magazén

International Journal
for Digital and Public Humanities

e-ISSN 2724-3923

Vol. 6 – Num. 2 December 2025

Edizioni
Ca'Foscari

e-ISSN 2724-3923

magazén
International Journal
for Digital and Public Humanities

Editors-in-chief
Franz Fischer
Diego Mantoan

Edizioni Ca' Foscari - Venice University Press
Fondazione Università Ca' Foscari
Dorsoduro 3246, 30123 Venezia
URL https://edizionicafoscari.unive.it/en/edizioni/riviste/magazen/

magazén
International Journal
for Digital and Public Humanities

Semestral journal

Head office Università Ca' Foscari Venezia | Venice Centre for Digital and Public Humanities | Dipartimento di Studi Umanistici | Palazzo Malcanton Marcorà | Dorsoduro 3484/D, 30123 Venezia, Italia | magazen@unive.it

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

[ve]dph
Venice Centre for Digital and Public Humanities

PEER REVIEW OPEN ACCESS

Certificazione scientifica delle Opere pubblicate da Edizioni Ca' Foscari: tutti i saggi pubblicati hanno ottenuto il parere favorevole da parte di valutatori esperti della materia, attraverso un processo di revisione anonima sotto la responsabilità del Comitato scientifico della rivista. La valutazione è stata condotta in aderenza ai criteri scientifici ed editoriali di Edizioni Ca' Foscari.
Scientific certification of the works published by Edizioni Ca' Foscari: all essays published in this issue have received a favourable opinion by subject-matter experts, through an anonymous peer review process under the responsibility of the Advisory Board of the journal. The evaluations were conducted in adherence to the scientific and editorial criteria established by Edizioni Ca' Foscari.

magazén e-ISSN 2724-3923
Vol. 6 – Num. 2 – December 2025

Table of Contents

URL https://edizionicafoscari.unive.it/en/edizioni4/riviste/magazen/2025/2/
DOI http://doi.org/10.30687/mag/2724-3923/2025/02

e-ISSN 2724-3923

magazén
Vol. 6 – Num. 2 – December 2025

A Very Brief Introduction and Summary

Franz Fischer
Università Ca' Foscari Venezia, Italia

Diego Mantoan
Università degli Studi di Palermo, Italia

Barbara Tramelli
Università Ca' Foscari Venezia, Italia

Reaching our sixth year in a row,[1] we are glad to close 2025 with the present issue that is yet another exciting chapter of *magazén*'s engagement in the field of digital and public humanities. The past year brought even wider recognition of the journal's continuous efforts to serve as a high quality publication venue for innovative research across traditional disciplinary boundaries. Scopus ranked *magazén* among the Top 25% of academic journals for Literature and Literary Theory. Simultaneously, the Italian ANVUR confirmed the journal's scientific status across the boards for Classics, Archaeology and History while it was classified top tier status 'Grade A' in the area of Art History. We are confident that other indices and disciplinary sectors will follow.

In the meantime *magazén* switched to a rolling basis publication mode, enabling continuous submission and publication of articles instead of waiting for specific issue deadlines. Even if articles are still gathered in issues and volumes, this practice allows for faster dissemination and increases the efficiency of the editorial workflow.

As an immediate result of this newly gained flexibility this issue brings together articles on Cultural Heritage, Digital Humanities and Literary Studies with two articles from the field of Classics which

1 This introduction paper was mutually agreed on by the editors of *magazén* with the precious support of the Journal Manager Elisa Corrò, who was instrumental in coordinating the editorial work of this issue.

stem from the international conference *Classical Texts in Digital Media II - Digital Methods for Editing and Studying Ancient Texts* held in the Venetian lagoon on San Servolo island in June 2025.[2] The proceedings of this conference will be scattered across the current and forthcoming issues while still recognisable as a coherent collection of conference papers focussing on the current state of digital methods for editing and studying ancient texts.

The present issue starts with a contribution by Ulrike Henny-Krahmer, Fernanda Alvares Freire, and Erik Renz, and it investigates the role of lecture series in the consolidation of Digital Humanities as a field of research. Through a quantitative analysis of DH lecture series conducted across European institutions in the past decade, the article shows how these formats function as infrastructures of scholarly exchange, connecting institutions, researchers, disciplines, and research topics.

The second paper, by Andreas Vlachidis, Isobel MacDonald, Foteini Valeonti, Julianne Nyhan and Kim Sloan, introduces the Collection Data Atlas as a methodological framework for navigating complex and dispersed cultural heritage data environments. Drawing on the Sloane Lab project, the article demonstrates how systematic data mapping enables multidisciplinary collaboration and supports data-driven research on the history and evolution of collections.

Subsequently, the article by Elsa Pereira, examines how the Semantic Web and Linked Open Data technologies can bridge methodological differences between archival practice and genetic criticism in the context of contemporary author archives. At the same time, she addresses the various challenges that currently limit the integration of genetic-oriented digital scholarly archives within the Web of Data.

The fourth article by Kostas Vlassopoulos and Kyriaki Konstantinidou, presents *SLaVEgents*, a large-scale DH project that combines big data, digital epigraphy, and history from below to investigate the agency of enslaved and freed persons in antiquity. It reveals the multiple identities of enslaved persons, the networks that they created and the different socio-cultural changes that occurred.

In the last contribution of the present issue, Armin Hoenen proposes a new computational approach to the study of literature before the nineteenth century, which combines image analysis, object recognition, and text mining; it offers proof-of-concept experiments

2 See conference website and programme at `https://www.unive.it/data/33113/2/103387`. The conference in itself was a follow up of the *Classical Texts in Digital Media*, which took place at the University of Patras (Greece) on 1-3 Sept. 2023. A more detailed report on the conference will follow as we collect and publish more conference papers in the forthcoming volume.

aimed at understanding the intellectual dynamics that led to the emergence of the stemmatic method.

As usual, our heartfelt gratitude goes to the many experts and scholars involved in the peer review process and to our advisory board members, the published authors, the members of our editorial board, as well as to our excellent publisher's team.

e-ISSN 2724-3923

magazén
Vol. 6 — Num. 2 — December 2025

Lectures that Link
Analyzing European Lecture Series as Nodes of Interaction in the Digital Humanities

Ulrike Henny-Krahmer
University of Rostock, Germany

Fernanda Alvares Freire
Technical University of Darmstadt; Berlin-Brandenburg Academy of Sciences and Humanities, Germany

Erik Renz
University of Rostock, Germany

Abstract This article aims to investigate the role of lecture series in Digital Humanities as a field of research within the European context over the past decade. Lecture series, widely used in higher education to facilitate scholarly exchange and to engage students, scholars, and broader audiences, have increasingly been adopted in DH since the late 2010s. By collecting and analyzing data from DH lecture series conducted across European institutions, we explore how they serve to connect institutions, researchers, disciplines, and research topics, employing quantitative data analysis.

Keywords Lecture series. Twenty-first century. DH community. Data visualization. Europe.

Edizioni Ca'Foscari

Peer review

Submitted 2025-06-02
Accepted 2025-06-10
Published 2025-09-09

Open access

Citation Henny-Krahmer, U.; Alvares Freire, F.; Renz, E. (2025). "Lectures that Link". *magazén*, 6(2), [1-22], 127-148.

DOI 10.30687/mag/2724-3923/2025/02/001

1 Introduction

For the Digital Humanities (DH) as an interdisciplinary field of research, scholarly exchange across disciplinary and institutional boundaries is essential. One type of event that is widespread in DH and that fosters scholarly exchange are lecture series, and these are the focus of this article.

A survey of the pertinent literature indicates that a 'lecture series' is generally characterized by a coherent thematic framework–ranging from a concrete topic to a broad question-delivered by a succession of speakers, who may come from different institutions and have diverse backgrounds. Each session addresses a sub-topic of the main theme (Linow, Führ, Kleihauer 2018, 177-9) and each speaker provides a specific point of view, thereby ensuring a plurality of perspectives on the overarching topic (Eberhardt 2010, 273). Moreover, Eberhardt emphasizes that lecture series are rarely confined to specialist audiences; they usually also address an interested public (273), and are therefore designed to be broadly accessible, requiring no prior disciplinary knowledge, and often function as an introductory platform for students across all fields as well as for non-university participants (276-7). Structurally, a lecture series most often spans one or more academic terms, with multiple sessions held at regular intervals (often weekly or monthly).

In research on higher-education didactics, lecture series have been conceptualized as a forum for discussion, an opportunity for self-improvement, or a format that supports the university's Third Mission (Eberhardt 2010; Dubs 2019; Nachtwei, Gierke 2023).[1] Overall, academic lectures have long been a cornerstone of scholarly life. As French and Kennedy observe, lectures remain "a valuable teaching method for both practical and pedagogical reasons" (2017, 640). Rooted in centuries-old university traditions, lectures bring learners together in a shared intellectual experience; as Palmer famously put it, "good teaching is always and essentially communal" (118). At their best, lectures excite curiosity and build community: students recognize themselves as part of 'something bigger' as they engage with a lecturer's expertise.

However, the majority of studies on lecture series debate the traditional lecture format, e.g. from a 'What value does it still have?' standpoint, criticizing it as a one-way transmission of information, prompting calls to make it more interactive, or shifting their focus to

1 The so-called 'Third Mission' refers to the social and economic mandate of universities, which lies beyond teaching and research. It includes the exchange of knowledge and technology between the university and society, further education and lifelong learning, as well as the university's engagement with the community.

more recent teaching practices such as e-lectures or hybrid lectures (Folley 2010; Dubs 2019, 18-37; Nørgård, Schreibman, Huang 2022). French and Kennedy report that for this reason in many institutions "the lecture has already evolved beyond the traditional idea of a unidirectional monologue" (2017, 640), incorporating active learning elements. In this light, lectures are understood not just as monologues, but as social, engaging events that adapt over time. Overall, aside from the occasional mention in higher-education research, lecture series have attracted surprisingly little attention regarding the wealth of information they offer. Explicit empirical investigations into lecture series are absent, in general and therefore also in the DH.

The traditional modes of public scholarship–including lectures and lecture series–may carry special significance for DH, which have often struggled with public perception and identity. As Nyhan (2016) has noted, despite numerous academic debates about 'What is DH?', the field is still frequently misrepresented outside its community. In this context, lectures and especially lecture series offer a format for showcasing DH work, for outlining its scope, for highlighting key areas, and for establishing and defining DH research communities inside of individual institutions and beyond. DH lecture series–typically cross-disciplinary but on a unifying theme–can serve as public forums where DH research is presented to students, scholars, and the public. Although such DH-themed series have proliferated across universities, they have not yet been systematically studied. In practice, these series usually bring together speakers from multiple institutions around topics such as, for instance, cultural heritage, computational analysis, media studies, or theoretical debates, so in principle they could reveal both the thematic contours of DH and the networks of scholars who participate.

Our hypothesis is that analysing the structure and content of DH lecture series can provide information about the state and development of this field of research, in a similar way to how contributions to DH conferences or DH journals have already been analyzed in national or international contexts to gain insights into the structure, networks, topics, and developments of DH.[2] Abstracts from conferences and journal articles are readily available as sources for such analyses, both in the form of the texts themselves, in addition to the corresponding metadata and bibliographies, depending on how they are published by conference organizers and journal editors. The fact that such sources are often made available in DH in open access and standard formats

2 See Weingart, Eichmann-Kalwara 2017 for an analysis of ADHO conference abstracts and Henny-Krahmer, Sahle 2018, Cremer et al. 2024, and Guhr 2025 for analyses about contributions to the DH conference of the German-speaking area. Additionally, see Kirtania 2021, and Spinaci, Colavizza, Peroni 2022 for bibliometric analyses of DH journal publications.

could explain why they have already been used frequently to track trends in DH. However, such structured and comprehensive data sources were not previously available for DH lecture series. In this article, we present a collection of data on European DH lecture series that we have newly compiled to fill this gap. Due to the abundance of existing DH lecture series, we have limited ourselves to the European context and to a period of a good ten years, between 2014 and 2025. Although there were already DH lecture series before 2014, our observation is that they became more numerous from the mid-2010s onwards. The restriction to a period of just over 10 years also enables us to cover as many series as possible within this scope. Data on events such as lecture series is often not permanently available–so one of the aims of our data collection is to secure it, so that it can be analysed in a transparent and reproducible way and not only by us but by all researchers interested in these events.

We assume that the analysis of lecture series in comparison to conference abstracts and research articles can provide new insights into DH structures and topics, since lecture series have a different character than conferences and publication venues. Conferences focus on a short period of time. This is where those academics who already identify with DH as a subject come together to present their latest work. In the case of publication organs such as journals, there is usually no direct contact between those involved, except via email. Lecture series, as conferences, are a social event with direct contact between organizers, speakers, and the audience, whether in person or online. Unlike conferences, however, these take place over longer periods of time, at many different, individual locations and institutions, and, as stated above, they can also have a connection to university teaching or be offered to the general public. We therefore assume that an analysis of the lecture series will produce different results than an examination of the other sources and that lecture series are another important building block in the sociological structure of DH that is worth researching.

Each DH lecture series has its own topics, objectives, and forms of implementation, something that also became apparent when collecting the data. Some series are part of a teaching program with students, others serve exclusively for the exchange between scholars, and still others integrate these elements and also include a general public. The topics and objectives may also differ from series to series. From this perspective, qualitative studies on DH lecture series will also be beneficial. In our contribution, however, we focus on the unifying, structural elements of all series and strive for a quantitative analysis. Against that background, we formulate the following central research question, which we pose in this article when examining the collection of data on European lecture series: To what extent and how do DH lecture series contribute to the

networking of researchers, institutions, disciplines, and topics in the field? Are they isolated events at individual institutions or is there a high degree of interconnection? As the title of our article suggests, we believe that 'lectures link', the question is, to what degree and how exactly they do. Local embedding of lecture series, for instance, is not an obstacle to networking, on the contrary. It can connect local actors with other regional, national, and international participants and thus expand the entire DH network in a different way than, for example, specialist conferences do.

In the remainder of this article, we will examine the questions posed as follows: in section 2, we present the data collection of European DH lecture series that we created, explaining which sources we collected, how the data was modeled, and what the current state of the collection is. In section 3, this data is analyzed with quantitative methods to investigate the links between researchers, institutions, disciplines, and topics that can be found in the data collection. The final section 4 serves to discuss our findings, keeping in mind the limitations of our analysis and pointing out future possible directions for research on DH lecture series.

2 Data Collection

We collect our data on European lecture series in a public GitHub repository (Henny-Krahmer, Alvares Freire, Renz 2025). In the following, we explain which DH lecture series we selected for data collection, and how we captured and modeled the data to create the database for our analyses.

2.1 Collection Criteria and Considerations

Across European institutions, the same concept of a lecture series is referred to by different terms in different languages. In English-speaking academia, labels such as 'Lecture Series', 'Guest Lecture Series', 'Seminar Series', 'Webinar Series', and more flexible formulations like 'Seminars in …', 'Talks in …', or 'Lectures in …' are prevalent. In German, one encounters *Ringvorlesung* (literally 'ring lecture') or *Vorlesungsreihe*, as well as *Kolloquium*. In French, the equivalent is *cycle de conférences* (conference cycle), and in Italian, *ciclo di conferenze* (cycle of conferences). Spanish similarly uses *ciclo de conferencias* (sometimes *serie de conferencias*). In Scandinavia, the terms are aptly descriptive: Swedish *föreläsningsserie* and Norwegian *foredragsrekke* (both meaning series of lectures). Despite this linguistic variety, the structure and function of these series are broadly similar. Therefore, when conducting a multilingual analysis,

we treat all these terms as equivalent, as they are identified as the same type of academic event in our study.

One advantage for our collection and analysis purposes is that, in accordance with the lecture series' aim to be a forum of public discussion, information about their programs is usually publicly accessible, i.e. published on the websites of the institutions that host them. Even if these are not designed for long-term availability, they can usually still be found beyond the period of the events themselves, so that we can collect this data.

An important, practical question of definition is when several lectures become a series. For our data collection, it was crucial that there is a thematic connection, expressed for example by an overarching title of a series, that there are recognizable organizers who are not the sole speakers and that at least three lectures have taken place that are related in this sense. These events should also not have taken place over a short period of time, i.e. at least over several months and usually no more than one event per week. This is how we draw the line between lecture series on the one hand and conferences or other, shorter events with several lectures, such as individual workshops or thematic weeks, on the other hand.

Once we have established what we consider a 'lecture series', we still have to decide which ones are DH lecture series. Our criterion for this was that the keyword 'Digital Humanities' or closely related terms such as 'Cultural Heritage' or 'Digital Heritage' must appear either in the title or in the general description of the series, i.e. that there must be an explicit reference to the subject. For this, we have considered all linguistic variants, not only the English terms (e.g., *humanités numériques* in French or *digital humaniora* in Swedish). In this respect, the decisive factor for us is whether a series explicitly declares itself to be a DH series or a series that is situated under the 'umbrella' of DH and thus aims to contribute to the DH as a research field and community. In our collection and analysis, we concentrate on general DH series, i.e., those that address the field as a whole. Beyond that, there are more specialized series on specific subfields of DH, for instance, on digital history.[3] We assess these, as well, as long as they refer to DH as a field, but it has to be taken into account that

3 Examples for such series are the *Offenes Forschungskolloquium Digital History* organized at the Humboldt University in Berlin (see `https://dhistory.hypotheses.org/digital-history-forschungskolloquium`) or the series *Voices Unbound: Lecture Series on Digital Oral History,* co-organized by the Technical University of Darmstadt, the University College London, the Luxembourg Centre for Contemporary and Digital History (C²DH) and the Max-Planck-Institut für Wissenschaftsgeschichte (see `https://hdsm.hypotheses.org/3657`). In both cases, the series focus on digital history as a specific subfield of DH but 'digital humanities' is mentioned in the general description of the series as a point of reference.

the scope of such series is narrower than of the ones addressing DH in general, which has effects on the people and institutions involved and the topics addressed. To sum up, we include lecture series that, by their very self-conception, aim to contribute to DH.

At the outset of our study, we conducted a systematic survey of lecture formats relevant to our data collection. Guided by the structural characteristics outlined above, we focused on events held in European countries. We realize that there are DH lecture series all around the world and we quickly saw that we would not be able to capture them all in a manageable time, so the limitation to the European context is meant as a first step and a starting point, beyond which we ourselves or other researchers can expand later.

In our collection, we focused on series in those languages that we have at least a reading comprehension of. We are aware that our language skills are influenced by our origins, educational contexts, and personal backgrounds and that this limits the data collection in a specific way. So far, we have considered lecture series in English, German, French, Italian, Spanish, and Swedish. In each of these languages, we identified instances of lecture series that conformed to the defining features as outlined in the previous section, through free keyword searches on the web and considering announcements in DH mailing lists and blogs.

Our assessment of DH lecture series in the scope defined above revealed a sample of 60 series from 14 countries.[4] Germany is the most frequently represented country (with 33 series), whereas some other countries appear only once. In Germany, DH is well represented as a discipline and is already comparatively well established, even though it is still considered a 'minor subject' in terms of institutionalization. Temporarily, the series we found in our assessment have a notable concentration in the 2020s: 37 of the 60 series started in or after 2020. It is quite conceivable that the COVID-19 pandemic plays a role in this context, as many activities have moved online since then, including many DH lecture series, which are in part either hybrid or purely online.

Of the 60 series, we have captured 30 for analysis so far (Germany: 14, France: 3, Austria: 2, Italy: 2, Switzerland: 2, Portugal and Sweden: 1, Spain: 1, Sweden: 1, UK: 1, UK and Ireland: 1, Belgium: 1). As mentioned in the introduction, we have only recorded lectures that took place between 2014 and the end of March 2025, thus covering approximately the last ten years. There are also earlier lectures, existing series are currently continuing, and new series started in 2025. In this respect, our database can be expanded in the future.

4 See the table in the appendix to this article listing the lecture series that are part of our data basis.

e-ISSN 2724-3923

2.2 Data Capture, Data Enrichment, and Data Modelling

Our data model takes into account the following basic information for recording individual lectures, which can be found on almost all of the websites related to the events: the date of a lecture, its title, the name of the speaker or speakers and their institutional affiliation, and the name of the institution where the lecture took place. If available, we also include the following details, even though they are often missing or only partially or inconsistently provided: an abstract summarizing the content of a lecture, the academic degree or title of the speaker, and whether the lecture was held online, in person, or in a hybrid format. Although lecture series formats can generally be considered homogeneous, the information provided in the presentations on their websites often differs significantly from one another.

In addition to the information about individual lectures, we capture details about lecture series terms and the whole overarching series: the name of the whole series, a description of its goals, its theme, and topics, if available, and the time frame and organizers of the series terms. Wherever possible, we document our sources by providing links archived with the WaybackMachine to make sure that our data sources are secured and transparent. We archive websites of whole series descriptions, of the programs of individual terms of a series, and of individual lectures and their details, if available. Furthermore, if there are links to related blog posts or videos of the lectures and these are easily recognizable, we collect these, as well, even though we do not archive videos for practical reasons.

A typical example of how information about a DH lecture series term is presented on a corresponding website is shown below **[fig. 1]**. In this example from the series *DH-seminariet*, there is a box listing the lectures of a specific term, indicating the date of each lecture and their speakers. The names of the speakers are linked to individual subpages where more information on each lecture can be found.

e-ISSN 2724-3923

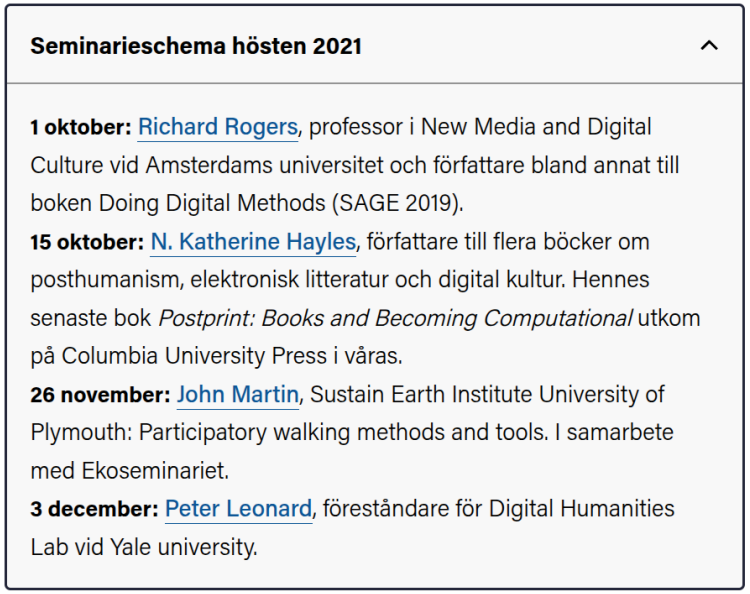


Figure 1 Information about a term of the Swedish lecture series *DH-seminariet*[5]

However, we do not only collect information about the lecture series that we find on the corresponding websites, but also supplement this with information that we can derive from the directly visible data, with the aim of having an expanded basis for analyzing the lecture series. We add the following information:

- gender of the speakers: we interpret this from the names of the speakers and from further information that may be provided; possible values are 'female', 'male', and 'non-binary';
- authority data about the speakers: we add ORCID or Wikidata IDs for those speakers whose ID we can easily find;
- places of organizations: we connect the mentioned organizations (host institutions and speaker affiliations) with information about their locations (cities, which in turn are referred to the respective countries and continents);
- authority data of organizations: we add Wikidata IDs to organizations where we can find them;

5 For figure 1, see the website of the lecture series at: `https://web.archive.org/web/20250526140442/https://www.gu.se/digital-humaniora/aktuellt/dh-seminariet#accordion=0d914e3d-55d8-4923-a9cb-1c7f6393f7fa`.

