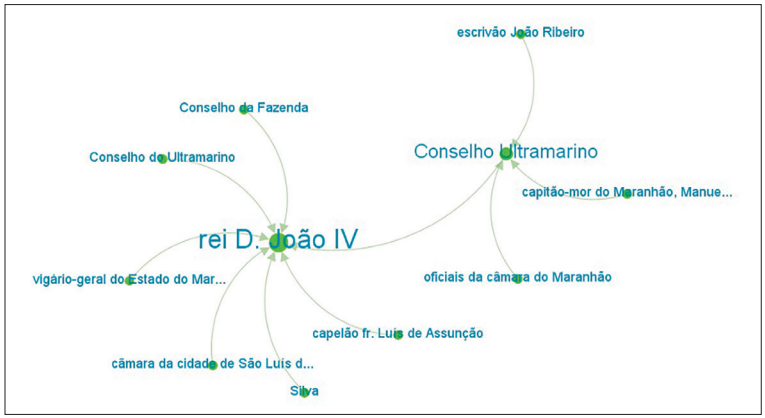

Figure 8 Example of a network connecting senders to recipients of correspondences from our corpus

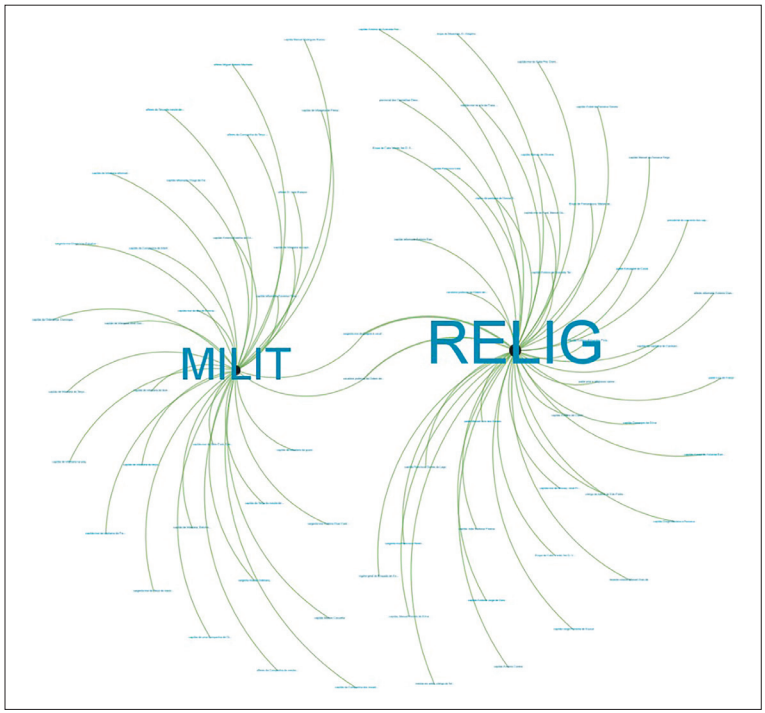

Figure 9 Example of a bipartite network of people and their respective organisation type

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