- authority data of places: we add IDs of the Getty Thesaurus of Geographic Names (TGN);
- disciplines: we add information about the various disciplines to which the talks are connected for every lecture;
- topics: we add lists of topics that the talks address to every lecture.

In general, some of the above information is encoded at the individual lecture as a specific instance at a certain point in time, for example, the affiliation and titles of the speakers, which may change. In contrast, we capture the names and gender of people and the places of institutions in a general index, assuming relative (even if not absolute) stability of these values.

To add the information about disciplines covered by a lecture, we created a list ourselves that includes 'digital humanities' as well as disciplines that are part of or related to DH as a broader research field, for example, 'library and information science', 'computer science', 'literary studies', 'linguistics', 'history', 'archaeology', and so on. When assigning the disciplines to the lectures, we only list 'digital humanities' if this is explicitly mentioned in the title or abstract of the talk or if no other more specific discipline is appropriate. Furthermore, several disciplines can be assigned to each lecture.

When it comes to assigning topics, we generated them using the LLM Mistral AI via a custom Python script. Every talk was enriched with five topics generated by the LLM. All the topics are provided in English.[6] The LLM topics are not from a controlled list. However, they tend to capture the content of the lectures quite well and even provide contextual information from the language model. Assigning topics to the lectures is easier when abstracts are present and not only the title is available. We decided to use an LLM topic assignment because this could be done automatically and effectively. In a future version of our data set, we aim to complement this with a manual assignment using the TaDiRAH taxonomy.[7]

All the resulting data that we collected is encoded following the Guidelines of the Text Encoding Initiative (TEI). The collection methodology and details of the encoding are described in an ODD file, of which a formal data schema can be derived to control the data. For an example of how an encoded lecture looks like, see Code 1.

6 The prompt for generating the LLM topics is documented in the corresponding script in the GitHub repository.

7 For a description of the TaDiRAH taxonomy, see `https://web.archive.org/web/20250624200607/https://de.dariah.eu/en/tadirah`.

```
<event xml:id="ls33_t4_l1" type="lecture" when="2021-10-01">
  <eventName xml:lang="en">Visual media analysis for Instagram and other online
    platforms</eventName>
  <ptr type="programme"
    target="https://web.archive.org/web/20250526155710/https://www.gu.se/evenemang/dh-seminariet-richard-rogers"/>
  <ptr type="video"
    target="https://play.gu.se/media/DH-seminariet%3A%20Richard%20Rogers%2020211001/0_4nl9nqre"/>
  <note type="abstract" xml:lang="en">
    <p>Instagram is currently the social media platform most associated with online images (and
      their analysis), but images from other platforms also can be collected and grouped, arrayed by
      similarity, stacked, matched, stained, labelled, depicted as network, placed side by side and
      otherwise analytically displayed. In the following, the initial focus is on Instagram,
      together with certain schools of thought such as Instagramism and Instagrammatics for its
      aesthetic and visual cultural study. Building on those two approaches, it subsequently focuses
      on other web and social media platforms, such as Google Image Search, Twitter, Facebook and
      4chan. It provides demonstrations of how querying techniques create online image collections,
      and how these sets are analytically grouped through arrangements collectively referred to as
      metapictures.</p>
  </note>
  <note type="keywords">
    <term type="discipline" corresp="#media-studies"/>
    <term type="topics-llm">image analysis, social media, instagram, visual culture, web
      platforms</term>
  </note>
  <note type="realization">
    <term type="speech">online</term>
    <term type="audience">online</term>
  </note>
  <listPerson>
    <person role="speaker" corresp="#rogers_richard">
      <name>
        <roleName type="title">Prof. Dr.</roleName>
      </name>
      <affiliation corresp="#uni-amsterdam"/>
    </person>
  </listPerson>
  <org role="host-institution" corresp="#uni-gothenburg"/>
</event>
```

Code 1 Example of an individual lecture encoded in XML TEI

3 Analysis of Lecture Series

In our analysis, we aim to examine the lecture series as a means of scholarly communication with the aim of promoting cross-disciplinary interactions and establishing links within the field of DH, be it inter-institutional, regional, national, or international networks. To this end, we use the data about DH lecture series that we collected to measure the degree to which researchers (speakers and organizers), institutions, disciplines, and topics are linked through their participation and occurrence in the lectures. In addition, we have collected information about places and times: where and when the lectures took place, where speakers came from and when they intervened. We can now use all of this data to examine DH lecture series in Europe over the course of the last decade. We do this by quantitatively evaluating the data in the form of statistical visualizations.

When analyzing the various lecture series we have recorded, we must take into account that they are not structurally homogeneous phenomena. As can be seen in Figure 2, the scope of the series varies, which naturally also influences their connections. The number of terms per series ranges between only a single term up to 12 terms, which means that some series are only organized once, whereas others have a longer tradition and are regularly held. The median in

our data set is a duration of 2.5 terms. The number of lectures per term also varies and lies between only a single lecture in one term up to 44 lectures, the latter being an exceptional case where no real term-structure could be recognized. The medium number of lectures per term is 6, showing that term lengths and frequencies of lectures held in a term vary in the different lecture series.

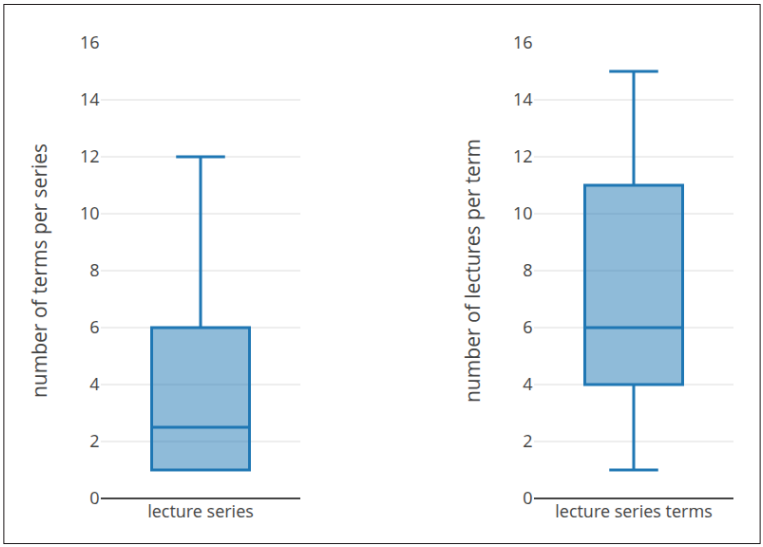


Figure 2a-b Structure of the lecture series by: (a) number of terms per series; (b) number of lectures per term

The 30 lecture series for which we have captured the data so far include 863 individual lectures. 85 people were involved as organizers and 820 as speakers. There are 30 different host institutions and 305 institutions involved as affiliations of speakers. Regarding places, the lectures took place in 29 cities in 11 countries and speakers from 207 cities in 36 countries were involved in the events. These figures give an impression of the scope of our data set, but also provide initial quantitative insight into the diversity of the individuals, institutions, and locations involved. We can directly see, for instance, that some speakers are involved several times, as the overall number of speakers is lower than the total number of lectures given, even though there are some lectures held by groups of speakers. Figure 3 shows that most speakers only held an individual lecture. In our dataset, 134 speakers held two or more lectures in all the series. The speaker who was involved most held nine lectures. If we take a closer look at this individual case, the following picture emerges: all the lectures were held between 2020 and 2024, at six different lecture series in Germany (in the cities Berlin, Potsdam, Rostock, Stuttgart, and

Erlangen/Nuremberg). During this time, the speaker was affiliated with four different German universities (Hamburg, Darmstadt, Regensburg, Stuttgart). The most active speaker in our current data set is therefore very well networked within Germany. However, there are also speakers that establish connections internationally. One speaker in our dataset who is affiliated with the University of Victoria in Canada was involved in four lectures, one in Antwerp (2017), Cologne (2018), London (2020), and Venice (2021), respectively, of which the lecture in London was planned but canceled. In summary, it can be stressed that relatively few speakers are active in several lecture series and that each case can be considered individually, as well, to understand the kind of linking between institutions and places that took place through the speakers.

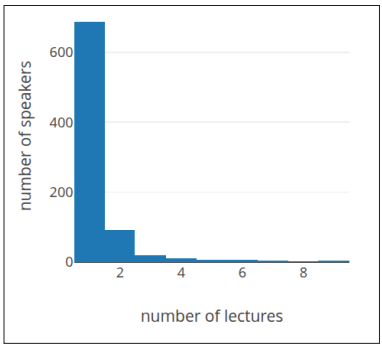
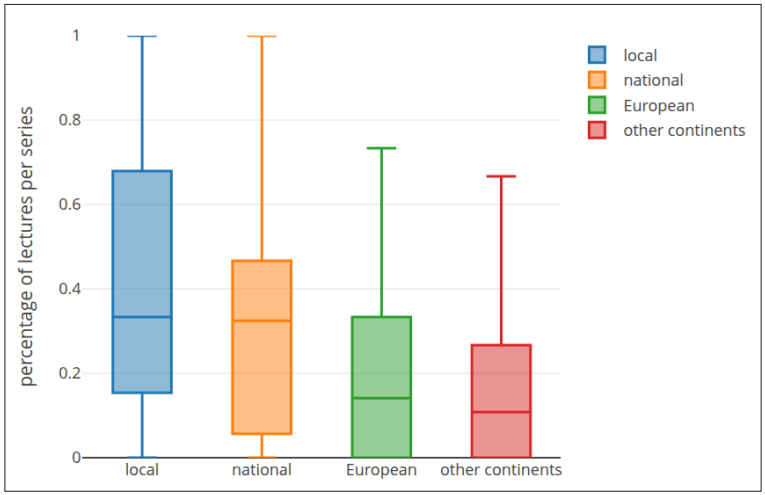
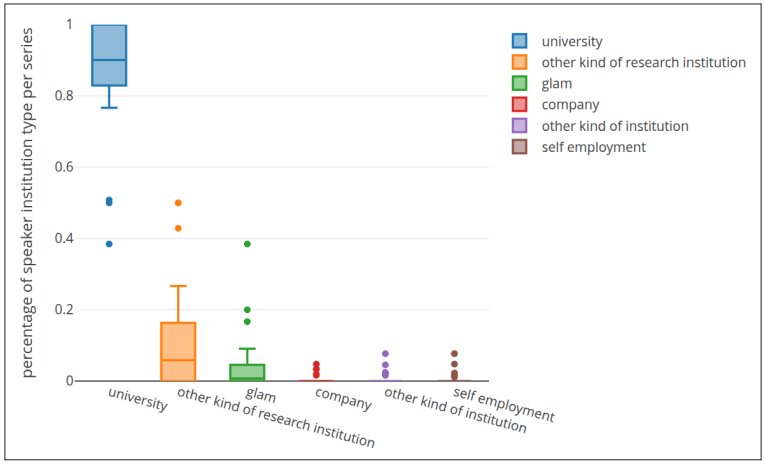


Figure 3
Number of lectures per speaker

Many speakers are invited from places other than the host institution, as the number of speaker cities (207) is almost seven times higher than the number of host cities (29). Speakers from 36 countries are involved in the lecture series organized in 11 different countries, which shows that international speakers are invited from a broader context. But how far does the reach of the lecture series extend in terms of location if we analyse the whole distribution of data? Figure 4a shows a series of box plots that visualize the percentages of lectures held by local, national, and European guests and by speakers invited from other continents. The shares are calculated per lecture series. We see that the lecture series primarily serve to connect various speakers from the same city and from other cities in the same country, with a median of 33% and 32% of invited speakers from these geographic contexts, respectively. Speakers from other European countries participate with a median of 14% throughout all the lecture series, and speakers from other continents with a median of 11%. The box plots show that there is some variance, especially in the proportion of local and national speakers. For instance, there are purely local series and series with no local speaker at all. Overall,

the degree of internationalization is relatively low throughout the lecture series, so that we can speak of 'lectures that link locally and nationally' in the first place.





Figures 4a-b Percentages of lectures held by: (a) local, national, European, and non-European speakers; (b) speakers from different types of institutions (both shown per lecture series)

Another type of information we captured in our dataset is the type of institution a speaker belongs to: university, other kind of research institution, GLAM, company, other kind of institution, or self-employment. Figure 4b shows a corresponding analysis of the percentages of speaker institution types per lecture series. With a

median of 90%, speakers are affiliated with universities; speakers from other kinds of research institutions are present with a median of only 6%, speakers from GLAM institutions with a median of less than 1%, and speakers from companies, other kinds of institutions (for instance, newspapers or foundations), and self-employed speakers are the exception. This shows that DH lecture series mainly take place within the academic field, at least as far as the invited speakers are concerned. In the sense of public humanities or a link to the non-academic professional environment, a further opening could take place here by inviting speakers from other areas more often.

Regarding disciplines, we analyze how disciplines are represented in the whole dataset of the DH lecture series and how strongly individual lecture series are influenced by specific disciplines. This serves to analyze to what extent the lecture series connects researchers and listeners in a highly interdisciplinary manner or only within specific research areas within DH. We recorded 36 different disciplines to which the lectures in the series contribute **[fig. 5]**. Most often, lectures can be associated with DH as a general field, followed by history, literary studies, computer science, and library and information science. Only individual lectures address specific fields such as Yiddish, African, Slavic, or Chinese studies, or connections between DH and mathematics or the natural sciences, for instance.

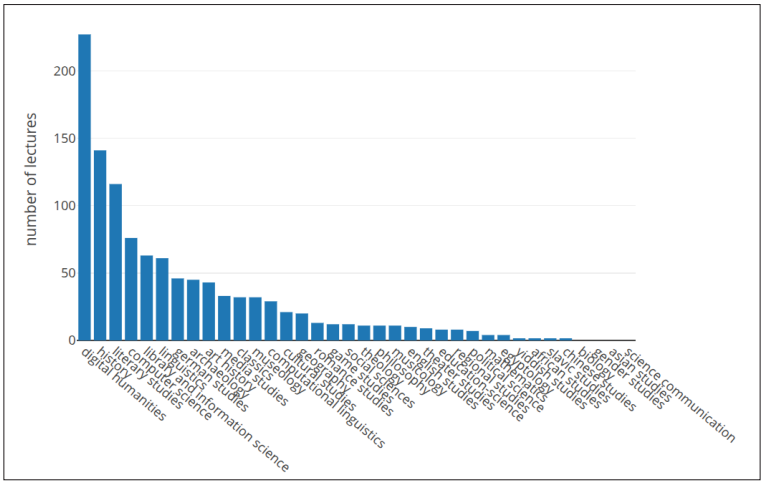
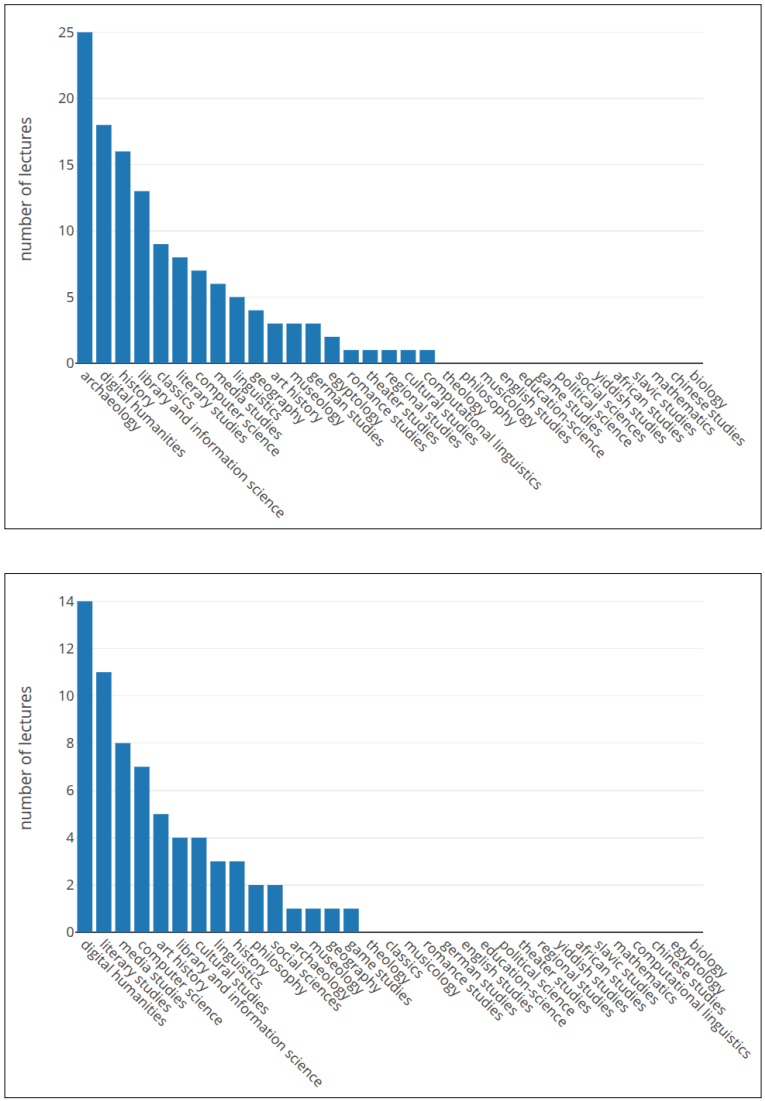


Figure 5 Disciplines of the lectures held in DH lecture series (overview)



Figures 6a-b Disciplines of the lectures held: (a) in the series *Colloquium in Digital Cultural Heritage* (2018-25); (b) in the series *DH-seminariet* (2020-24)

Each lecture series has its own profile of disciplines that it covers. The series *Colloquium in Digital Cultural Heritage*, for instance, which takes place at the University of Cologne since 2018, has an inclination towards the field of archaeology **[fig. 6a]**, whereas the series *DH-seminariet*, which ran between 2020 and 2024 at the

University of Gothenburg, favors DH in general and literary studies **[fig. 6b]**. Overall, most series have specific disciplinary focuses, but also address general DH topics and other fields to a certain degree.

Regarding the topic keywords that we collected, they cannot be directly used as indicators for the degree to which a lecture series enables networks between researchers, but they may give insight into the degree of interdisciplinarity and thematic diversity of the DH lecture series. Since we had the thematic keywords generated by an LLM, the results are a non-standardized list of keywords. This must be taken into account in the analysis. In a first, non-exhaustive approach, we compare four thematic areas of DH: digitization and cultural heritage, digital edition, analysis, and artificial intelligence **[fig. 7]**. We analyze the proportions of lectures in each series that were marked with these topic keywords.[8] The results show that most lecture series are thematically diverse, which one would expect. The median proportion of lectures in a series concerned with digitization and cultural heritage is 4%, with digital edition 1%, with analysis 21%, and with artificial intelligence 16%, with some variance and outliers visible in the box plots. The fact that the fourth area is relatively strongly represented is quite striking. One hypothesis to be examined in more detail is that this topic area is currently being dealt with heavily in DH lecture series.

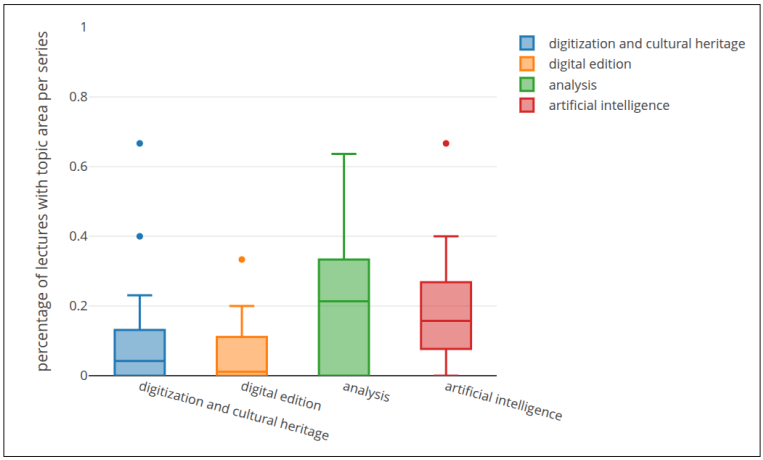
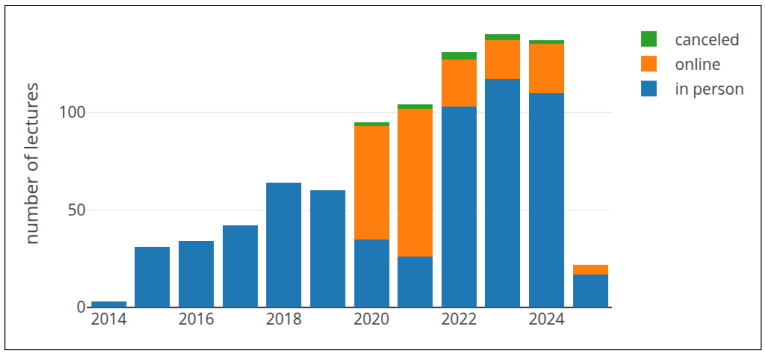


Figure 7 Percentages of topic areas per lecture series

8 For this purpose, the following terms were searched for inside of the keywords: 'cultural heritage' and 'digitization'; 'edition', 'editing', 'editorial'; 'analysis'; 'machine learning', 'ai', 'artificial intelligence', 'llm'.

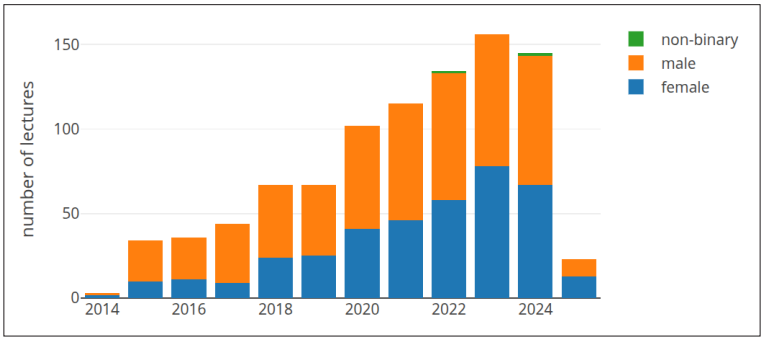
All the analysis we have done so far could also be considered over time, but there is no room for that in this article. We will therefore limit the analysis over time to two general aspects: how many lectures took place over time at all **[fig. 8a]**, including the indication of the mode that the lectures were delivered in (in person, online, or canceled), and how many lectures were given by speakers of which gender **[fig. 8b]**. We clearly see that the number of lectures increases with the beginning of the COVID-19 pandemic in the year 2020, and online lectures begin to happen in the same year. After 2021, the share of online lectures decreases again but about one-fifth the lectures continue to be delivered online. The pandemic may have fostered DH lecture series because online formats became common as a result.[9] Obviously, the kind of 'linking' and 'networking' that takes place in the context of a lecture is influenced by the presentation mode. Overall, we see an increase in the number of lectures, reflecting the growth and further establishment of DH as a research field and an increase in public lecture series activities.

Regarding the gender distribution of speakers, has there been a trend toward greater balance over time? Figure 8b shows that this is the case, but only moderately. The number of lectures per year with female and male speakers was only the same in 2023 – in other years, there were more male speakers, and for 2025, we do not have any reliable results yet. Non-binary speakers, as far as we could assess, are a minority.[10]



9 The question of how the presentations were delivered can also be examined from the perspective of the audience, for whom we recorded the values 'in person', 'online' or 'hybrid'. Even if speakers presented on site, there may have been an additional online broadcast. However, for reasons of space, we will not pursue this further here.

10 A possible extension of the analysis of gender distributions would be to analyze the genders of the lecture series organizers.



Figures 8a-b Number of lectures per year categorized by: (a) mode of delivery (in person, online, canceled); (b) gender of the speakers (female, male, non-binary)

4 Conclusions

The structural, quantitative evaluations of the series have shown, on the one hand, that each series has its own character in terms of its structural scope and composition. Speakers from different places and institutions are invited, referring to different disciplines and topics within the 'big tent' DH. On the other hand, there are some general characteristics that can be observed. The extent to which local, national, or international guests are invited shows that the lecture series tend to primarily connect local and national actors in the field. Only some speakers are repeatedly present in different lecture series, and most speakers are affiliated with universities, not other types of institutions. Through our insights into how these series are currently organized and realized, we can learn how they could be implemented differently in a targeted manner, for instance, to be more international and more open beyond academia. By organizing lecture series, we can actively shape DH networks.

With this study, we presented a large dataset on European DH lecture series, showing how such data can be collected and modeled as a basis for quantitative analyses of lecture series as events that bring together researchers, students, and the public in an interdisciplinary setting, enabling community-building and networking in and beyond individual institutions. The dataset still needs to be completed, though, to include all the data that we could identify as relevant for our study, and it can also be expanded to cover lecture series from further countries and in more languages. Our analysis is only an initial exploration of how a dataset on DH lecture series can be researched and what insights it can bring. Further types of analyses are possible: Future steps could, for instance, focus more on developments over time, which we have only done marginally

in this article. All analyses could also be refined by looking at the situation for individual countries, locations, institutions, or lecture series. Many perspectives and focuses between distant and close examination of the data are possible. Further options for displaying the relationship between places, institutions, and people involved in the lecture series include the use of maps and networks. In addition, audience information could be collected: currently, we have no data on participant numbers or backgrounds, and it remains unclear whether audiences primarily consisted of students, fellow scholars, or laypersons. The data on DH lecture series that is usually publicly available does not provide this information. Going beyond what can be easily collected, quantitative analysis could be combined with qualitative investigations by focusing on individual lecture series more in-depth and by including other sources of data. For instance, interviews could be conducted with organizers, speakers, and participants. This would provide a clearer picture of how different stakeholders experience the lecture series and how they can be further developed as networking events for the community of DH researchers, students, and the public.

Appendix

Table 1 List of lecture series in the data collection

No.	Lecture Series	Countries	Venues	Years
1	*ACDH-CH Lectures*	Austria	Vienna	2015-
2	*The Digital Humanities Guest Lecture Series / Introduction to Digital Humanities*	Austria	Vienna	2019-
3	*Platform{DH} Lecture Series / platform{talks}*	Belgium	Antwerp	2014-22
4	*Cycle de conférences – Humanités numériques*	France	Rouen	2021-22
5	*Les humanités numériques au coeur du patrimoine culturel*	France	Corte	2021-22
6	*Rendez-vous du Centre des Humanités Numériques*	France	Paris	2022-23
7	*Einführung in die Digital Humanities*	Germany	Münster	2022
8	*Digital Humanities: Anwendungsbereiche, Möglichkeiten, Werkzeuge*	Germany	Tübingen	2024-
9	*Digital Humanities – Theorie und Methodik*	Germany	Leipzig	2014-
10	*Digital Humanities – Aktuelle Forschungsthemen*	Germany	Cologne	2015-
11	*Digital Humanities in den Geisteswissenschaften / Digitale Geisteswissenschaften*	Germany	Stuttgart	2015-20
12	*DH-Kolloquium an der BBAW*	Germany	Berlin	2017-
13	*Colloquium in Digital Cultural Heritage*	Germany	Cologne	2018-
14	*Phänomenologie der Digital Humanities*	Germany	Berlin	2021-
15	*Grundlagen und anwendungsorientierte Methoden der Digital Humanities*	Germany	Dresden	2022-
16	*Kulturwissenschaften und Digital Humanities. Konzeptionelle Annäherungen*	Germany	Vechta	2023
17	*Digital Humanities im Fokus: Methoden, Anwendungen und Perspektiven*	Germany	Rostock	2023-
18	*DHSS Vortragsreihe*	Germany	Erlangen / Nuremberg	2023-24
19	*Offenes Master-Kolloquium Digital Humanities*	Germany	Stuttgart	2024-
20	*Vortragsreihe „Code & Kultur“*	Germany	Potsdam	2024-
21	*Seminars in Digital and Public Humanities*	Italy	Venice	2019-
22	*Doing digital humanities @ DiSSGeA*	Italy	Padua	2020-22
23	*Knowledge Organisation and Digital Humanities: An International Webinar Series*	Portugal / Sweden	Porto / Växjö / Kalmar	2021
24	*Ciclo UC3M de Humanidades Digitales*	Spain	Madrid	2024
25	*DH-seminariet*	Sweden	Gothenburg	2020-24
26	*Einblicke in die Digital Humanities*	Switzerland	Bern	2020-
27	*Cultural Heritage in the Digital Age*	Switzerland	Basel	2023-
28	*The Susan Hockey Lecture in Digital Humanities*	United Kingdom	London	2015-19
29	*The Digital Humanities Lecture Series*	United Kingdom	London	2022
30	*Trust and Authority in the Digital Age*	United Kingdom / Ireland	Birmingham / Dublin	2021

Bibliography

Cremer, F.; Blessing, A.; Helling, P.; Henny-Krahmer, U.; Jung, K.; Reiter, N. (2024). "DHd Chronicles – Anreicherung und Analyse der Beiträge zu den Jahrestagungen der Digital Humanities im deutschsprachigen Raum 2014–2023". Weis, J.; Haider, T.; Bunout, E. (eds), *Book of Abstracts – DHd2024*. Passau: Zenodo, 154-160. https://doi.org/10.5281/zenodo.10698356.

Dubs, R. (2019). *Die Vorlesung der Zukunft. Theorie und Praxis der interaktiven Vorlesung*. Stuttgart: utb. https://doi.org/10.36198/9783838552699.

Eberhardt, U. (2010). "Ringvorlesungen als Diskussionsforum und Instrument hochschuldidaktischer Weiterbildung". Eberhardt, U. (Hrsg.), *Neue Impulse in der Hochschuldidaktik*. Wiesbaden: VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-531-92319-2_12.

Folley, D. (2010). "The Lecture is Dead Long Live the e-Lecture". *Electronic Journal of e-Learning*, 8(2), 93-100. https://academic-publishing.org/index.php/ejel/article/view/1590.

French, S.; Kennedy, G. (2017). "Reassessing the Value of University Lectures". *Teaching in Higher Education*, 22(6), 639-54. http://dx.doi.org/10.1080/13562517.2016.1273213.

Guhr, S. (2025). "Trends in den Computational Literary Studies bei den DHd-Jahrestagungen". *DHd Blog*. https://dhd-blog.org/?p=22316.

Henny-Krahmer, U.; Sahle, P. (2018). "Einreichungen zur DHd 2018". *DHd Blog*. https://dhd-blog.org/?p=9001.

Henny-Krahmer, U.; Alvares Freire, F.; Renz, E. (2025). *Lectures that Link*. Version 0.5. https://doi.org/10.5281/zenodo.15769951.

Kirtania, D.K. (2021). "Bibliometric Analysis of Open Access Digital Humanities Publications". *Library Philosophy and Practice (e-journal)*. https://digitalcommons.unl.edu/libphilprac/6667.

Linow, S.; Führ, M.; Kleihauer, S. (2018). "Aktivierende Ringvorlesung mit begleitender Konzept-Werkstatt Herausforderung: Nachhaltige Entwicklung – Klimaschutz in und um Darmstadt". Leal Filho, W. (ed.), *Nachhaltigkeit in der Lehre. Theorie und Praxis der Nachhaltigkeit*. Berlin; Heidelberg: Springer Spektrum, 203-18. https://doi.org/10.1007/978-3-662-56386-1_11.

Nachtwei, J.; Gierke, L. (2023). "Im Sinne der Dritten Mission. Über die Ringvorlesung als Lehrformat". *Forschung & Lehre*, 30(7), 502-3.

Nørgård, R.T.; Schreibman, S.; Huang, M.P. (2022). "Digital Humanities and Hybrid Education: Higher Education in, with and for the Public". Schwan, A.; Thomson, T. (eds), *The Palgrave Handbook of Digital and Public Humanities*. Cham: Palgrave Macmillan, 11-29. https://doi.org/10.1007/978-3-031-11886-9_2.

Nyhan, J. (2016). "It is time to Address the Public Communication of DH". *Digital Humanities Quarterly*, 10(3). http://www.digitalhumanities.org/dhq/vol/10/3/000261/000261.html.

Palmer, P.J. (2017). *The Courage to Teach: Exploring the Inner Landscape of a Teacher's Life*. San Francisco: Wiley.

Spinaci, G.; Colavizza G.; Peroni, S. (2022). "A Map of Digital Humanities Research across Bibliographic Data Sources". *Digital Scholarship in the Humanities*, 37(4), 1254-68. https://doi.org/10.1093/llc/fqac016.

Weingart, S.B.; Eichmann-Kalwara, N. (2017). "What's Under the Big Tent?: A Study of ADHO Conference Abstracts". *About Digital Studies / Le champ numérique*, 7(1). http://doi.org/10.16995/dscn.284.

e-ISSN 2724-3923

magazén
Vol. 6 – Num. 2 – December 2025

A Data Atlas Method for Analysing and Visualising Dispersed Cultural Heritage Collections

Andreas Vlachidis
University College London (UCL), United Kingdom

Isobel MacDonald
The British Museum, United Kingdom

Foteini Valeonti
University College London (UCL), United Kingdom

Julianne Nyhan
TU-Darmstadt, Deutschland

Kim Sloan
The British Museum, United Kingdom

Abstract The history of collecting shows how people and institutions have tried to interpret and shape their worlds, but tracing objects across heritage collections is hindered by fragmented, uneven data. Issues of scope, digitisation, legacy metadata and dispersed systems limit data-driven research and digital reconnection efforts. This paper proposes the Collection Data Atlas, a framework for mapping complex informational landscapes across institutions and epochs. Using the AHRC-funded Sloane Lab, it applies the Sloane Data Atlas to visualise and synthesise the intricate data environment of Sir Hans Sloane's collection.

Keywords Collections as Data. Cultural Heritage. Data Atlas. Sloane. Digital Humanities.

Edizioni Ca'Foscari

Peer review

Submitted 2025-08-18
Accepted 2025-10-30
Published 2025-12-23

Open access

Citation Vlachidis, A.; MacDonald, I.; Valeonti, F.; Nyhan, J.; Sloan, K. (2025). "A Data Atlas Method for Analysing and Visualising Dispersed Cultural Heritage Collections". *magazén*, 6(2), [1-30], 149-178.

DOI 10.30687/mag/2724-3923/2025/02/002

1 Introduction

The proliferation of digital collection databases built by cultural heritage institutions over the past six decades, and the many digitisation projects that have been undertaken in the heritage sector in recent years, have acted to dramatically expand the empirical basis upon which research on the history of collections can be pursued (Institute of Museum and Library Services 2018; Jones 2022, 2). Such developments and digital advances, open the theoretical possibilities to pursue data-driven, long durée analysis of how processes of fluidity have shaped the heritage collections at our disposal today. Understanding the information environment of such collections, which encompasses both historic, handwritten, manuscript-based descriptions of acquired objects as well as present-day digital databases is imperative for examining how cultural heritage collections have grown over time. This can help us examining the development of collections beyond the lens of acquisitions made by individuals or institutions that "captures only one side of collections' history" (Cornish, Driver 2020, 327). Moreover, better understandings of these information environments have the potential to cast new light on the fluidity of collections, the role that circulation played in their formation and mobilisation, and to signify the histories of objects, their acquisition, location and movement over time.

The information that describes heritage objects and collections tends to be highly complex, often having been catalogued by many people, over a long period in a range of analogue and digital formats. It is often siloed across multiple institutions and different institutional recording systems, with varying levels of interoperability and accessibility. This can lead to a lack of understanding of what material is held within a given institution. At present, the ability to undertake such research is reliant on an understanding of the idiosyncrasies of collection databases and related material located across different institutions. Research is moving beyond the accumulation of collections, seeking a deeper examination of the dynamic movement and organisation of objects across information systems, and individual and institutional agents and actors (Driver et al. 2021, 8). Thus, our research aims to address the following questions: How can the review and mapping of dispersed datasets be undertaken in a systematic way, using a methodological framework that can capture shared understandings of complex data landscapes, that have been created by multidisciplinary teams, institutions, and systems with varying data structures? Additionally, how can the data landscape of such heritage collections, including their intersections and the interactions of their constituent parts be synthesised and visualised in a structured and transparent way?

The collection of Sir Hans Sloane is a paradigmatic example of potential that may be unlocked by answering the above questions. Assembled from the 1680s onwards, and in part financed by profits from the transatlantic trade in enslaved human beings, Sloane's vast collection of natural history, pharmaceutical specimens, books, manuscripts, prints, drawings, coins, and antiquities from across the world was made as Britain became a global trading and imperial power (MacGregor 1994, 53). Today Sloane's physical collections, and their historical collection records, are spread across three national institutions: the National History Museum (NHM), British Library (BL) and BM. They are, in turn, recorded in five different digital cataloguing systems across those institutions, each shaped by the individual disciplines and institutional histories that gave rise to them. The *Sloane Lab: Looking back to build future shared collections* (hereafter Sloane Lab) project, funded by the AHRC's *Towards a National Collection Discovery* programme, re-establishes the broken links between Sloane's catalogues and collections across the Natural History Museum (UK), the British Library, and the British Museum, devising automated and augmented methods that are relevant not only for Sloane's collection but also for cultural heritage collections in museums, galleries, libraries, and archives (Nyhan et al. 2025). The project positions Sloane's collection as a microcosm through which to investigate the technical, infrastructural, conceptual, historical and social challenges faced in bringing together digital cultural heritage collections.

A crucial precursor of researching Sloane's collection is the mapping of its informational landscape, both historical and present day, to move from a fractured, institution-specific view of the individual parts of the collection to a view of the collection as a whole, allowing its cross-institutional landscape to be understood in its entirety for the first time. We accordingly present the *Collection Data Atlas* as an instrument that can further this aim, while likewise functioning as a methodological instrument that can be used to identify, represent and ultimately support the mobilisation of complex and cross-institutional collections (see e.g., Institute of Museum and Library Services 2018; Padilla et al. 2019; Towards a National Collection).[1] The focus is not at the lower level of data characteristics and metadata that can be easily addressed by schemas and conceptual models of unified data representations. Instead, the Collection Data Atlas addresses softer silos, which relate to the scope, size, availability, coverage, legacy attributes, and manifestation of collection data, which often persist both within and between institutions. Such attributes cannot be adequately addressed by conceptual data models and metadata

1 See Towards a National Collection: `https://www.nationalcollection.org.uk/`.

mappings alone. Our approach offers a holistic metaphor to allow for a multidisciplinary understanding, inventorying analysis, and conception of collection landscapes.

This paper begins with a synthesis of the wider research context of this paper, including a brief discussion of research that uses data-driven methods to understand the history of collections. An overview of the way in which digital collections data has been used in various projects, alongside the move for transparency in the creation of datasets for analysis, situates the *Collection Data Atlas* as an intentional instrument designed for the early stages of large-scale data aggregation. The Sloane Lab Data Atlas then exemplifies the design rationale and application of the *Collection Data Atlas* in the data aggregation process, demonstrating how it aids data auditing, data ingestion prioritisation and decision-making within the complex data landscape of the Sloane collection. We conclude with a critical reflection on the usefulness of proposed methodological framework within the Sloane Lab and, crucially, its transferability to other collections as data projects. The reflection addresses technical and broader efforts to understand collection histories and presents how other projects can effectively adopt the proposed framework to chart dispersed collection landscapes within the cultural heritage domain and beyond. The Sloane Lab Data Atlas is available on GitHub[2] and a snapshot of it is appended **[Appendix]**.

2 Research Context

2.1 Collections as Data

In recent years, cultural heritage collections have become key sites of technological encounter. As Thomas Padilla has argued, "Collections as data entails thinking about ways to increase meaning making capacity by making collections more amenable to use across an expanded set of methods and tools, typically but not exclusively computational in nature" (Padilla 2017, 2). Efforts to support computationally driven research of cultural heritage collections can cultivate open documentation and content-sharing approaches, with useful examples including the HathiTrust Research Centre, the National Library of the Netherlands Data Services and APIs, the Library of Congress' Chronicling America, and the British Library (15). However, it is worth noting that although Collections as Data projects foreground the practice, theory, and ethics surrounding the use of digital collections data by cultural heritage institutions, many

2 Access the resource online: `https://github.com/sloanelab-org/data-atlas`.

smaller institutions have not built digital collections or designed access to them with the aim of supporting computational use (10).

Meanwhile, data-driven projects tend to use digital collections data for provenance research. The Provenance Lab at Leuphana University, Lüneburg, for example, pursues provenance research through a network-based approach, highlighting social and economic trends in collection formation and change with digital methods (Rother, Mariani, Koss 2023). The *Digital Benin* project, hosted at Markk Museum am Rothenbaum, Hamburg, has created an online knowledge forum, reconnecting object data from 131 institutions worldwide of over 5,000 objects that were looted in the British military campaign against Benin City in 1897 (Luther 2024). With a similar focus on colonial heritage held in museums, the *Pressing Matter: Ownership, Value and the Question of Colonial Heritage in Museums* hosted at Vrije Universiteit in Amsterdam, is creating a Knowledge Graph to aggregate entities across institutional silos (Shoilee 2022). The *CUDAN Open Lab*, based out of Tallin University, is using collection data from twelve European contemporary art museums to explore the acquisition of contemporary art works and associated museum profiles at an aggregate level (Solà et al. 2023). The *Between Canon and Coincidence*, hosted at Leiden University, is studying the provenance of Latin American collections in Europe through the collection and analysis of collection data from 12 museums in 9 countries using network analysis, big data computer models and AI (Leiden University 2023; Berger 2023). These examples highlight the growing use of digital collections data and data-driven methods to augment and challenge our understanding of object provenance and broader patterns of collection growth and dispersal.

In the UK a growing number of research projects use data-driven approaches to examine the history of institutional collections. These have ranged from an examination of relational patterns between the museum and communities associated with the collection of the Pitts Rivers Museum, Oxford (Gosden, Larson, Petch 2007; Petch 2006; 2002); an analysis of objects from a particular part of the world within one museum collection (Wingfield 2011); an analysis of specific areas of a national collection (Phillipson 2019); and the mapping of certain collector typologies (Penn, Cafferty, Carine 2019). A two-year research project at the British Museum conducted a comprehensive quantitative examination of the history of its current collection records (MacDonald 2023). Its aim was to trace patterns across time, material and departments and to provide an institutional context within which individual objects, collections and collectors could be understood. Whilst most of these projects focus on the digital collections data held by one cultural heritage institution, some have traced patterns running across institutions. For example, Charlotte Dixon analysed

data across 13 UK museums to compare the collecting practices of Indian Ocean boat models by British institutions (Dixon 2023, 51).

As this wider research context suggests, within the cultural heritage sector there is an increasing emphasis on virtual reunification and the interlinking and analysis of data from dispersed collections (Punzalan 2014; Hyvönen 2012). These efforts are often led by cross-institutional, large-scale aggregators, such as Europeana (de Boer V et al. 2012). The key stages of a data integration workflow tend to be described as comprising the following main steps: Harvesting, Ingestion, Mapping, Indexing, Storing, Monitoring, Enriching, Displaying, and Publishing (Siqueira, Martins 2022). Yet, the decision-making process for selecting data to be included in (and excluded from) projects has not been documented clearly, with internal dynamics and priorities often defining the scope of data ingestion. Importantly, this is interdisciplinary question that needs curators and technical teams to work on together something that is not always possible or straightforward given the university-museum partnerships, and associated funding schemes it requires.

Data-driven projects are as strong as the data they examine, and within cultural heritage institutions these datasets hold various nuances, biases and limitations. Digital collection databases may be incomplete if part of a collection is uncatalogued. An example of this can be seen in Sloane's collection of coins and medals, catalogued in "Coins vol 1" to "Coins vol 10" and "Medals". They were presumed to be among the few remaining manuscript items in the BM Coins and Medals department when it was hit by an incendiary bomb on the night of 10 May 1941, and they have not been located since the war. These catalogues were described in 1933 as "ten bound foolscap volumes, with indexes in several volumes, […] bound in ten volumes according to countries, apparently all coins" (MacGregor 1994, 164). As this suggests, the indices may have been included in the ten volumes and the medals must have been listed separately. There is very little in the BM's database to represent this vast area of Sloane's collection and any analysis of the data held on the collections database will in turn be skewed, though researchers are not necessarily alerted to this when they search the BM's database. This reflect how data within systems, both historical and contemporary, has been input from different human sources with varying degrees of bias and accuracy. If projects do not acknowledge, or are not aware of this, analyses, and indeed findings, may be flawed, skewed and biased as a result (Institute of Museum and Library Services 2018; Jones 2022, 7).

2.2 Collection Data Atlas

The term 'Data Atlas' is predominantly used within the domains of computer science and information science and can take several forms usually influenced by domain or project characteristics. Consequently, there is no single, universally agreed definition of the term 'Data Atlas'. It has been used across several digital projects as a high-level abstraction to enable exploration, analysis, and synthesises, or visualisation, of usually disparate datasets.[3] However, it has also been used within a single domain of homogeneous datasets, as in the case of the *Atlas of Digitised Newspapers and Metadata*. This atlas, part of the *Oceanic Exchanges* project examined in depth the metadata of digitised newspapers across 10 international datasets and was created in response to the need for standardisation to enable meaningful connection across different datasets (Beals et al. 2020). The atlas provided a comprehensive visualisation of all the metadata fields and established a set of mappings that showed how metadata fields related to one another across different datasets.

Wong 2102 defined 'Data Atlas' as a compendium of diverse data objects, including maps, lists tables, illustrations, or analysis, resulting in a total overview of an organisation's information systems. The study proposed the development of an atlas and described its development methodology as comprising three key stages: the organisation of a multidisciplinary team; identification of information in all repositories and systems; and use of findings to build the atlas (Wong 2012). Similarly, the "Virtual Laboratory of neuroscientific data" presented the design and development of a technologically competent but conventional system architecture, entailing the term *Data Atlas* as an abstraction of the collective relation of the two main layers of the system architecture (Munir et al. 2015).

The role of a *Data Atlas* has also been understood as an interactive web platform for standardising data collection from Public Libraries. This allows for the organisation of measurable results under the Common Impact Measurement System (CIMS), with the aim of enabling public libraries to quantify their impact on individuals and communities beyond the metrics of services. The example differs from those above, acting more like a framework of data collection across seven areas of the CIMS: inclusion, culture, education, communication, economic development, health, and governance (Schrag 2015). Similarly, within the domain of biodiversity conservation, *Data Atlas* has been situated as a standard method in biodiversity fieldwork, holding strong ties to the domain of biological observation, and

3 See Beals et al. 2020; Wong 2012; Vroom 2019; Schrag 2015; Robertson, Cumming, Erasmus 2010; Parimbelli et al. 2022; Munir et al. 2015

providing a well-defined procedure of data collection at the point of field species observation (Robertson, Cumming, Erasmus 2010).

The term *Data Atlas* also appears in archaeological fieldwork data, as a tool to archive, store, link, and make accessible data which combines artefacts, written sources and pictorial evidence as information sources (Vroom 2019). It is also applied to an information system produced by PERISCOPE partners of integrated data about the COVID-19 Pandemic and its effect on health, economics, policy-making and society (Parimbelli et al. 2022). This atlas makes data readily available to the research community, decision makers and the public as a means of amplifying research and its impact, acting as a central solution for exploration and analysis of georeferenced data.

Table 1 The five characteristics of data atlas development as they appear in literature across 7 reviewed projects

	Multidisciplinary Team	Inventory quality	Review & Inspection	Data Collection	Integration & Organisation
Beals	×	×	×		×
Munir			×	×	
Parimbelli	×		×	×	
Robertson		×	×	×	
Schrag	×	×	×	×	
Vroom		×	×		×
Wong	×	×	×		×

Across all definitions of *Data Atlas* discussed above, we can apply a fine distinction between approaches focused on a data collection framework that drives implementation (Robertson, Cumming, Erasmus 2010; Parimbelli et al. 2022; Munir et al. 2015), and those that enable integration, mapping or organisation of pre-existing data or resources under a commonly understood arrangement.[4] These two distinct approaches are summarised in Table 1, where the first column reflects data collection, and the second represents integration and organisation. The table further summarises the projects in terms of key characteristics, including whether they engage a multidisciplinary team, use the Data Atlas as an inventory instrument, or employ it for review and inspection of data landscapes.

In recent years, there has been increasing interest in the creation of methodological tools that can increase transparency around the creation and use of datasets within a range of digital projects, from AI to Natural Language Processing (NLP) and Machine Learning

4 See Beals et al. 2020; Schrag 2015; Vroom 2019; Wong 2012.

(ML) (Mitchell et al. 2019). This has led to the adoption of systematic documentation approaches aimed at transparency and accountability by gathering information about dataset motivation, composition, collecting, pre-processing, labelling, intended uses, distribution and maintenance (Bender, Friedman 2018). By operating at the level of datasets such documentation approaches increase understanding of the contents and uses of datasets for internal and external project stakeholders, aiming to mitigate potential bias, overgeneralisation, or exclusion within project results. Whilst sharing the motivation to increase understanding of datasets used within digital projects, our approach to documentation of cross-institutional collection-as-data environments differs in a key respect. Rather than working at the level of the individual dataset alone, our approach focuses also on entities (institutions or digitisation projects) holding physical and digital data. This aims to augment our understanding of complex, cross-institutional data environments and to manage complexity within highly heterogeneous and immense datasets, resulting in better decision-making, particularly when navigating large and dispersed collection landscapes. More broadly, this approach as discussed in the section below offers a methodological instrument that other collections-as-data projects can adopt as an intentional data audit stage, informing key decisions about data aggregation, prioritisation, and access.

3 Methodological Framework

We propose *Collection Data Atlas* as a documentation paradigm and methodological instrument for the panoptic representation of cultural heritage collections, which aims to support understanding and decision making with regards to the status, scope and accessibility of collection and sub-collection datasets. Within the context of large-scale aggregation projects, our definition holds specific relevance for data audit and prioritisation, the decision-making process for selecting data to be included (or excluded) from aggregation, digital tools, techniques and workflows created for projects. Concurrently, it is a trans-institutional information tool, integral to efforts to comprehend the history of dispersed collections. It provides an understanding of data relating to a collection beyond the parameters of institutional holdings, in turn giving a deeper understanding of the documentation that has shaped knowledge of a cultural heritage collection over its history.

Drawing on the state-of-the-art review discussed above, we realise the following key characteristics that crosscut and give rise to the otherwise many and varied instances of Data Atlases presented above **[tab. 1]**: *Multidisciplinary Team*, *Data Atlas Metaphor, Review*

and Inspection, and *Inventory Quality*. The *Multidisciplinary Team* brings together experts, human intervention, consultation, and the practice of iterative development. The *Atlas Metaphor* refers to the definitions of the units used within the framework and can either hold a Geo-referencing connotation, an all-encompassing view, or a high-level unification. The *Review and Inspection* characteristic of the atlas facilitates decision-making and inspection of a collections landscape to aid understanding about the origin, availability, use and other relevant domain characteristics of interest. The *Inventory Quality* captures the order of arrangement of the data and resources and how they are grouped and labelled, as well as how complete and consistent is the inventory. The following sections address the four key elements of the methodology based our experience developing a Collection Data Atlas for the Sloane Lab project.

3.1 Multidisciplinary Team

The pivotal role and significance of a multidisciplinary team in the development of a Collection Data Atlas is recognised as integral to its successful creation. Multidisciplinary and cross-institutional teams can facilitate the creation of comprehensive instruments that foster shared understanding of a complex historic collections' landscape among team members and disciplines.

Crucial to this is the collaborative process, supporting a shared understanding of classification, use of language, definitions, membership, and the overall scope of the atlas.

Our initial conception of an atlas arose from the need to sketch a list of relevant data resources for ingestion and unification within the data aggregation environment of the project (i.e. the Sloane Lab Knowledge Base). Previous studies focusing on the original Sloane collection and historical manuscripts served as the starting point for generating a list of valuable resources (MacGregor 1994; Nickson 1988; Walker 2022). The first iteration of the atlas collated information about datasets and helped the multidisciplinary team to realise their breadth, availability and state. Further iterations led the team to the recognition of a broader landscape of resources beyond the historical set of the original Sloane Manuscript Catalogues, including collection guides, body of objects and other auxiliary and tertiary resources. The involvement of a multidisciplinary team in the design process of the atlas promoted shared understanding between the Technical Team (consisting of computer, data and information scientists) and the Collections Team (consisting of historians and museum curators), making both individual members and their institutions aware of the broader landscape of the Sloane collection.

3.2 Atlas Metaphor

The term 'atlas' often reflects the idea of all-encompassing and comprehensive representation of a world. In its conventional sense, an atlas conveys spatial and geographical information, carrying maps, charts, graphical representations and associated information like place names, statistics, and other textual descriptions. Geographically referenced datasets containing demographic and environmental information are commonly referred to as a data atlas (Robertson, Cumming, Erasmus 2010). However, the term data atlas can also have a metaphorical meaning in the context of information and data science, signifying a comprehensive and holistic representation of a system or a data framework. In this sense, data atlases draw upon the traditional concept to describe modules, components, and attributes of synthesised representations, offering a rich and multidimensional view of complex systems and data structures.

The atlas metaphor as a comprehensive representation of a 'world' is a highly versatile and abstract concept, making it particularly well-suited for integration within the diverse field of heritage and critical heritage studies. Central to this metaphor is the notion of continents as main sections that hold material and curate the vast scale of a collection. Its inherent abstraction enables flexible adaptation to specific requirements of diverse projects, ensuring an efficient approach to fit the complexities of a cultural heritage collection landscape.

3.3 Review and Inspection

A Collection Data Atlas fundamentally aims to help users understand and navigate complex collection landscapes using visualisations that support decision-making regarding data modelling, aggregation, ingestion, and resource prioritisation. Accordingly, the atlas should make affordances that balance richness of information and simplicity of representation. The high-level organisation of a Data Atlas should allow the grouping of collections under major categories like institutions and projects, with further aspects of collection particulars communicated through arrays of additional information. The combination of collection grouping and array casting of attributes enables a robust framework for insightful review and inspection. It supports users to select and scrutinise collections based on their distinctive characteristics, facilitating a comprehensive understanding of their origin, availability, and utilisation. Furthermore, this can accommodate evolving changes to meet the growth of collections, modifications in their characteristics,

and any further digitisation of the collection and its related documentation.

3.4 Inventory Quality

A Collection Data Atlas should support comprehension and navigation of complex and extended data landscapes. Thus, an inventory quality that can communicate organisation and membership across various collections and units is vital not only for the use of an atlas but also for its maintenance and growth. During the development of the Sloane Lab Data Atlas, we realised five design principles to ensure the inventory quality of the atlas namely *Clarity*, *Consistency*, *Analytical Arrangement*, and *Flexibility*.

Clarity seeks to prevent user confusion through cognitive overload and to support analysis and decision making based on a commonly shared and understood set of semantics about the type, state and availability of collection dataset. This is further reinforced by the Data Atlas Taxonomy discussed in the section below that supports *Consistency* via a hierarchical organisation and grouping of the various levels of item granularity. *Analytical Arrangement* refers to the use of an array of features per type of collection to ease the inspection and examination of individual units, allowing users to gain insights at a more granular level. Such arrangements reflect format, level of digitisation, digital availability, physical location and other useful attributes of items. Maintaining *Flexibility* allow accommodation of nuanced adjustments critical to the all-encompassing aims of an atlas, ensuring that it remains responsive to evolving needs and diverse requirements. This is achieved through an adjustable array of features and attributes that can be tailored to each item type, whether physical or digital. However, while catering for flexibility, there is a potential trade-off with design economy as it may lead to item replication.

Our approach focuses on the acquisition and collective view of datasets within the Sloane collection with the aim to integrate datasets into the Sloane Lab's knowledge base. In this process we define the scope and accessibility of these resources and identify their digitisation status and format, which enable us to design suitable data mapping and ingestion methods. Furthermore, our objective is to provide a methodological abstraction, particularly valuable to the field of DH, to support future projects in managing and comprehending complex landscapes of collections dispersed across, various institutions and systems of varying accessibility status.

4 The Sloane Lab Data Atlas

Building on the Collection Data Atlas method, we implement the Sloane Lab Data Atlas to provide a comprehensive view of the historical catalogues and modern datasets related to the Sloane collection, which are currently dispersed across various national institutions and individual projects. The implementation delivers a series of design affordances that balance the richness of information with ease of use across a complex cultural heritage collection landscape. It is also scalable, allowing continuous improvements and expansion as our understanding of the boundaries of the Sloane Collection develops. The Sloane Lab Data Atlas is realised as a tabular visualisation in the form of a Microsoft Excel Spreadsheet, because this allows for continuous improvement and expansion, whilst offering a range of options for formatting and high-resolution exports. Critically, Excel files are common tools that scholars and practitioners are familiar with (e.g. for searching, filtering, exporting data), flattening the learning curve for end users. The atlas **[Appendix]** is also scalable, allowing continuous improvements and expansion as our understanding of the boundaries of the Sloane Collection develops and has been made available on GitHub[5] as an open-source project and also directly available in PDF format[6] from the Sloane Lab portal.

Our approach focuses on integration, with the overarching aim of bringing together the many records, historical and modern, that describe the Sloane collection within a unified information environment called the Sloane Lab knowledge base. To achieve this, we establish the scope and accessibility of these records and identify their digitisation status and format in order to design suitable data mapping and ingestion methods. Furthermore, our objective is to provide real word case and working example to support future projects in managing and comprehending complex landscapes of collections dispersed across various institutions and systems with varying levels of accessibility.

4.1 Data Atlas Taxonomy

The Sloane Lab Data Atlas represents a diverse range of resource types and items, each having distinct forms, sizes, purposes, and granularity levels. Such resources expand beyond the original

5 See Github project page at `https://github.com/sloanelab-org/data-atlas`.

6 See PDF version of the of the Sloane Lab Data Atlas at `https://sloanelab.org/wp-content/uploads/2024/09/Data-Atlas-Vector.pdf`.

e-ISSN 2724-3923

Sloane manuscript catalogues, including historical resources such as collection guides, correspondence, minutes from when Sloane served as Secretary of the Royal Society (1693 to 1713), as well as modern resources such as digital surrogates, digitally born documents and databases. The Data Atlas taxonomy is designed for organising and categorising these resources based on their inherent characteristics and relationships. It is a hierarchical, structured framework designed to facilitate consistent grouping and provide clear definitions, enabling a comprehensive understanding of the breadth and depth of items, aiding classification, and supporting scalability. Consequently, the taxonomy serves as a system that allows the development of a shared language and structure to support communication and analysis of the participating resources within the Data Atlas.

The principal and foundational concept within the taxonomy is the 'Collection Unit' which refers to a physical or digital entity in a collection. It is not an abstract notion but rather a distinct entity characterised by specific attributes that are instantiated for both physical and digital items, including but not limited to size, location, degree of digitisation, transcription type, availability, accessibility, metadata schema, and programmable access. The notion of the collection unit originates from the NHM's *Join the dots* collections assessment exercise, where collections are arranged into discrete units that reflect how curators organise, index and work with their collections (Miller 2020). Our definition is elastic to allow the use of the Collection Unit as a building block to create a visual representation of the historical and contemporary Sloane collections. The taxonomical arrangement of the Collection Unit is presented in Figure 1.

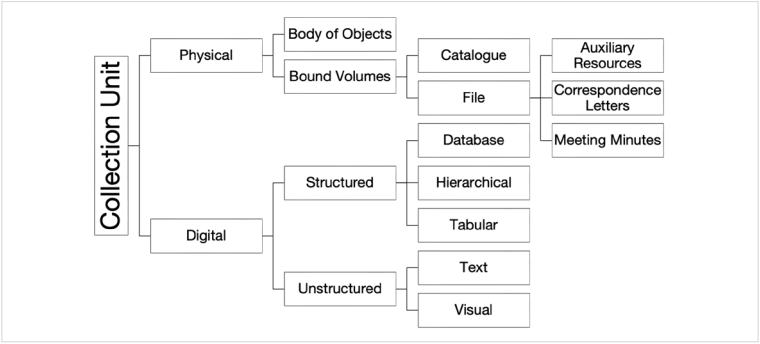


Figure 1 The Data Atlas Taxonomy

The Collection Unit may be Physical or Digital. Physical units are tangible and characterised by materiality. They can be touched or observed. Because the Sloane collection is historical, most units

contained in the Data Atlas are physical, for example, manuscript catalogues, collections of printed books, artifacts, artworks, specimens, etc. The digital surrogates of physical units, generated through a digitisation process such as scanning, photography and OCR (Optical Character Recognition) are considered attributes intrinsic to the physical unit rather than independent digital entities. This taxonomic design maintains a crucial distinction: between digital records of physical items (surrogates) and digital records about physical items (eg. representations and digital born material). This separation ensures that surrogate and born-digital materials remain distinct categories. The rationale reflects the atlas's primary focus on organising datasets and records of collections (i.e. the data records themselves) rather than the physical objects they describe.

Digital Units are retro-digitised or born-digital entities, either separate and distinct from any physical unit in the collection or born out of a digitation process, containing distinct and enriched attributes that augment and add value to the physical unit beyond its original form. An example is the fully searchable transcriptions in Text Encoding Initiative (TEI) format of the Sloane manuscript catalogues produced by the Enlightenment Architectures project (Enlightenment Architectures 2020). They introduce a set of important features and attributes such as Data Schema, Interactive Edition, Programmable Access, and Digital Location that are distinct from those associated with the original Physical Unit (i.e. manuscript catalogue). Similarly, the Sloane Letters Project[7] as a distinct digital collection unit, extends the volumes of Sloane correspondence held at the BL by adding summaries of letters, notes, item information and descriptive metadata (Sloane Letters n.d.). Inevitably, the differentiation between Physical and Digital Units may result in duplications throughout the Data Atlas, especially for collection units involved in extensive digitisation projects that yield substantial and enriched outputs. To establish a connection between these separate units originating from the same historical resource, the attribute *Origin* is introduced for digital units. This ensures a link between the distinct manifestations of the unit across different continents of the data atlas.

Additional categories are incorporated into the taxonomy to further classify Physical and Digital units. The Physical Unit is divided into Body of Objects and Bound Volumes. The Body of Objects consists of individual tangible items such as artefacts, artworks, specimens, etc. which are collectively understood as a curatorial unit but have not been arranged into a written catalogued form. Conversely, Bound Volumes are tangible objects, often comprised of written material bound together into single or multiple volumes or books, including

7 See project website The Sloane Letters Project. `https://sloaneletters.com/`.

materials that adopt the form of volumes or books, such as herbaria, where the written component supplements plant specimens pressed into the folios of a volume.

Bound Volumes are further categorised into Catalogue and File. The Catalogue follows the conventional definition of a systematic organisation of items arranged in a specific order, typically accompanied by descriptive information, providing details about the item's attributes, specifications, or characteristics. Conversely the File, includes all written components usually produced for record-keeping purposes that have a structured collection of information but are not catalogues. Files can be further classified into Meeting Minutes, Correspondence Letters, and other Auxiliary Resources.

The Digital Collection Unit is classified into Structure and Unstructured Units. The Structured Unit is organised and formatted according to a predefined schema, which may be well-defined and standardised. Structured units typically establish relationships and properties of data elements, making them amenable to querying and analysis using computational methods. Further categorisation of the structured units includes all types of Databases (i.e. relational and NoSQL), Tabular datasets such as spreadsheets, and Hierarchical XML data structures such as TEI files.

The Unstructured Unit does not offer the same level of structure and organisation to data as the Structured Unit, but it can be complex and rich in information. Such units are digitally born and can be further divided into Document and Visual. The Text is predominantly a text file that may contain figures and other pieces of information arranged in tabular format, available as a Word document, PDF file, Blog post and similar. The Visual unit is presented in form of images, maps, 3D models and other visualisations. Such visualisations may be the result of a tertiary implementation aimed at digitally reconstructing and preserving physical resources.

4.2 Data Atlas Continents

The Sloane Lab Data Atlas is arranged in rectangular panels, which are referred to as continents as per the atlas metaphor. Each continent visualises an institution or a major project that holds Collection Units related to the Sloane Collection. Collection units are visualised as different rows, which are grouped using appropriate categories based on the Data Atlas Taxonomy. To visualise different manifestations of the same Collection Unit (e.g. digital surrogates of a historic manuscript catalogue), a combination of colour coding and the use of separate columns within the same row is used. Footnotes clarify duplication of units appearing in more than a single continent. For example, the TEI manifestations of all collection units visualised

on the continent *Enlightenment Architectures* are also visualised on the Natural History Museum (NHM), the British Museum (BM) and the British Library (BL) continents respectively.

The section below outlines the details of the Sloane Lab Data Atlas, which organises both physical and digital collection units under distinct atlas 'continents'. Four institutional continents and three project continents are discussed whilst the details of an eighth continent, *Sloane Manuscript Catalogues Today* are presented, which aggregates all manuscript catalogues located across various heritage institutions.

4.2.1 Institutional Continents

The Institutional continents of the Sloane Lab Data Atlas include collection units from the NHM, BM, BL, and the Royal Society, with all institutions listing both Bound Volumes and Groups of Objects, except for the Royal Society, which lists only Bound Volumes. The NHM continent **[fig. 2]** holds 20 original Sloane manuscript catalogues, mostly related to fossils, insects, minerals, vegetables, and other specimens. It also contains 6 collection guides, including 3 volumes of the John Ray Historia Plantarum, 265 Sloane Herbarium Volumes, and 2 additional 'Bound Volumes' attributed to Leonard Plukenet (1641-1706) and James Petiver (1665-1718). The material has varying levels of digitization, transcription, and availability, as shown on the map. In addition, the NHM continent contains various contemporary datasets, including algae, fungi and plants, insects, invertebrates, palaeontology, mineralogy, and vertebrates.

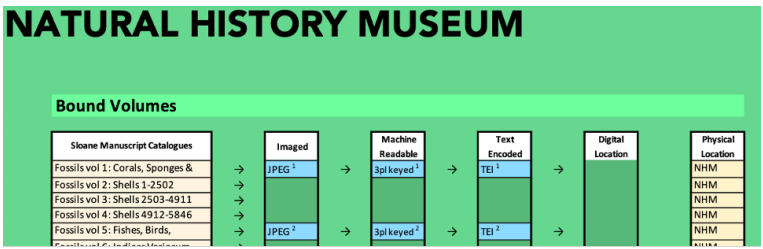


Figure 2 Part of the Natural History Museum dataset holding the original Sloane manuscript catalogues

The BM continent **[fig. 3]** includes four volumes of original Sloane manuscript catalogues, primarily focused on physical objects such as gems, cameos, amulets, and various other miscellaneous items. These manuscripts have varying levels of digitization and transcription and are physically housed at the British Museum. In addition, the continent organises approximately 15,000 objects from the BM's

Sloane collection across the eight curatorial departments where they are currently held: Prints & Drawings, Britain, Europe and Prehistory, Greece and Rome, Middle East, Asia, Coins and Medals, Africa, Oceania and the Americas, and Egypt and Sudan.

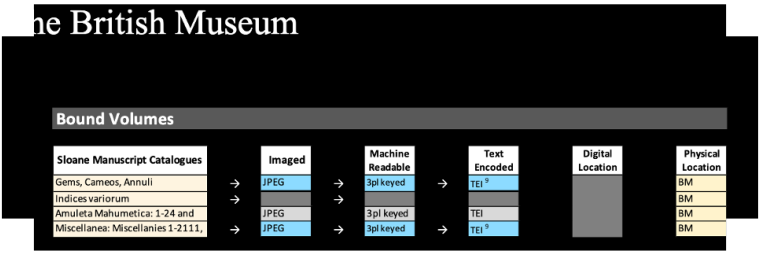


Figure 3 The British Museum dataset holding the original Sloane manuscript catalogues

The BL institutional continent **[fig. 4]** contains 21 volumes of original Sloane manuscript catalogues, each with varying levels of digitisation and transcription, detailing the large collection of printed books and manuscripts gathered by Sir Hans Sloane. At the time of his death, Sloane's library was estimated to hold 50,000 volumes, of which approximately 47,200 were printed items. His manuscript collection, comprising around 5,200 items, covers a wide range of subjects, including medicine, alchemy, chemistry, botany and horticulture, exploration and travel, mathematics, natural history, magic, and religion. The entirety of Sloane's collected printed books and manuscripts is catalogued in the British Museum's online catalogue and organized into relevant datasets. Additionally, the BL continent includes two separate entries for datasets containing approximately 1,400 of Sloane's correspondence letters.

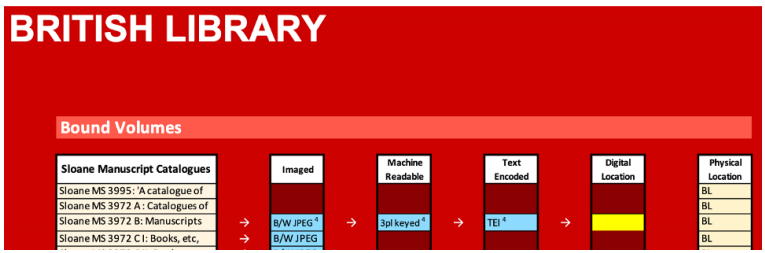


Figure 4 Part of the British Library dataset holding the original Sloane manuscript catalogues

The Royal Society's collection **[fig. 5]** includes five separate units of meeting minutes spanning the years 1686 to 1711, containing summaries of Royal Society meetings, documenting discussions on

experiments, publications, and natural curiosities. The majority of the manuscripts were compiled by, or on behalf of, Hans Sloane, who served as Secretary of the Royal Society from 1693 to 1713, and later as President (1727-41). The minutes from 1686 to 1711 have been digitised, imaged, and transcribed, and are available online through the Royal Society's.

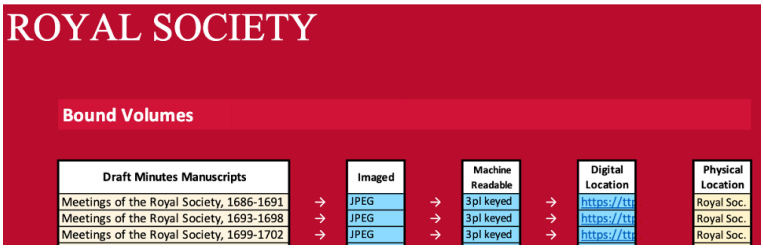


Figure 5 Part of the Royal Society dataset holding the original Sloane manuscript catalogues

4.2.2 Project Continents

The project continents contain collection units related to the work of significant research projects that investigated, organised, and digitised Sir Hans Sloane's original manuscript catalogues and other relevant materials from his collection. The Adam Matthew Library's digital data includes the Commonplace book manuscripts collected by Sloane, all of which are imaged and transcribed into machine-readable formats. The Sloane Letters Project is a database encompassing thirty-eight volumes of Sloane's correspondence held at the British Library (MSS 4036-4069, 4075-4078). It includes descriptions and metadata for Sloane MSS 4036-4035 and 4075, along with some letters and transcription from MSS 4054-4055, 4066, 4068-4069, and 4076. The Enlightenment Architectures continent includes the Text Encoding Initiative (TEI) outputs of the project, funded by the Leverhulme Trust. The project was a collaboration between the British Museum and University College London, with contributions from the British Library and the Natural History Museum. The project transcribed and created open-access, TEI-encoded remediations of five of Sloane's catalogues: two volumes on 'fossils', one volume of printed books and ephemera, one volume of 'miscellanies', and one cataloguing his collection of manuscripts (Ortolja-Baird et al. 2019, 44).

4.2.3 The Sloane's Manuscript Catalogues Today

The 'Continent' aggregates and indexes the surviving Sloane manuscript catalogues, which have been used and annotated by subsequent curators for three centuries now. During his lifetime, Sir Hans Sloane produced catalogues of his entire collection, which by 1725 had grown to 32 in number. The size of the catalogues varied, with some catalogues listing thousands of items and others being relatively short and probably combined or even bound by Sloane into single larger volumes. At his death in 1753, the numbers of items in each of his 32 catalogues had increased substantially, as recorded by the list transmitted to his executors after his death (MacGregor 1994, 28-9). The original catalogues continued to be used over the next centuries by museum curators. As curatorial departments grew, merged and changed, catalogues were copied, split up and recombined for ease of reference. In 1998 Peter M. Jones published the book *A preliminary checklist of Sir Hans Sloane's catalogues* which set the basis of a fully descriptive listing of Sloane's Catalogues (Jones 1988), and together with MacGregor A. in 1994 the book *Sir Hans Sloane, Collector, Scientist, Antiquary, Founding Father of the British Museum*. Both list 31 catalogues of the Sloane collection in total. The Jones/MacGregor numbering of catalogues (i.e. 1-31) reflects the complex bibliographic and curatorial history of the collection, which in some cases led to the creation of new catalogues for ease of reference, based on the organizational structure of heritage institutions. In other cases, multi-volume catalogues, such as the catalogue of printed books and manuscripts (12 volumes) were assigned a single number, while other catalogues, originally bound together by Sloane, such as Minerals (5 volumes) and Fossils (6 volumes) were split into individual numbers. Jones and MacGregor also included in their numbering volumes that were indices to various hand-written catalogues created by Sloane's assistant Thomas Stack (d.1756).

The Sloane Lab Data Atlas diverges from the Jones/MacGregor numbering system by listing Sloane's original catalogues by individual volume when these are predominantly by him as a catalogue of his collection **[fig. 6]**. However, the Jones/MacGregor numbering is maintained as a concordance to facilitate reference to earlier projects and publications. This explains why the number of catalogues in the atlas exceeds the 31 catalogues listed by Jones and MacGregor, while also omitting some catalogues they included.

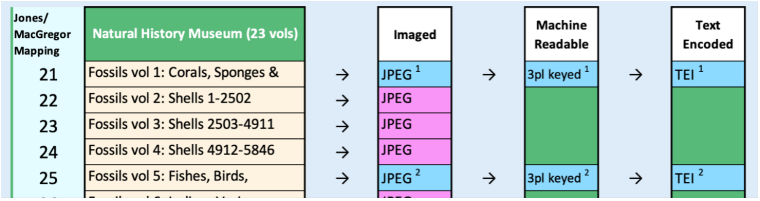


Figure 6 Part of the Sloane's Manuscript Catalogues Today held by the Natural History Museum (NHM), illustrating the MacGregor mappings

4.3 Data Atlas Attributes and Organisation

The atlas implements the inventory design principle by organising collection units by continent and arranging the continents into distinct sections while using colour coding, grouping, and attribute assignment to further enhance structure and clarity. Each continent is assigned a unique background colour to distinguish between institutions or digitisation projects and is organised into sections based on collection unit types, such as 'Bound Volumes' and 'Group of Objects' **[Appendix]**. Sections contain one or more tables, representing the different groups of collection units, with each table row containing an individual collection unit. The table columns represent various attributes assigned to collection units: for Bound Volumes these distinguish between a physical manuscript catalogue and its digital surrogates. For Group of Objects they reflect attributes related to state, size and availability. Moreover, cell colour is used to represent further attributes of collection units related to format and current availability, such as digitally available (blue), digital in the process of becoming available (grey), tangible objects (ivory), and online URL (yellow). Figure 7 illustrates an example arrangement of British Museum collection units, divided into 'Bound Volumes' and 'Group of Objects'. The different levels of digitisation of historical manuscripts are shown as Imaged JPEG, Machine Readable triple keyed transcription, and Text Encoded in TEI while the contemporary, databased catalogue, of Prints and Drawing contains a separate set of attributes about the state, size and availability of the collection.

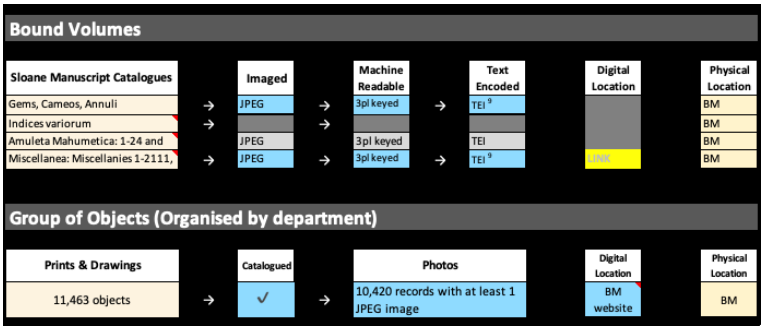


Figure 7 Part of the British Museum atlas continent illustrating organisation of collection units for Bound Volume and Groups of objects

5 Discussion and Reflections

5.1 Benefits and Limitations

The Sloane Lab Data Atlas presents a panoptic visualisation of a complex bibliographic and data environment and represents the first time that the contemporary data and institutional landscape of Sloane's collections has been brought together. As a 'collections as data' project, clarity around the scope, volume, availability, nature and digital accessibility of the project's data is of central importance to the Sloane Lab, its team, partners and stakeholders. The Atlas delivers this, bringing together the expert knowledge of curators and information scientists to act as an organisational tool for infrastructure development. It has aided the design of appropriate data mapping, modelling and ingestion approaches for the Sloane Lab project and wider development of its Knowledge Base. It has enabled partner institutions to identify data absences in their digitised collections that led to revisiting digitisation pipelines to cover such absences. Furthermore, it has given cultural heritage institutions and universities alike an important further case study of the importance of institutional collaborations between curators, academics and technical teams.

Most importantly, the Collection Data Atlas is a methodological instrument that can be beneficial to large-scale digital projects that deal with complex data environments in three distinct ways. Firstly, it augments a project team's understanding of different information systems, what data they hold, and their level of digitisation and availability is. Secondly, it facilitates the scoping out of complex data environments and improves project-wide decision-making, namely the prioritisation of datasets for ingestion into, for example, a Knowledge

Base and the effective budget allocation concerning digitisation for a given project and institutional partners alike. And thirdly, it allows for these complex data environments to be better communicated with stakeholders such as funders, institutional collaborators and with end-users.

These three key benefits are embodied by the Sloane Lab Data Atlas. However, to achieve and maintain this balance, a series of accommodations were made. Whereas the high-level organisation of the Data Atlas is well defined, at the lower level (i.e., closer to the data) the conventions of the Sloane Lab Data Atlas are purposely flexible. A prime example of this is how the different columns are understood differently depending on the section, as explained above. Moreover, whilst the Sloane Collections are vast, the Sloane Lab Data Atlas visualisation is agile and scalable to accommodate future changes and modifications. Accordingly, though a graphic visualisation made using a design tool may be subjectively more appealing, the Sloane Lab Data Atlas has been developed using Microsoft Excel, balancing these desiderata. Added to this, with the current approach we have encountered the issue of double instancing. For the purposes of simplicity, the Sloane Lab Data Atlas addresses this issue with explanatory footnotes, something that a technically strict visualisation may not allow for.

A significant future avenue for the development of the Data Atlas would be to bring into question the unnamed, enslaved people who related to some of the objects in Sloane's collection (see Introduction) and thus with the historical records of it that are mapped by the Atlas. In their 2022 paper on the question of silence and bias in the early-modern archive, Ortolja-Baird and Nyhan extracted the names of c. 3,000 people and 600 geographical locations from just two of Sloane's Manuscript Catalogues, noting how this quantitative examination of historical records ":attest a discrepancy of people and places in colonial and imperial contexts, and hints at just how many persons are absent from Sloane's network" (Ortolja-Baird, Nyhan 2022, 855). As the authors note, the individuals most named within Sloane's catalogues are British or European individuals. This raises a question to be addressed in future work of how the Data Atlas can make visible the quantity of marginalized voices who were central to the way in which Sloane's knowledge of his collections was formed and catalogued within his lifetime, in turn influencing the cataloguing and curation of his collections within the institutions in which they are now held, and ultimately the digitisation of historical records. With each stage in this knowledge making process the polyvocal sources of Sloane's knowledge have been further silenced. Following Lawther, this question has the potential to bring people back to the centre of Sloane's collections, in turn addressing power dynamics and hierarchies that have been present within the

collection and its record for c. 340 years (Lawther 2023). This it only acts to emphasise the importance of recognising the genealogy of museum documentation (Goskar 2024), and the impact that the history of a collection or institution's documentation has on the future ability to identify provenance (MacDonald 2023, 317). The Collections Data Atlas addresses these challenges by providing a comprehensive documentation framework that clarifies the scope, volume, availability, nature, and digital accessibility of collection and sub-collection datasets, supporting this way both historical understanding and future provenance research.

6 Conclusion

In this paper, the Collection Data Atlas has been positioned as an instrument that enables data-driven researchers to work with historians, curators, bibliographic specialists and others to establish what is extant, how it has been recorded, and, crucially, who has created it. The atlas presented is thus an instrument for data collection, inventory and multidisciplinary exploration. In the applied instance (i.e. Sloane Lab Data Atlas) explored in this paper, it is also the first synthesis of its kind, and visual representation of the historical and contemporary collection records of Sloane, now dispersed across different institutions, information systems and infrastructures.

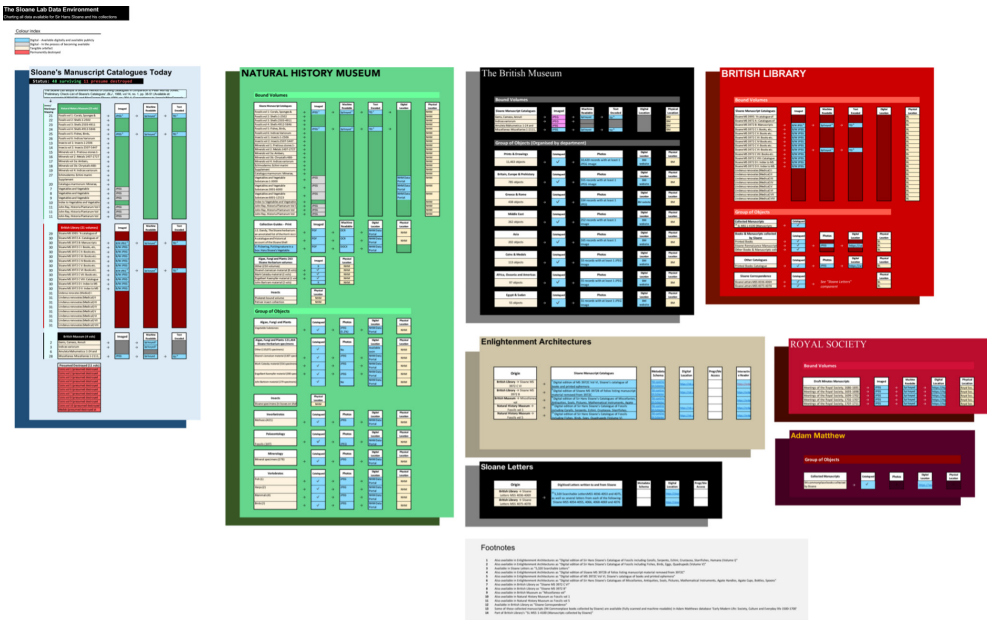
Underlying our discussion is an acknowledgement of three axes of the history of Sloane's documentation: firstly, the way in which Sloane and his contributors catalogued his collection within his lifetime; secondly, how the collection has been recorded and curated by the varying institutions it is now housed within and, thirdly, the digitisation practices by cultural heritage institutions and digital projects of the historical and institutional documentation relating to the Sloane collections. Sloane's collection was created through the economic, political and cultural processes of Britain's increasing global entanglements of the seventeenth and eighteenth centuries. As such, the question of who has created these historical records is central to further an understanding of the dynamics of power that have shaped knowledge around Sloane's collections, and the impact this has on the contemporary contexts within which his collections are located. This has been identified for a next step in the development of this tool. As it stands, the Sloane Lab Data Atlas is the first step in the wider project's development of new and better computational approaches to the detection and visualisation of these past entanglements ever present through gaps, biases and exclusion within the digitally available historical record. Likewise, following the transferability of the data atlas, we hope that it can be a tool that

can further conversations and computational interventions about this in other data-driven digital projects.

So too, the range of documents in which the Sloane collection is described and contextualised is vast. It extends beyond some 40 manuscript catalogues compiled by Sloane and his assistants and the records of the national institutions mentioned above. They include historical resources such as collection guides, correspondence, minutes from when Sloane served as Secretary of the Royal Society (1693 to 1713), as well as modern resources such as digital surrogates, digitally born documents and databases.

Appendix: Snapshot of the Data Atlas

The appendix presents the Sloane Lab Data Atlas in full deployment, comprising all eight data contents that contributed to the extended version of the Atlas for the data review purposes of the Sloane collection. The datasets include the Sloane manuscript catalogues as they appear today, the datasets held by the Natural History Museum, the British Museum and the British Library, as well as datasets that have been handled by individual projects including the Enlightenment Architectures, the Sloane Letters, the Minutes of the Royal Society and the Adam Matthew's digitisation project.



Bibliography

Beals, M.; Bell, E., Cordell, R.; Fyfe, P.; Russell, I.G.; Hauswedell, T.; Neudecker, C.; Nyhan, J.; Oiva, M.; Pado, S.; Pimentel, M.P. (2020). *The Atlas of Digitised Newspapers and Metadata: Reports from Oceanic Exchanges*. Online: Figshare, Loughborough University. https://doi.org/10.6084/m9.figshare.11560059.

Bender, E.M.; Friedman, B. (2018). "Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science". *Transactions of the Association for Computational Linguistics*, 6, 587-604.

Berger, M. (2023). "Between Canon and Coincidence". *Journal for Art Market Studies*, 1. https://doi.org/10.23690/jams.v7i1.147.

Caygill, M. (1994). "Sloane's Will and the Establishment of the British Museum". MacGregor, A. (ed.), *Sir Hans Sloane: Collector, Scientist, Antiquary, Founding Father of the British Museum*. London: British Museum Press.

Cornish, C.; Driver, F. (2020). "'Specimens Distributed': The Circulation of Objects from Kew's Museum of Economic Botany, 1847-1914". *Journal of the History of Collections*, 32(2), 327-40. https://doi.org/10.1093/jhc/fhz008.

de Boer, V. et al. (2012). "Supporting Linked Data Production for Cultural Heritage Institutes: The Amsterdam Museum Case Study". Simperi, E. et al. (eds), *The Semantic Web: Research and Applications*. Berlin: Springer, 123-34.

Dixon, C. (2023). *Sailing the Monsoon Winds in Miniature: Understanding Indian Ocean Boat Models*. London: British Museum Press.

Enlightenment Architectures (2020). "Enlightenment Architectures: Sir Hans Sloane's Catalogues of His Collections". *Reconstructing Sloane*. https://reconstructingsloane.org/enlightenmentarchitectures/2017/07/19/featured-content/.

Gosden, C.; Larson, F.; Petch, A. (2007). *Knowing Things: Exploring the collections at the Pitt Rivers Museum, 1884-1945*. New York: Oxford University Press.

Hauswedell, T.; Nyhan, J.; Beals, M.H.; Terras, M.; Bell, E. (2020). "Of Global Reach yet of Situated Contexts: An Examination of the Implicit and Explicit Selection Criteria That Shape Digital Archives of Historical Newspapers". *Archival Science*, 20(2), 139-65. https://doi.org/10.1007/s10502-020-09332-1.

Hodel, T. (2023). "Konsequenzen Der Handschriftenerkennung Und Des Maschinellen Lernens Für Die Geschichtswissenschaft. Anwendung, Einordnung Und Methodenkritik". *Historische Zeitschrift*, 316(1), 151-80. https://doi.org/10.1515/hzhz-2023-0006.

Hyvönen, E. (2012). "Publishing and Using Cultural Heritage Linked Data on the Semantic Web". *Synthesis Lectures on the Semantic Web: Theory and Technology*, 2(1), 1-159.

Institute of Museum and Library Services (2018). *The Santa Barbara Statement on Collections as Data (v2)*. https://collectionsasdata.github.io/statement/.

Jones, M. (2022). *Artefacts, Archives and Documentation in the Relational Museum*. Oxford: Routledge.

Jones, P.M. (1988). "A Preliminary Check-List of Sir Hans Sloane's Catalogues". *The British Library Journal*, 14(1), 38-51.

Lawther, K. (2023). *People-Centred Cataloguing*. http://www.kathleenlawther.co.uk/wp-content/uploads/2023/01/people-centred-cataloguing-final.pdf.

Leiden University (2023). *Archaeologist Martin Berger explores Latin American collections with an ERC grant*. https://www.universiteitleiden.nl/

en/news/2023/09/archaeologist-martin-berger-explores-latin-american-collections-with-an-erc-grant.

Luther, L. (2022). "Introduction: Digital Benin". *Digital Benin*. https://digitalbenin.org/documentation/introduction.

MacDonald, I. (2023). "Counting When, Who and How: Visualising the British Museum's History of Acquisition Through Collection Data, 1753-2019". *Journal of the History of Collections*, 35(2), 305-20.

MacGregor, A. (ed.) (1994). *Sir Hans Sloane: Collector, Scientist, Antiquary, Founding Father of the British Museum*. London: British Museum Press.

Miller, G. (2020). *Collection Unit Definition and Guidelines (From "Join the Dots" Collections Assessment Exercise)*. Natural History Museum. https://data.nhm.ac.uk/dataset/join-the-dots-collections-assessment-exercise/resource/c3c83129-27f7-4bb9-9d3f-e85f4d550f55.

Mitchell, M.; Wu, S.; Zaldivar, A.; Barnes, P.; Vasserman, L.; Hutchinson, B.; Spitzer, E.; Raji, I.D.; Gebru, T (2019). "Model Cards for Model Reporting" = *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. New York, NY: Association for Computing Machinery, 220-9. https://doi.org/10.1145/3287560.328759.

Modern Migrants Project Team (s.d.). "The Research Project". *Modern Migrants: Painting from Europe in US Museums*. https://www.modernmigrants.art/.

Munir, K.; Ahmad, KH.; McClatchey, R. (2015). "Development of a Large-Scale Neuroimages and Clinical Variables Data Atlas in the neuGRID4You (N4U) Project". *Journal of Biomedical Informatics*, 57, 242-62.

Nickson, M.AE. (1988). "Hans Sloane, Book Collector and Cataloguer, 1682-1698". *The British Library* Journal, 14, 52-89.

Oceanic Exchanges Project Team (2017). *Oceanic Exchanges: Tracing Global Information Networks in Historical Newspaper Repositories, 1840-1914*. https://doi.org/10.17605/OSF.IO/WA94S.

Nyhan, J.; Vlachidis, A.; Flinn, A.; Pearlman, N.; Carine, M.; Hill, J.; Humbel, M.; Jansari, S.; Sloan, K.; Pickering, V.; Valeonti, F. (2025). "Final Report - Sloane Lab: Looking Back to Build Future Shared Collections". *Towards a National Collection*. https://doi.org/10.5281/zenodo.14771754.

Ortolja-Baird, A.; Pickering, V.; Nyhan, J.; Sloan, K.; Fleming, M. (2019). "Digital Humanities in the Memory Institution: The Challenges of Encoding Sir Hans Sloane's Early Modern Catalogues of His Collections". *Open Library of Humanities*, 5(1), 44.

Ortolja-Baird, A.; Nyhan, J. (2022). "Encoding the Haunting of an Object Catalogue: On the Potential of Digital Technologies to Perpetuate or Subvert the Silence and Bias of the Early-Modern Archive". *Digital Scholarship in the Humanities*, 37(3), 844-67.

Parimbelli, E.; Larizza, C.; Urosevic, V.; Pogliaghi, A.; Ottaviano, M.; Cheng, C.; Benoit, V.; Pala, D.; Casella, V.; Bellazzi, R.; Giudici, P. (2022). "The PERISCOPE Data Atlas: A Demonstration of Release v 1.2". Michalowski, M (ed) = *Proceedings of the 20th International Conference on AI in Medicine (AIME 2022)*. Cham: Springer International Publishing, 412-15.

Penn, M.G.; Cafferty, S.; Carine, M. (2018). "Mapping the History of Botanical Collectors: Spatial Patterns, Diversity, and Uniqueness Through Time". *Systematics and Biodiversity*, 16(1), 1-13.

National Gallery of Art. "National Gallery of Art Collaborates with Researchers to Analyse Permanent Collection Data". https://www.nga.gov/press/2019/datathon.html.

Padilla, T. (2017). "On a Collection as Data Imperative". *UC Santa Barbara*. https://escholarship.org/content/qt9881c8sv/qt9881c8sv.pdf.

Padilla, T.; Allen, L.; Frost, H.; Potvin, S.; Russey Roke, E.; Varner, S. (2019). "Always Already Computational, Collections as Data Final Report". https://doi.org/10.5281/zenodo.3152935.

Petch, A. (2002). "Today a Computerised Museum Catalogue: Tomorrow the World". Special issue, *Papers Originating from the MEG Conference 2001* of *Journal of Museum Ethnography*, 14, 94-9.

Petch, A. (2006). "Counting and Calculating: Some Reflections on Using Statistics to Examine the History and Shape of the Collections at the Pitt Rivers Museum". *Journal of Museum Ethnography*, 18, 149-56

Phillipson, T. (2019). "Collections Development in Hindsight: A Numerical Analysis of the Science and Technology Collections of National Museum Scotland Since 1855". *Science Museum Group Journal*, 12.

Punzalan, R.L. (2014). "Understanding Virtual Reunification". *Library Quarterly: Information, Community, Policy*, 84(3), 294-323.

Rother, L.; Mariani, F.; Koss, M. (2023). "Hidden Value: Provenance as a Source for Economic and Social History". *Economic History Yearbook*, 64(1), 111-42.

Robertson, M.P.; Cumming, G.S.; Erasmus, B.F.N. (2010). "Getting the Most Out of Atlas Data". *Diversity and Distributions*, 16, 363-75.

Schrag, T. (2015). "Building a Public Library Impact Data Hub: A Global Libraries 'Data Atlas' for Storytelling, Strategy Development, and Collaboration". *International Federation of Library Associations (IFLA) World Library and Information Congress (WLIC 2015)*.

Shoilee, S.B.A. (2022). "Knowledge Discovery for Provenance Research on Colonial Heritage Objects". *Doctoral Consortium at ISCW 2022 co-located with 21st International Semantic Web Conference*.

Siqueira, J.; Martins, D.L. (2022). "Workflow Models for Aggregating Cultural Heritage Data on the Web: A Systematic Literature Review". *Journal of Association for Information Science and Technology*, 73, 204-24.

Solà, M.C.; Korepanova, A.; Mukhina, K.; Schich, M. (2023). "Quantifying Collection Lag in European Modern and Contemporary Art Museums". *VINCI '23: Proceedings of the 16th International Symposium on Visual Information Communication and Interaction*, 39, 1-8.

Vroom, J.A.C. (2019). "Data Atlas of Byzantine and Ottoman Material Culture: Archiving Medieval and Post-Medieval Archaeological Fieldwork Data from the Eastern Mediterranean (600-2000 AD), Phase 1". *Research Data Journal for the Humanities and Social Sciences*, 1-12.

Wingfield, C. (2011). "Donors, Loaners, Dealers and Swappers: The Relationship between the English Collections at the Pitt Rivers Museum". Byrne, S.; Clarke, A.; Harrison, R.; Torrence, R. (eds), *Unpacking the Collection: Networks of Material and Social Agency in the Museum*. New York: Springer.

Walker, A. (2022). "Sir Hans Sloane's Books: Seventy Years of Research". *British Library Journal eBLJ*, Article 6.

Wong, W. (2012). "Mapping Your Way to Compliance with a Data Atlas". *Information Management*, 46(1).

Zaagsma, G. (2022). "Digital History and the Politics of Digitization". *Digital Scholarship in the Humanities*, 38(2), 830-51. https://doi.org/10.1093/llc/fqac050.

e-ISSN 2724-3923

magazén
Vol. 6 – Num. 2 – December 2025

The Genetic Dossier in the Web of Data
From Documentary Collections to a Scholarly Archive

Elsa Pereira
University of Porto, Faculty of Arts and Humanities, CITCEM, Portugal

Abstract While archivists and genetic scholars differ considerably in their methodological frameworks, the digital turn in archival preservation and scholarly editing provides an opportunity to narrow the gap. This article examines how Semantic Web technologies can bridge differing approaches to documentary collections of contemporary authors, while also outlining two current challenges to this pursuit: some limitations of LOD in representing genetic dossiers in informative ways and a series of legal issues that prevent digital scholarly archives of genetic orientation from realising their full potential in the Web of Data.

Keywords Archive. Recordkeeping. Scholarly editing. Genetic dossier. Semantic web. Copyright.

Summary 1 Introduction. – 2 Literary Archives in Recordkeeping vs. Genetic Criticism. – 3 Transitioning Genetic Dossiers into the Web of Data. – 4 Conclusion and Outlook.

This article was supported by Portuguese national funds through FCT – Fundação para a Ciência e a Tecnologia, I.P., within the scope of UIDB/04059/2025 (DOI: 10.54499/UID/04059/2025).

Edizioni Ca'Foscari

Peer review

Submitted 2025-09-29
Accepted 2025-11-28
Published 2025-12-23

Open access

Citation Pereira, E. (2025). "The Genetic Dossier in the Web of Data". *magazén*, 6(2), [1-16], 179-194.

DOI 10.30687/mag/2724-3923/2025/02/003

1 Introduction

Archivists and genetic scholars differ considerably in their methodological frameworks, which may explain why they

> are largely not taking part in the same conversations, not speaking the same conceptual languages, and not benefiting from each other's insights. (Caswell 2016, 2)

The digital turn in archival preservation and scholarly editing presents a window of opportunity to bring the two fields together. As more GLAM institutions responsible for preserving and providing access to literary archives of modern or contemporary authors undertake large-scale digitisation of material under their custody, new possibilities emerge for digital scholarly projects to reconnect those resources within a Semantic Web ecosystem, fostering closer collaborations between genetic scholars and archival institutions. However, the Web of Data also presents specific challenges that warrant further research and attention.

After comparing differing organising approaches to literary archives in recordkeeping and genetic criticism, this article will consider how Semantic Web technologies may bridge the methodological frameworks of both fields, while outlining two current obstacles to this pursuit: some limitations of LOD in representing genetic dossiers in informative ways and a series of legal issues that prevent digital scholarly archives of genetic orientation from realising their full potential in the Web of Data.

2 Literary Archives in Recordkeeping vs. Genetic Criticism

Typically comprising the working and personal papers of modern or contemporary authors, literary archives are characterised by a wide variety of material,[1] whose significance and appreciation set them apart from most other types of archival repositories. Increasingly cherished by literary enthusiasts and collectors since the beginning

1 "The ideal literary collection captures the full gamut of a writer's work – background notes and research; annotated books and critical editions; literary drafts; photographic components; audio material; personal journals; literary logs; objects like keepsakes or awards; correspondence with publishers, editors and friends; editors' and printers' proofs, and final copies" (Molloy 2019, 328). The Group for Literary Archives & Manuscripts at the University of Manchester (GLAM) and the Group for Literary Archives and Manuscripts – North America (GLAMNA) have further systematised the material typically found in literary archives: `http://glam-archives.org.uk/?page_id=1731` (12/12/2022).

e-ISSN 2724-3923

of the Romantic *Geniezeit*,[2] the relevance of literary archives lies in "the insights they give into the act of creation", which translates into "a higher financial value" and may explain why an author's holographs end up being "scattered in diverse locations" worldwide (Sutton 2014, 295-6):

> Literary archives […] tend to travel much further than other types of papers and to be housed in unpredictable locations – often […] determined by market forces rather than by internal archival logic. […L]iterary papers are usually found […] to be divided between several collecting institutions. This phenomenon, which we have come to call 'split collections' […] represent[s] an essential part of the world of literary manuscripts. (Sutton 2018, 7-8)

Indeed, very few writers have their manuscripts entirely preserved in a single institution, whether a public or university library, a state archive, a private foundation, a literary house, or a museum. Not only do authors themselves disseminate documentary evidence of their writing, offering drafts as a memento or gift of friendship (Boie 1993, 42-43) to those with whom they correspond throughout their lives, but, after passing away, their estates are also frequently divided among heirs and subject to the sort of posthumous plunder that is implicit in the Latin root of the Portuguese term designating literary archives: *espólio* (from the Latin *spolia*, 'spoils' or 'stolen treasures').

While archival studies and literary genetic criticism approach the *spoils* of an author with very different methodologies, we believe their perspectives and understandings "overlap and can be brought together" (Bunn, Rayner 2019, 360).

Archival studies is a subfield of information science dedicated to "the nature, management, and uses of records", defined as "persistent representations of activities, created by participants or observers" (Caswell 2016, 3, 5). Traditionally, recordkeepers curating literary archives are guided in their work by a foundational principle of "respect for the fonds" (Muller, Feith, Fruin 2003, 54), which implies preserving the so-called "provenance" of archival documents that bear an organic relationship to one another. In practice, this means that documentary pieces shall not be aggregated by subject or any

2 With the privatisation of intellectual property and the concomitant valorisation of the creative genius during the Romantic period, writers started to preserve draft manuscripts systematically. However, it was not until the second half of the twentieth century that extensive literary archives proliferated, as "increasing attention was being given to the processes of literary composition and revision in their own right. At the forefront of this practice were French scholars associated with the Centre d'Analyse des Manuscrits in Paris, formed after the accession of Heinrich Heine's papers by the Bibliothèque Nationale in 1966" (Anderson et al. 2021, 8).

other interpretative criteria; instead, records created by different individuals must be "kept separately", with their "original order" and context preserved (Caswell 2016, 7), while classified according to genre, type, or material support (Lopes 2007, 55).

This traditional understanding of "provenance" as an organising principle of archival studies contrasts with the "speculative approach" (Drucker, Nowviskie 2004, 431) of scholarly editors with a "genetic orientation" (Van Hulle, Shillingsburg 2015, 36), who, conversely, view records not as a "fixed product" (Bunn, Rayner 2019, 369), but as "dynamic objects in motion" (Caswell 2016, 6).

Originating in France during the 1960s, in connection with the Centre d'Analyse des Manuscrits in Paris (later evolved in the current ITEM – Institut des Textes et Manuscrits Modernes), genetic criticism succeeded in adding a temporal, paradigmatic dimension to the literary text, regarded as a process rather than a product, by drawing attention to its variations in draft form and all the transformations that result from the author's writing or rewriting activity over time. This type of "archaeology of the manuscript" (Van Mierlo 2013) relies on literary archives, mainly from the nineteenth century onward, to provide insight into the compositional development of a literary work and expand the interpretation possibilities of the text:

> manuscripts […] offer up new and unseen material, and also suggest, in their very physicality, the writing methods and processes unique to the subject of study. They can further 'solve factual problems like the dating of a poem or establishing an accurate text' and 'illuminate the broader meanings of a literary work' (Gioia 2004: 36). Beyond this, archival materials offer us other conduits of research and knowledge, […revealing], as Cook argues, the 'context behind the text, the power relationships shaping the documentary heritage, and indeed the document's form and content'. (Stead 2016, 4)

For that to be in place, scholars must compile a *genetic dossier* (Grésillon [1994] 2016, 286) comprising all the physically dispersed documents of an author's writing project that "bear witness to the evolution of the work" (De Biasi 2004, 38). This may include the version records that preceded publication (e.g. notes, drafts, revised manuscripts, typescripts, print proofs), as well as other correlated evidence of the broader interpersonal networks that contribute to the author's creative process, such as his library and correspondence. The methodology involves not only collecting all extant genetic documentation (*recensio*) but also comparing the respective textual variants to infer the genealogical relationships among the collected pieces and organising the work's *avant-texte* (Bellemin-Noël 1972) according to the writing chronology.

Whereas recordkeepers use systematic guidelines in their archival practice, based on such principles as provenance, collective control, or original order of records, genetic scholars aggregate different authorial material of various provenance (e.g. marginalia in books, notebooks, draft manuscripts, typescripts, letters exchanged with other people), subjectively organising the jigsaw pieces into speculative archives – the *genetic dossier* – aimed at reconstructing the author's creative process.

In recent years, the digital medium has significantly facilitated the constitution of these interpretative *scholarly archives*,[3] allowing researchers to aggregate facsimiles and transcriptions of material scattered across different institutions and model their textual relations within a dedicated virtual environment for specific academic purposes.

3 Transitioning Genetic Dossiers into the Web of Data

Over the past decade, scholarly editorial initiatives, such as the Samuel Beckett Digital Manuscript Project,[4] the Shelley-Godwin Archive,[5] or the Gustave Roud: Textes & Archives,[6] have succeeded in digitally reuniting dispersed documentation of modern and contemporary authors, facilitating the examination of the genetic dossier of their works.

3 For more systematic definitions of "scholarly archive" in DH projects, distinguishing it from the traditional notion of "archive" in archival studies, see e.g. Theimer 2012; Adema, Stoyanova 2015.

4 Van Hulle, Nixon 2011-present. The project was developed by the Centre for Manuscript Genetics at the University of Antwerp, the Beckett International Foundation at the University of Reading, the Oxford Centre for Textual Editing and Theory at the University of Oxford, and the Harry Ransom Humanities Research Center at the University of Texas at Austin, with the permission of the Estate of Samuel Beckett.

5 Fraistat et al. 2013-present. The project aims to unite online the widely dispersed handwritten legacy of Percy Bysshe Shelley, Mary Wollstonecraft Shelley, William Godwin, and Mary Wollstonecraft. It is the result of a partnership between the New York Public Library and the Maryland Institute for Technology in the Humanities, in cooperation with Oxford's Bodleian Library, the Huntington Library, the British Library, the Houghton Library, and the Victoria and Albert Museum.

6 Jaquier, Maggetti 2022. Developed at the University of Lausanne and supported by the Swiss National Science Foundation (2017-2022), the project provides a critical print edition of Gustave Roud's complete works, and a genetic digital archive of the authorial material housed at the Centre des Littératures en Suisse Romande..

Although Wout Dillen has rightly noted that many of these DH projects call themselves *archives*,[7] their "archival impulse" (Eggert 2019) diverges from the principles followed by librarians and recordkeepers, who are invariably bemused by the term used in this context.[8] Instead of attending to the provenance of documents, digital scholarly archives of genetic orientation are "hermeneutical instruments" (Ramsay, Rockwell 2012, 79) aimed at organising a "purposeful collection of surrogates" (Price 2008) to reveal interpretative connections between dispersed textual witnesses. Presenting themselves as a "work-site" (Eggert 2005, 433), a "knowledge site" (Shillingsburg 2006, 88), or a "platform for learning" (Theimer 2014, 146), those projects actively engage with the dynamics of variation through a range of digital tools to allow readers to navigate multiple versions of a text and follow the author's compositional development over a more or less extended period of experimentation and revision.

In addition to XML-TEI markup and algorithmic collation, participatory editorial projects, such as the LdoD Archive,[9] have been exploring social editing functionalities, supported by structured databases and Web 2.0 environments that enable users to create *virtual editions* of the documentation.[10] Scholars are invited to manipulate the "dynamic layer of the archive" (Portela 2022, 191), either performing editorial script acts such as annotations, or reconfiguring the writing sequence of texts, in what some have also been labelling as new interactive "forms of analysis and creativity",

7 "[…] projects that are generally considered as digital scholarly editions often do not shy away from calling themselves archives […] – think, for instance, of the *William Blake Archive*, the *Piers Plowman Electronic Archive*, the *Walt Whitman Archive*, and, more recently, the *Shelley-Godwin Archive*. […] As the digital medium started to break down the borders between archives and editions, […] the user can decide how to use the digital resource: as an archive of textual documents and image reproductions; as a (genetic) dossier that organises these documents and exposes their internal logic; or as an edition, a curated and edited collection of texts that informs the reader on the textual tradition of the work" (Dillen 2019, 265, 267).

8 As Kate Theimer observed, "[a]rchivists would not refer to online groupings of digital copies of non-digital original materials, often comprised of materials (including published materials) located in different physical repositories or collections, purposefully selected and arranged in order to support a scholarly goal, as an 'archives' – and so the confusion of an Archivist tourist in the land of Digital Humanities" (Theimer 2012).

9 Portela, Silva 2017-present. Developed at the Centre for Portuguese Literature at the University of Coimbra and funded by the Portuguese Foundation for Science and Technology and the European Regional Development Fund, the LdoD Archive is a collaborative digital archive of the *Book of Disquiet* by Fernando Pessoa. It contains images of the autograph documents, transcriptions of those documents, and also transcriptions of four editions of Pessoa's work.

10 See Silva, Portela 2015.

in line with the poststucturalist "esthetic of the possible" (Gooding et al. 2019, 386, 376).

More recently, digital scholarly archives of genetic orientation have also been drawing inspiration from advancements in Linked Open Data (LOD) and other technologies that follow up on Berners-Lee's vision of a Semantic Web of Data (Berners-Lee et al. 2001; Berners-Lee 2006), to reveal and enhance the complex network of relationships devised among various documents of a genetic dossier. In a nutshell, these projects use persistent URIs to identify resources and apply web ontologies to formally represent relationships or the underlying logic among different nodes in the documentary network. Resource Description Framework (RDF) datasets, represented as subject-predicate-object triples, will model a graph structure that computers can interpret, while users can interact via a SPARQL endpoint with a graphical interface that leverages the semantic layer for querying and manipulating the graph database.[11]

Among the projects that have been applying Semantic Web technologies to genetic dossiers,[12] the *Gustave Roud: Textes & Archives* (Jaquier, Maggetti 2022) deserves special mention, as the team designed a new data model, formalised in the Web Ontology Language, specifically for Genetic Criticism.[13] This GeNO ontology effectively describes the interwoven networks within and outside an author's genetic dossier and can be queried using cURL and Gravsearch, a virtual graph search based on SPARQL that allows researchers to find, for instance, which diary entry of the author ended up in his fictional work.

The *Shelley-Godwin Archive* (Fraistat et al. 2013-present) is another initiative worth mentioning, as it builds on linked data principles and the Shared Canvas data model to support a participatory platform where anyone on the web can describe, discuss, and reuse facsimiles and transcriptions of archival material, within a global,

11 For a comprehensive perspective on graph data-models and Semantic Web technologies in scholarly digital editing, see Spadini, Tomasi, Vogeler 2021.

12 A noteworthy project developed in Italy is the digital edition of Paolo Bufalini's notebook (Daquino et al. 2020).

13 Geno – the Genetic Networks Ontology (Spadini 2023), which builds as an extension of the knora-base ontology. Other existing ontologies for semantic editions, such as CAO – Critical Apparatus Ontology (Giovannetti 2019) and CEO – Critical Edition Ontology (Martignano 2023) did not adequately describe the network of textual witnesses in relation with other witnesses in the genetic dossier. See Christen, Spadini 2019, 84.

interconnected network of information that aligns with the 5S model[14] and an interdisciplinary vision of "Linked Research" (Capadisli 2016). Interestingly, the project stems from a partnership with several public and university libraries, expanding still-rare collaborations between literary scholars and recordkeepers[15] into the Web of Data and opening up possibilities for further cooperation.

In fact, as more GLAM institutions lead large-scale digitisation projects that adhere to protocols such as the International Image Interoperability Framework (IIIF) and the Text Encoding Initiative (TEI), new opportunities emerge for archivists to incorporate digital scholarly archives into local descriptions, enhancing or contextualising their records. Conversely, scholars should also be able to connect genetic dossiers to the authors' archival repositories and personal libraries,[16] allowing users to navigate the virtual research interface without losing contact with the provenance and archival order of the material records. But while steps have been taken along the path, the vision of a global web of literary archives remains "far away on the utopian horizon" (Fordham, cited in Anderson et al. 2021, 7), hindered so far by at least two main obstacles.

The first issue that stands out is the lack of shared vocabularies, ontologies, and "good human-usable interfaces for the Semantic Web" (Brown, Simpson 2015). Recent initiatives promoting Linked Open

14 The 5S model refers to the fundamental concepts of Streams, Structures, Spaces, Scenarios, and Societies (5S) that formally model digital libraries, regarded as "a managed collection of information with associated services involving communities where information is stored in digital formats and accessible over a network" (Gonçalves 2004, 19). For a comparative approach between the frameworks of digital libraries, archives, and editions, see e.g. Meschini 2020, chapter 3.

15 "Think for example of *Litteraturbanken*, the 'Swedish Literature Bank' […]. In an impressive collaborative effort between literary and linguistic scholars, research libraries, and editorial societies and academies, this project contains a wide range of digital facsimiles and their (corrected OCR based) transcriptions of documents pertaining to Swedish literary works from the Middle Ages to the present. Alongside their edited texts available in HTML (and, when possible, EPUB), these are contextualised further by means of scholarly introductions, presentations, other didactic materials, and even allow for basic text analysis functionalities through a collaboration with Språkbanken, the 'Swedish Language Bank'" (Dillen 2019, 265).

16 Many twentieth-century authors have not only their manuscripts but also their libraries preserved and digitised. Among other examples, it is worth mentioning the Private Library of Portuguese author Fernando Pessoa (1888-1935), comprising roughly 1300 books once owned by the poet and currently available on the website of Casa Fernando Pessoa: `https://bibliotecaparticular.casafernandopessoa.pt/index/index.htm`. Although the fully digitised collection has been available for several research projects on Pessoa's marginalia, the digital library is not IIIF-compliant, which makes it not ideal to incorporate within a semantic web environment.

e-ISSN 2724-3923

Data vocabularies for the description of manuscripts[17] and textual variation[18] are promising contributions to addressing the problem, but so far, digital scholarly projects experimenting with ontologies to express textual relationships across different authorial materials have designed their data models independently of the archival records underpinning the projects.[19] One suggestion to overcome the current disconnect between literary archives and genetic dossiers, enabling rich, interlinked data to be shared and repurposed by third-party applications, could involve open knowledge bases such as Wikidata, as well as Solid Pods specifically designed for GLAM institutions to share archival records and promote decentralised data networks, as this kind of resource enables different web APIs to provide new views into the knowledge graph.[20] Still, while knowledge graph visualisation for the Solid ecosystem is making progress and paving the way for further research,[21] experts in graph technologies recognise that network graphs in general do not present complex datasets of textual information in a clear and intelligible manner, mainly because Linked Data is a machine-readable format not intended for humans,[22] and "[m]any levels of discursive mediation are

17 See e.g. the efforts developed by the Working Group for Linked Manuscript Descriptions, whose goal is to create a common Linked Open Data vocabulary for the description of medieval manuscripts from the Middle East. The working group met for its first sessions online on 15-16 December 2021 as part of the Linked Pasts VII Symposium, hosted by Ghent University. `https://www.ghentcdh.ugent.be/linked-pasts-vii-symposium`.

18 In this regard, see Bleeker et al. 2025. The authors have also recently established a working group on Visualizing and Investigating Differences In Texts (VIDIT), aimed at building a global community of scholars, developers, and designers interested in studying and visualising variation in historical and literary texts. `https://wg-vidit.github.io/`.

19 The *Gustave Roud: Textes & Archives*' data model, for instance, is independent of the archival online inventory, available at the Centre for Literatures in French-speaking Switzerland: `https://atom-archives.unil.ch/index.php/ch-000225-8-p73`.

20 Solid (SOcial LInked Data) is a set of technology specifications for the Web of Data, which includes decentralised online pods, ("often referred to as data vaults), standard communication between apps, and the use of a universal data format in the form of a Resource Description Framework (RDF) […]. The central notion of Solid is the technical and organizational separation of data, services, and identity. In this way, Solid as a set of technology specifications enables the creation of decentralized applications using W3C standards and protocols […], which counterweights the current dominant architecture of the internet" (Theys et al. 2025, 505).

21 In this regard, see e.g. Dedecker et al. 2022.

22 "ancora tanta strada c'è da fare nella realizzazione di applicazioni che siano in grado di utilizzare in modo sapiente quei dati per restituire all'utente sotto forma di nuova conoscenza. Partiamo dal presupposto che i LOD non sono pensati originariamente per l'utente, ma per l'elaborazione da parte della macchina" (There is still a long way to go in the realisation of applications to wisely use data and return it to the user in the form of new knowledge. LOD is not originally designed for the user but for processing by the machine) (Tomasi 2022, 132-3).

needed for the methods of close and distant reading to productively inform one another" (Stoyanova 2023, 39). The *Gustave Roud: Textes & Archives*, for instance, drew inspiration from celestial maps to explore the centrality of the author's diary within the genetic dossier, achieving positive usage test results among experts (Elli et al. 2023, 32, 36), but interpreting graph representations of such complex data sets is incredibly difficult for literary scholars without the technical knowledge for querying the ontology.[23] As such, we need archivists, genetic scholars, and data scientists to come together and codevelop graphic user interfaces that make graph network visualisation more accessible and "informative for a wider audience".[24]

In the case of modern and contemporary literary archives, however, another major obstacle to reconnecting authorial material within a Semantic Web ecosystem stands out, due to a fundamental conflict between the free availability of distributed data, implicit in the concept of LOD,[25] and a series of legal restrictions affecting authorial repositories, particularly in countries with a legal tradition of *Droit d'Auteur*. Despite usually being deposited in institutions funded by public resources, manuscripts and other archival material of twentieth and twenty-first-century writers is subject to a series of copyright and non-copyright restrictions that protect the privacy and moral rights of authors, forcing GLAM institutions to "curtail the widespread digitisation of whole collections" (Anderson et al. 2021, 7) and restrict access to "a minority of researchers who have the time and funding" to view the documents on site (Jaillant 2019, 290). In fact, those willing to use contemporary literary materials for research purposes often find themselves in a never-ending maze of bureaucracy, including formal authorisations from both copyright owners and custodians of the material, which implies dealing with different propriety, authority, dependency, and privacy restrictions:

> The owner of copyright for material in the Manuscripts Collection is the writer or creator of the material, or the creator's legal heir(s). Note that the donor of the material is not always the copyright

23 See e.g. "Constellation génétique de Campagne perdue de Gustave Roud". `https://roud.unil.ch/resources/http%3A%2F%2Frdfh.ch%2F0112%2FpKBOXI-GSEyVBBECW1Xgkw`. A simple way to improve the legibility of this genetic network would be incorporating weblinks to the different nodes connected in the graph, allowing users to navigate the digital scholarly archive taking the celestial map as a reference.

24 Statement issued by the VIDIT: `https://wg-vidit.github.io/`. For an overview of current graph visualisation techniques, specifically applied to collation outputs, see Birnbaum, Dekker 2024.

25 Linked Open Data is a combination of two basic concepts: linked data (a method of storing information based on the connections and relationships between items) and open data (signifying data that has been made freely available for distribution).

e-ISSN 2724-3923

> owner. In addition, many collections contain a variety of letters, diaries, documents owned by multiple copyright owners. […] Should you wish to publish material from the Library's Manuscript Collection, you will need to: declare your intention to the Library as custodian of the material, obtain copyright clearance from the copyright holder(s).[26]

As demonstrated in a previous article dedicated to major legal obstacles encountered by European genetic scholars, recent exceptions introduced by the CDSM Directive did not meet the requirements of ongoing advancements in digital humanities,[27] making the path towards publishing and providing online international access to contemporary literary archives difficult to navigate, especially for unpublished works, where "the waters are particularly muddy" (Dillen, Neyt 2016, 788). Before taking further steps towards a much-anticipated vision of genetic dossiers in the Web of Data, scholars investigating twentieth and twenty-first-century authors therefore need bold policy-making adjustments to ensure that their work will not be rendered worthless by someone refusing publication permission:

> we need to extend the scope of the available exceptions […] to allow for scholarly publication in the digital age – or otherwise, a legal license designed with scholarship in mind so that academic researchers may work with published texts and holographic materials in public archive libraries, disclosing research results (in person, on paper, and online) without interference from heirs or successors. Moreover, we also need national or European management systems led by independent copyright boards to facilitate the clearance of orphan works for different uses and reduce the randomness of our current authorisation system. (Pereira 2023, 523-4).

26 National Library of Australia, "Rights and the Manuscripts Collection". https://www.library.gov.au/services/copyright-library-collections/rights-and-manuscripts-collection.

27 European Directive 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market. Digital literary approaches affected by the TDM exception include: "classification and clustering of texts (e.g. for authorship attribution and stylometry), extraction of distinctive features, semantic analysis with topic modelling, analysis of polarity with sentiment analysis, character relationships with network analysis, and analysis of relationships between texts (e.g. in text reuse). However, we should note that only those materials to which scholars have lawful access can be mined, and experiences in countries where TDM exceptions have been in force show that copyright issues will subsist: 'Despite the TDM exception in German copyright law, Text and Data Mining (TDM) with copyrighted texts is still subject to restrictions, including those concerning the storage, publication and follow-up use of the resulting corpora'" (Pereira 2023, 521-2).

4 Conclusion and Outlook

In his 2014 book on memory and scholarship in the age of digital reproduction, Jerome McGann argued that to study literary creativity, scholars needed "cultural records to be comprehensive, stable, and accessible" while being able to augment "that record with our own contributions" (McGann 2014, 131-2). The proposal implied shifting the idea of archival records as fixed informational resources to embrace the digital scholarly perspective on the term, regarded as "a complex system inhabited by all the different agents involved in the production of academic work" (Bunn, Rayner 2019, 369).

In this regard, Semantic Web technologies present a window of opportunity to enhance interdisciplinary collaboration among archivists, textual scholars, and genetic critics, reapproaching their methodological frameworks to think anew about the working methods of prominent writers. While the persistence of practical obstacles to this pursuit leaves the full potential of genetic dossiers in the Web of Data untapped, much work has been done to overcome those shortcomings and interdisciplinary working groups must be formed to carry on the efforts, co-designing archival software for the Semantic Web ecosystem and simultaneously allowing for so-called "digital forensic work" on literary archives, i.e.:

> multiple modes of ordering and interpreting while also, at the same time, securing the collections that underpin this innovative work. (Gooding et al. 2019, 376)

Meanwhile, scholars and recordkeepers must come together and exert pressure on legislators to introduce the policy changes necessary to allow greater freedom in using copyrighted works for the preservation of cultural heritage in the Digital Single Market. The recent (albeit insufficient) TDM exception introduced into European legislation demonstrated that only through sustained commitment can we achieve the legal measures necessary to enable ongoing developments in computational literary studies. Allowing contemporary documentary collections and genetic dossiers to transition into the Web of Data should be our next goal.

Bibliography

Adema, J.; Stoyanova, S. (2015). "The Multidimensional Scholarly Archive". *Open Reflections*. https://openreflections.wordpress.com/2015/10/28/the-multidimensional-scholarly-archive/.

Anderson, L.; Byers, M.; Warner, A. (2021). "Introduction: Poetry, Theory, Archives". Anderson, L.; Byers, M.; Warner, A. (eds), *The Contemporary Poetry Archive: Essays and Interventions*. Edinburgh: Edinburgh University Press, 1-24.

Bellemin-Noël, J. (1972). *Le Texte Et l'Avant-Texte: Les Brouillons d'Un Poème De Milosz*. Paris: Larousse.

Berners-Lee, T.; Hendler, J.; Lassila, O. (2001). "The Semantic Web: A New Form of Web Content that is Meaningful to Computers Will Unleash a Revolution of New Possibilities". *Scientific American*, 284, 1-5. https://www.researchgate.net/publication/225070375_The_Semantic_Web_A_New_Form_of_Web_Content_That_is_Meaningful_to_Computers_Will_Unleash_a_Revolution_of_New_Possibilities.

Berners-Lee, T. (2006). "Linked Data". www.w3.org/DesignIssues/LinkedData.html.

Birnbaum, D.J.; Dekker, R.H. (2024). "Visualizing Textual Collation: Exploring Structured Representations of Textual Alignment". *Proceedings of Balisage: The Markup Conference 2024*. Balisage Series on Markup Technologies, vol. 29. https://doi.org/10.4242/BalisageVol29.Birnbaum01.

Bleeker, E.; Spadini, E.; Nava, B.; Oostveen, B.; Dekker, R.H. (2025). "'Here's strangeness.' A Collaborative Approach to Visualising Textual Variation". https://doi.org/10.5281/ZENODO.15387538.

Boie, B. (1993). "L'écrivain et ses manuscrits". Hay, L. (ed.), *Les Manuscrits des Écrivains*. Paris: Hachette, 34-53.

Brown, S.; Simpson, J. (2015). "An Entity by Any Other Name: Linked Open Data as a Basis for a Decentred, Dynamic Scholarly Publishing Ecology". *Scholarly and Research Communication*, 6(2). http://src-online.ca/index.php/src/article/view/212/409.

Bunn, J.; Rayner, S.J. (2019). "Observing the Author-editor Relationship: Recordkeeping and Literary Scholarship in Dialogue". *Archives and Manuscripts*, 47-3, 359-73. https://doi.org/10.1080/01576895.2019.1609363.

Capadisli, S. (2016). "Where is Web Science? From 404 to 200". https://csarven.ca/web-science-from-404-to-200.

Caswell, M. (2016). "'The Archive' Is Not an Archives: On Acknowledging the Intellectual Contributions of Archival Studies". *Reconstruction: Studies in Contemporary Culture*, 16(1). https://escholarship.org/uc/item/7bn4v1fk.

Christen, A.; Spadini, E. (2019) "Modeling Genetic Networks: Gustave Roud's Oeuvre, from Diary to Poetry Collections". *Umanistica Digitale*, 7, 77-104. https://doi.org/10.53681/c1514225187514391s.31.176.

Daquino, M.; Dello Buono, M.; Giovannetti, F.; Tomasi, F. (2020). *Paolo Bufalini: Appunti*. https://projects.dharc.unibo.it/bufalini-notebook/introduction.

De Biasi, P-M. (2004). "Toward a Science of Literature: Manuscript Analysis". Deppman, J.; Ferrer, D.; Groden, M. (eds), *Genetic Criticism: Texts and Avant-textes*. Philadelphia, PA: University of Pennsylvania Press, 36-68.

Dedecker, R.; Slabbinck, W.; Wright, J.; Hochstenbach, P.; Colpaert, P.; Verborgh, R. (2022). "What's in a Pod? A Knowledge Graph Interpretation for the Solid Ecosystem". Saleem, M.; Ngonga Ngomo, A.-C. (eds), *Proceedings of the 6th*

Workshop on Storing, Querying and Benchmarking Knowledge Graphs, 81-96. https://solidlabresearch.github.io/WhatsInAPod/.

Dillen, W. (2019). "On Edited Archives and Archived Editions". *International Journal of Digital Humanities*, 1, 263-77. https://doi.org/10.1007/s42803-019-00018-4.

Dillen, W.; Neyt, V. (2016). "Digital Scholarly Editing Within the Boundaries of Copyright Restrictions". *Digital Scholarship in the Humanities*, 31-4. https://doi.org/10.1093/llc/fqw011.

Drucker, J.; Nowviskie, B. (2004). "Speculative Computing: Aesthetic Provocations in Humanities Computing". Schreibman, S.; Siemens, R.; Unsworth, J. (eds), *A Companion to Digital Humanities*. Oxford: Blackwell Publishing Professional, 431-47. https://doi.org/10.1002/9780470999875.ch29.

Eggert, P. (2005). "Text-Encoding, Theories of the Text, and the 'Work-Site'". *Literary & Linguistic Computing*, 20(4). http://doi.org/10.1093/llc/fqi050.

Eggert, P. (2019). "The Archival Impulse and the Editorial Impulse". *Variants*, 14, 3-22. http://doi.org/10.4000/variants.570.

Elli, T.; Benedetti, A.; Pallacci, V.; Spadini, E.; Mauri, M. (2023). "Designing Network Visualizations for Genetic Literary Criticism". *Convergências*, 16(31). https://doi.org/10.53681/c1514225187514391s.31.176.

Folsom, E. (2007). "Database as Genre: The Epic Transformation of Archives". *PMLA*, 122(5), 1571-9. https://doi.org/10.1632/pmla.2007.122.5.1571.

Fraistat, N.; Viglianti, R.; Denlinger, E.C.; (dir.) (2013-present). *The Shelley-Godwin Archive*. http://shelleygodwinarchive.org.

Gooding, P.; Smith, J.; Mann, J. (2019). "The Forensic Imagination: Interdisciplinary Approaches to Tracing Creativity in Writers' Born-digital Archives". *Archives and Manuscripts*, 47-3, 374-90. https://doi.org/10.1080/01576895.2019.1608837.

Giovannetti, F. (2019). *The Critical Apparatus Ontology (CAO)*. Version: 0.9. https://w3id.org/cao.

Gonçalves, M.A. (2004). *Streams, Structures, Spaces, Scenarios, and Societies (5S): A Formal Digital Library Framework and Its Applications*. Blacksburg: Faculty of the Virginia Polytechnic Institute and State University. https://www.academia.edu/54086046/Streams_structures_spaces_scenarios_and_societies_5S_A_formal_digital_library_framework_and_its_applications.

Grésillon, A. [1994] (2016). *Éléments de Critique Génétique*. Paris: CNRS Éditions.

Jaillant, L. (2019). "After the Digital Revolution: Working with Emails and Born-Digital Records in Literary and Publishers' Archives". *Archives and Manuscripts*, 47-3, 285-304. https://doi.org/10.1080/01576895.2019.1640555.

Jaquier, C.; Maggetti, D. (dir.) (2022). *Gustave Roud. Textes & Archives*. https://roud.unil.ch.

Lopes, F. (2007). "Como se trabalha no Arquivo de Cultura Portuguesa Contemporânea". *As Mãos da Escrita: 25 anos do Arquivo de Cultura Portuguesa Contemporânea*. Lisboa: Biblioteca Nacional de Portugal, 51-74. https://purl.pt/13858/1/abertura/como-trabalha-acpc.html.

Martignano, C. (2023). *Critical Edition Ontology (CEO)*. Version: 1.0. http://purl.org/critical-edition-ontology.

McGann, J. (2014). *A New Republic of Letters: Memory and Scholarship in the Age of Digital Reproduction*. Cambridge: Harvard University Press.

Meschini, F. (2020). *Oltre il Libro: Forme di Testualità e Digital Humanities*. Milano: Editrice Bibliografica.

Molloy, K. (2019). "Literary Archives in the Digital Age: Issues and Encounters with Australian Writers". *Archives and Manuscripts*, 47-3, 327-42. https://doi.org/10.1080/01576895.2019.1631863.

Muller, S; Feith, J.A.; Fruin, R. (2003). *Manual for the Arrangement and Description of Archives*. Transl. by A.H. Leavitt. 2nd ed. Chicago: Society of American Archivists.

Pereira, E. (2023). "Authors' Rights vs. Textual Scholarship: A Portuguese Overview". *Journal of Intellectual Property, Information Technology and E-Commerce Law*, 14-4, 510-24. https://www.jipitec.eu/jipitec/article/view/19.

Portela, M.; Silva, A.R. (dir.) (2017-present). *LdoD Archive: Collaborative Digital Archive of the Book of Disquiet*. https://ldod.uc.pt/.

Portela, M. (2022). *Literary Simulation and the Digital Humanities: Reading, Editing, Writing*. New York: Bloomsbury.

Price, K.M. (2008). "Electronic Scholarly Editions". Siemens, R.; Schriebman, S. (eds), *A Companion to Digital Literary Studies*. Oxford: Blackwell. https://companions.digitalhumanities.org/DLS/?chapter=content/9781405148641_chapter_24.html.

Ramsay, S.; Rockwell, G. (2012). "Developing Things: Notes toward an Epistemology of Building in the Digital Humanities". Gold, M.K. (ed.), *Debates in the Digital Humanities*. Minneapolis: Minnesota Scholarship Online, 75-84. https://doi.org/10.5749/minnesota/9780816677948.003.0010.

Shillingsburg, P. (2006). *From Gutenberg to Google: Electronic Representations of Literary Texts*. Cambridge: Cambridge University Press.

Silva, A.R.; Portela, M. (2015). "TEI4LdoD: Textual Encoding and Social Editing in Web 2.0 Environments". *Journal of the Text Encoding Initiative*, 8. https://doi.org/10.4000/jtei.1171.

Spadini, E. (2023). *GENO, the Genetic Networks Ontology*. Version: 1.0. https://w3id.org/geno.

Spadini, E.; Tomasi, F.; Vogeler, G. (eds) (2021). *Graph Data-Models and Semantic Web Technologies in Scholarly Digital Editing*. Norderstedt: Books on Demand.

Stead, L. (2016). "Introduction". Smith, C.; Stead, L. (eds), *The Boundaries of the Literary Archive: Reclamation and Representation*. London; New York: Routledge, 1-13.

Stoyanova, S. (2023). "Articulating Intra- and Intertextual Relationships in the Fragment Collection. Working with the Digital Edition of Giacomo Leopardi's Zibaldone". *magazén*, 4(1), 13-42. https://doi.org/10.30687/mag/2724-3923/2023/01/001.

Sutton, D.C. (2014). "The Destinies of Literary Manuscripts, Past, Present and Future". *Archives and Manuscripts*, 42(3), 295-300. https://doi.org/10.1080/01576895.2014.948559.

Sutton, D.C. (2018). "Introduction: Literary Papers as the most 'Diasporic' of all Archives". Sutton, D.C., Livingstone, A. (eds), *The Future of Literary Archives: Diasporic and Dispersed Collections at Risk*. ARC Humanities Press.

Theimer, K. (2012). "Archives in Context and as Context". *Journal of Digital Humanities*, 1(2). https://journalofdigitalhumanities.org/1-2/archives-in-context-and-as-context-by-kate-theimer/.

Theimer, K. (2014). "The Role of Archives in a Digital Society: Now is What Matters". *Archivaria: The Journal of the Association of Canadian Archivists*, 78, 145-7. https://archivaria.ca/index.php/archivaria/article/view/13498.

Theys, T.; Mechant, P.; Maes, M.; Bourgeus, A.; Saldien, J.; De Marez, L. (2025). "Solid Pods: A Promising Approach to Enhance Users' Perception of Data Transparency and Control". *Interacting with Computers*, 37-6, 504-17. https://doi.org/10.1093/iwc/iwaf017.

Tomasi, F. (2022). *Organizzare la Conoscenza: Digital Humanities e Web semantico*. Milano: Editrice Bibliografica.

Van Hulle, D.; Nixon, M. (2021) "Editorial Principles and Practice". *Samuel Beckett Digital Manuscript Project*. https://www.beckettarchive.org/editorial.

Van Hulle, D.; Nixon, M. (dir.) (2011-present). *Samuel Beckett Digital Manuscript Project*. https://www.beckettarchive.org.

Van Hulle, D.; Shillingsburg, P. (2015). "Orientations to Text, Revisited". *Studies in Bibliography*, 37, 27-44. https://xtf.lib.virginia.edu/xtf/view?docId=StudiesInBiblio/uvaBook/tei/sibv059.xml;chunk.id=d25715e2576;toc.depth=1;toc.id=d25715e2576;brand=default.

Van Mierlo, W. (2016). "The Archaeology of the Manuscript: Towards Modern Palaeography". Smith, C.; Stead, L. (eds), *The Boundaries of the Literary Archive: Reclamation and Representation*. London; New York: Routledge, 15-29.

e-ISSN 2724-3923

magazén
Vol. 6 — Num. 2 — December 2025

Digital Epigraphy and the Study of Ancient Slavery

Kostas Vlassopoulos
University of Crete, Greece

Kyriaki Konstantinidou
Institute for Mediterranean Studies, Greece

Abstract The digitisation of the overwhelming majority of ancient evidence has made possible the emergence of Big Data and their utilisation by projects which concern the actions of millions of people. SLaVEgents represents the first large-scale project combining digital humanities, big data and history from below in order to explore the agency of enslaved persons in antiquity. It is building an open-access, interlinked digital prosopography that will provide a single point of entry for the study of all ancient slaves, freed persons and possible slaves attested between 1000 BCE-300 CE. Based on and documenting sources across multiple ancient languages, *SLaVEgents* researches the multiple identities of enslaved persons; the networks and communities that they created or participated in and the ways in which slave agency led to major political, social, economic and cultural changes in antiquity. This article offers an overview of the digital epigraphy of ancient slavery made possible by *SLaVEgents* and the surprising patterns that emerge from the collection of the evidence in regards to the distribution of manumission inscriptions, slave epitaphs and dedications, and occupational references.

Keywords Agency. Big data. Biography. Digital humanities. Epigraphy. Networks. Prosopography. Slavery

Summary 1 Digital Classics and Ancient History. – 2 *SLaVEgents*: A New Approach to Ancient Slavery. – 3 Methodology and Sources. –4 Ontology and Workflow. – 5 The Digital Epigraphy of Ancient Slavery.

Research for this article was supported by the project *SlaVEgents: enslaved persons in the making of societies and cultures in Western Eurasia and North Africa, 1000 BCE-300 CE*, funded by an Advanced Grant of the European Research Council (Grant Agreement no. 101095823).

Edizioni Ca'Foscari

Peer review

Submitted 2025-09-29
Accepted 2025-12-18
Published 2026-12-14

Open access

Citation Vlassopoulos, K.; Konstantinidou, K. (2025). "Digital Epigraphy and the Study of Ancient Slavery". *magazén*, 6(2), 195-214.

DOI 10.30687/mag/2724-3923/2025/02/004

1 Digital Classics and Ancient History

Classics is among the earliest disciplines in the Humanities to engage extensively with the digital revolution that emerged in the 1970s and 1980s (Bagnall, Heath 2018; Christensen 2022). As a result of the forward-thinking of some important pioneers, effectively every single Greek and Latin literary text now exists in one or more digital formats; the same largely applies to Greek and Latin papyri and ostraca. In the case of inscriptions, probably 90-95% of Latin inscriptions have been digitised, while the equivalent rate for Greek inscriptions is probably around 80%; similar percentages apply to ancient coins. It is only in the case of archaeological evidence apart from inscriptions and coins that digitisation lags substantially behind all other forms of ancient sources.

This large-scale digitisation makes possible the emergence of Big Data projects. Despite the constant complaint of ancient historians about the paucity of evidence, the actual reality is that the scale of the available evidence has long overgrown the capacity of any individual living scholar. A huge amount of pertinent evidence is known only to a few specialists of particular times and places; our conceptual models and general narratives tend to focus on certain well-known corpora and largely ignore the majority of the existing evidence, while the specialist work on particular pieces of evidence rarely tries or succeeds to build wider models and narratives on their basis. Thus, the digitisation of ancient evidence and the use of modern technological tools, like digital annotation, tagging and Social Network Analysis, open up the possibility of actually exploiting the Big Data of ancient evidence in ways which have been impossible with traditional scholarly methods.

At the same time, digitisation is particularly important for certain approaches to ancient history. Ever since its emergence in antiquity, historiography has overwhelmingly adopted a top-down perspective, focused on elites and the state apparatuses they controlled. It was only in the 1960s that history from below emerged as a major alternative, with the pioneering work of scholars like Eric Hobsbawm, E.P. Thompson and Eugene Genovese. While history from below has had a major impact on medieval, early modern and modern history, it was largely shunned by ancient historians. Nevertheless, over the last few years history from below has finally started to have a significant impact among ancient historians (Courrier, Magalhães de Oliveira 2021; Gartland, Tandy 2024). History from above can be based on the biographies of relatively limited numbers of eminent people provided by ancient literary sources, or the detailed descriptions of the *cursus honorum* of elite men provided by inscriptions. History from below can only rarely be based on such sources; and given the fact that it focuses on the lives of millions of ordinary people captured only

fragmentarily in the existing sources, any systematic study of ancient history from below must be based on different methods, which require the employment of masses of evidence. It is precisely at this point that the digitisation of ancient sources, Big Data projects and history from below can join hands and mutually benefit from the collaboration.

This article aims to present a large-scale digital project titled *SLaVEgents: enslaved persons in the making of societies and cultures in Western Eurasia and North Africa, 1000 BCE-300 CE*. Funded by an Advance Grant of the European Research Council, the 25-strong international team of the project aims to take advantage of the digitisation of ancient sources and the emergent Big Data this generates in order to make a major contribution to the study of history from below in antiquity by transforming the study of ancient slavery and enslaved persons and consequently the very study of ancient history.[1] The article also shows how digital *SLaVEgents* will influence the study of specific fields in ancient history, namely the epigraphy of ancient slavery, by presenting some surprising patterns that emerge from the collection of evidence.

2 *SLaVEgents*: A New Approach to Ancient Slavery

Slavery was an ever-present feature of ancient societies to the extent that numerous studies have explored its implications for writing the history of those societies (Schumacher 2001; Andreau, Descat 2006; Hunt 2018). Traditional approaches to the topic have overwhelmingly adopted a top-down perspective, in which slavery is seen as unilaterally determined by the masters (Finley 1980; Bradley, Cartledge 2011). Over the last decade, this status quo has come under increasing challenge, as studies from different theoretical traditions have started to complement the study of *what happened to* ancient slaves with the exploration of *what slaves did* (Vlassopoulos 2021). Building on these developments, *SLaVEgents* represents the first large-scale digital project to focus on the agency of enslaved persons and to explore how they actively shaped the ancient societies in which they lived. Slave agency (Johnson 2003; Schiel et al. 2017) consists of the strategies and actions of enslaved persons, shaped by the roles created for slaves by their masters and other slaving actors, as well as by the identities, networks and communities that slaves created for themselves.

To achieve in-depth analysis of slave agency, *SLaVEgents* is building a digital prosopography that will transform the study and understanding of ancient slavery across the board. This open-access,

1 See the project's webpage: `https://www.ims.forth.gr/en/project/view?id=272`.

interlinked prosopography will provide a single point of entry for the study of all slaves, freed persons and possible slaves attested between 1000 BCE and 300 CE from Mesopotamia to the Atlantic. *SLaVEgents* not only collects the names of all known enslaved persons from antiquity, but also identifies other pertinent factors, such as biographical information (masters, family and kinship, ethnicity, recorded activities, known associates). Its research objectives focus on identifying, tracing and investigating the multiple identities of enslaved persons (Vlassopoulos 2022); the networks and communities that they created or participated in (Taylor, Vlassopoulos 2015); and the ways in which slave agency led to major political, social, economic and cultural changes in antiquity (Vlassopoulos 2026).

In contrast to most existing digital prosopographies, which are effectively limited to providing lists of names accompanied by source references,[2] *SLaVEgents*' digital prosopography includes all relevant sources in the original ancient languages (Aramaic, Assyrian, Babylonian, Hebrew, Egyptian, Greek, Latin, Phoenician) and in modern English translation. In addition, it also records the relevant archaeological data, by offering links to online collections of archaeological materials, or references to printed sources. In this way, *SLaVEgents* creates the evidentiary foundation for innumerable future Big Data projects. At the same time, the open-access form of the database and the translation of the sources in English will expand massively the availability and accessibility of this mass of evidence to people without access to restricted resources and without the linguistic skills to understand all the various ancient languages.

3 Methodology and Sources

SLaVEgents is based on a wide range of sources, many of which have never been used for the study of slavery before. It draws upon published evidence from all kinds of sources: documentary (inscriptions, ostraca, papyri, curse tablets, letters, registers, contracts); legal (court records, juristic texts, law collections), and literary, both fictional (drama, novels, poetry) and non-fictional (historiography, biography, oratory, epistolography, philosophy, medicine, astrology, patristic texts); it also collects archaeological evidence attributed to individual ancient slaves (tombstones, votive reliefs, artefacts). The identification and collection of the relevant evidence is one of the major aims of *SLaVEgents*, not least because

2 E.g. *The Lexicon of Greek Personal Names*: https://www.lgpn.ox.ac.uk/home; *The Digital Prosopography of the Roman Republic*: https://romanrepublic.ac.uk/; *Prosobab*: https://prosobab.leidenuniv.nl/index.php.

e-ISSN 2724-3923

slave prosopographies for most ancient societies simply do not exist; currently there are only those for the cities of Athens and Rome (Fragiadakis 1988; Solin 1996). Most of the evidence for enslaved persons remains unidentified and scattered across all the kinds of primary sources mentioned above. The work of documenting those references utilises so far as it is possible open-access digital databases with large-scale collections of:

- literature (Perseus, `https://scaife.perseus.org/library/`)
- epigraphy (PHI, `https://epigraphy.packhum.org/allregions`; EDCS, `http://db.edcs.eu/epigr/epi.php`; EDR, `http://www.edr-edr.it/default/index.php`)
- papyrology (papyri.info, `https://papyri.info/`)
- documentary sources (CDLI, `https://cdli.ucla.edu/`).

Where necessary, these materials are supplemented by restricted-access digital collections (such as the TLG, `http://stephanus.tlg.uci.edu/`) and printed publications of original sources.

Although prescriptive sources give the impression that there was a clear dividing line separating slave from free in ancient societies, in reality it is often very difficult to establish the status of the individuals attested in our sources; this partly results from the descriptive vocabulary of the sources, which often uses categories which are vague or not specifically related to slaves (Zelnick-Abramovitz 2018). *SLaVEgents* does not explain away this complexity and ambiguity, but puts it at the centre of our attention; it aims to make a major contribution towards the systematic study of the vocabulary of slavery and the identification of criteria for distinguishing the status of individuals, as well as to explore the historical reasons for this complexity and ambiguity.

SLaVEgents draws on over a decade of work that has aimed at determining guidelines for a linked data ontology for historical prosopographies. Emerging out of the pioneering work of the Lexicon of Greek Personal Names (`https://www.lgpn.ox.ac.uk/`), which continues to collect and publish with documentation all known ancient Greek personal names, came, in 2014, the Standards for Networking Ancient Prosopographies: Data and Relations in Greco-Roman Names (SNAP:DRGN) project. Largely inspired by the Pelagios linked data initiative, which connects online resources through references to place (Vitale et al. 2021), SNAP:DRGN has sought to formulate a comparable method for linking people. Taking a pragmatic approach to the absence of any widely accepted database format or even print-based approach to the representation of ancient prosopographical information (let alone a standard linked data format), SNAP:DRGN has published a set of guidelines for representing core person disambiguation data in linked data RDF (Bodard et al. 2017). As explained below, the digital prosopography of *SLaVEgents* is based on these guidelines.

4 Ontology and Workflow

The database itself uses Nodegoat (`https://nodegoat.net/about`), a humanities' web-based research and data-visualisations environment. By being rooted in the world of the humanities, Nodegoat offers rich flexibility in the creation and development of a data structure for representing any given content; equally, however, it allows for data export in standard formats, which facilitates data sharing beyond the single project for reuse among the wider research and learning communities (van Bree, Kessels 2013). Borrowing from actor-network theory, Nodegoat treats people, networks, and sources as equal 'objects', offering powerful relational, spatial and temporal analysis and visualisation. This object-centred approach aligns well with *SLaVEgents*' focus on documenting the multiple identities of enslaved persons in a flat, non-hierarchical structure, based on the various ways in which they conceptualised their classification as slaves and their entanglement with a range of other identities that were partly related to slavery ('objects' such as work and function) and partly independent from it ('objects' such as gender, family, kinship, ethnicity, religion).

As a reflection of their importance, inputting data starts not with the enslaved person, but with the source material: all work stems from the primary sources themselves. Data entry generally takes the following two steps.

First, the *SLaVEgents* researcher navigates to the Source tab. After adding the source, they then work through a series of fields **[fig. 1]**:

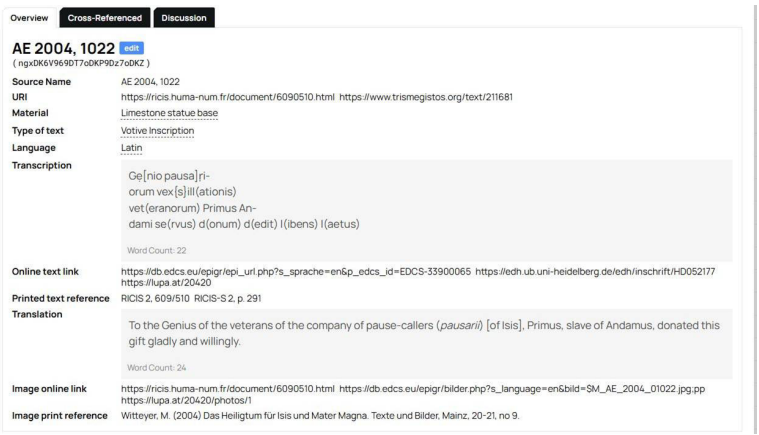


Figure 1 Epigraphic sources

e-ISSN 2724-3923

- source name
- URI
- material > e.g. stone altar, literary text, wax tablet, ceramic vase, papyrus, etc.
- type of text > e.g. legal, letter, philosophy, graffiti, epitaph, etc.
- language > e.g. Aramaic, Hebrew, Latin (including bi-lingual, tri-lingual for epigraphic evidence)
- transcription (the original text)
- online text link (sometimes = URI)
- printed text reference (if the text is not digitised)
- translation (English)
- translation online link
- translation print reference (if there is no online translation)
- online image link (e.g. reliefs, vases)
- image print reference (if there is no online image)

The Source tab divides the database according to ancient languages and to the material of the source and its genre. This ensures adequate attention is given to particular kinds of sources and issues that arise from them. Each source is also linked to other open-access databases of ancient documents where possible. All primary sources are available in English, thereby providing access to a wider audience and enhancing the comparative study of slavery across different ancient societies.

After the primary Source information is filled in, the *SLaVEgents* researcher navigates to the Network tab and works through a second set of fields **[fig. 2]**:

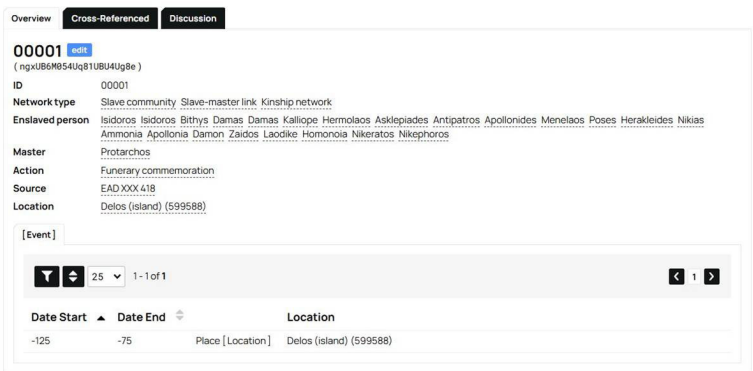


Figure 2 Networks

- ID (given by system)
- network type > e.g. slave group, slave-master link, kinship network, work community

- enslaved person > existing (type name) or new: opens 'enslaved person' category from object (see below)
- master > ditto (new master category: name, family name, master URI, identical with enslaved person)
- third party > ditto
- slave group
- action (e.g. what's the action in which the enslaved person is involved, e.g. work, sale, sexual liaison, punishment, theft, etc.)
- source (connecting network to the source or sources)
- cult (documenting the participation of members of the network in religious activities)
- then: subobjects > add event: period and location (when and where the incident takes place)

Network types are an important aspect of the project's investigation into slave agency, and in this respect alone Nodegoat delivers on its value as a network-based research environment. Built on the different persons involved in an action in which an enslaved person participates in each source, the Network tab shows all the social networks and communities that slaves created on the basis of their various roles and identities. By virtue of the Network tab, examination is not limited to the broader groups and communities to which a slave belonged and acted or the vocabulary that is related to and used for the slaves; it is also possible to explore the great variety of activities (through the 'action' option) in which the slaves were involved, as well as the similarities and differences in all the above domains over time.

As already mentioned, *SLaVEgents* models the object 'enslaved person' in ways that build on the SNAP:DRGN recommendations; equally, the researcher can also take into consideration particular features that relate to the figure of the enslaved person **[fig. 3]**. Each enslaved person has:

e-ISSN 2724-3923

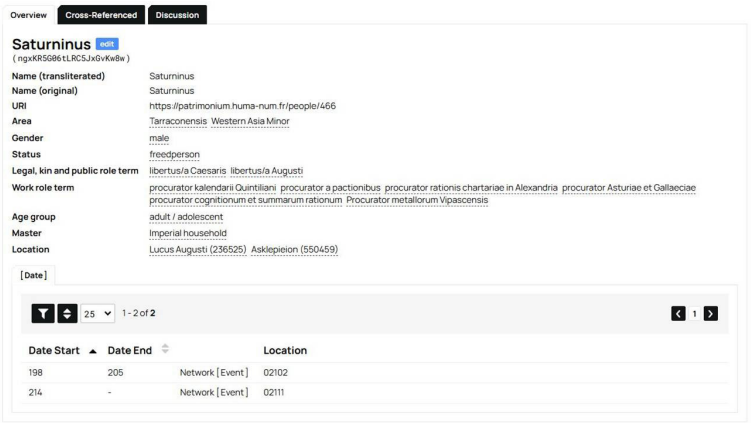


Figure 3 Enslaved persons entry

- a canonical URI for publication, type (enslaved person), and citation;
- names (both transliterated and in the original);
- area (associated place of origin), time period (associated date), and other external URIs.

Additionally, the enslaved person object has the following categories: gender; status; legal, kin and public role term; work role term; age group; specific age in years; price; fictional or real status; and, finally, associated manumission conditions.

5 The Digital Epigraphy of Ancient Slavery

SLaVEgents' digital prosopography currently includes 28,000 enslaved and freed persons, 15,000 masters and 12,000 free third parties. These individuals are recorded in 19,000 sources, 14,000 of which are Greek and Latin inscriptions, thus illustrating the fundamental role of epigraphy in our database. Our projection is that, when finally completed, the prosopography will include upwards of 50,000 enslaved and freed persons and an equivalent number of masters and free third parties, recorded in upwards of 35,000 sources. These numbers demonstrate how *SLaVEgents* combines digital humanities, big data and history from below. Digital humanities provide a number of tools like digital annotation, tagging, and social network analysis in order to make the data amenable to discovery, processing, and quantitative and qualitative interpretation. Big data offer the opportunity to move beyond normative and structuralist models of ancient societies and study relations and interactions distributed across space and time.

Finally, the evidentiary foundation of *SLaVEgents* is quintessential for studying the agency of millions of subaltern people and tracing its conjunctural and cumulative historical consequences.

In this respect, it is important to point out two important contributions of digital *SLaVEgents* to the study of ancient history. The first concerns our insistence on creating Linked Open Data, rather than just another self-enclosed database (Middle 2024). Our digital prosopography includes systematic interlinking with all relevant digital databases of ancient Open Data: collections of literary, epigraphic, papyrological, numismatic and archaeological sources; prosopographies; encyclopaedias; and gazetteers of ancient settlements. This is a crucial step for opening up the study of ancient slavery and enslaved persons to the study of all other aspects of the ancient world. To give one example, all inscriptions recorded in our digital prosopography are linked to their relevant URI in Trismegistos. Through Trismegistos, the user can find references to most printed or digital editions of the relevant inscription; at the same time, Trismegistos includes digital tagging of the place at which each inscription has been found, while also listing all other inscriptions that have been found at the same place. As a result, the interlinking of our digital prosopography with Trismegistos makes possible the study of a particular inscription mentioning enslaved persons alongside the complete epigraphic output of the place involved; it will thus facilitate the study of local epigraphic habits and their patterns, a crucial issue, as the discussion below shows (Nawotka 2020). It will also enable the study of enslaved persons alongside the totality of the recorded local population and the study of slavery alongside all other institutions and practices recorded in the local epigraphic evidence. Our digital prosopography aims precisely to break the conceptual apartheid within which slavery studies in antiquity have been largely pursued and to open up a way in which it can have an impact on the study of all other aspects of ancient history.

The second digital contribution concerns changing the experience of how to conduct research in ancient history. Our digital prosopography is shaped by the parameters of space, time and interaction. By using the digital work of *Pleiades*, it is possible to locate enslaved persons, masters and third parties on a map, which also includes temporal co-ordinates **[fig. 4]**.

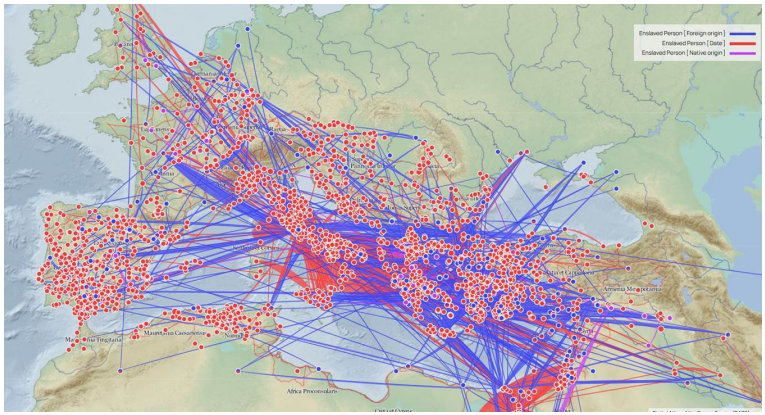


Figure 4 Enslaved and freed persons, 1000 BCE -300 CE

Users can select which settlements or regions they want to be depicted on the map, as well as the temporal duration that is of interest to them. As a result, users will be able to find within seconds in a visual form the answer to questions like 'in which places are manumission inscriptions recorded', 'how many and which enslaved persons are attested in Larissa between 100 BCE and 150 CE', 'in which places are enslaved and freed persons belonging to Roman soldiers attested', or 'how many and which enslaved persons are attested across the ancient world between 500-200 BCE? At the same time, the incorporation of the tools of Social Network Analysis in our digital prosopography makes possible the visualisation of the various networks involving slaves, masters and third parties and their complexity; the social network of imperial slaves and freed persons is a telling example **[fig. 5]**.

It is a radically different experience of approaching the material than the printed text of ancient sources or modern scholarly literature that still accounts for the vast majority of scholarly work.

We would like to illustrate these features of the project by tracing a number of patterns that are already emerging from the collection of data, their digital processing that we described in the previous sections, alongside the digital mapping of the evidence in spatial and temporal terms. These patterns are often highly surprising, and they raise important methodological questions that we need to discuss in order to be able to interpret historically the relevant data. Given the overwhelming preponderance of Greek and Latin inscriptions among our collected evidence, we shall focus here on digital epigraphy and the various epigraphic habits associated with enslaved and freed persons.

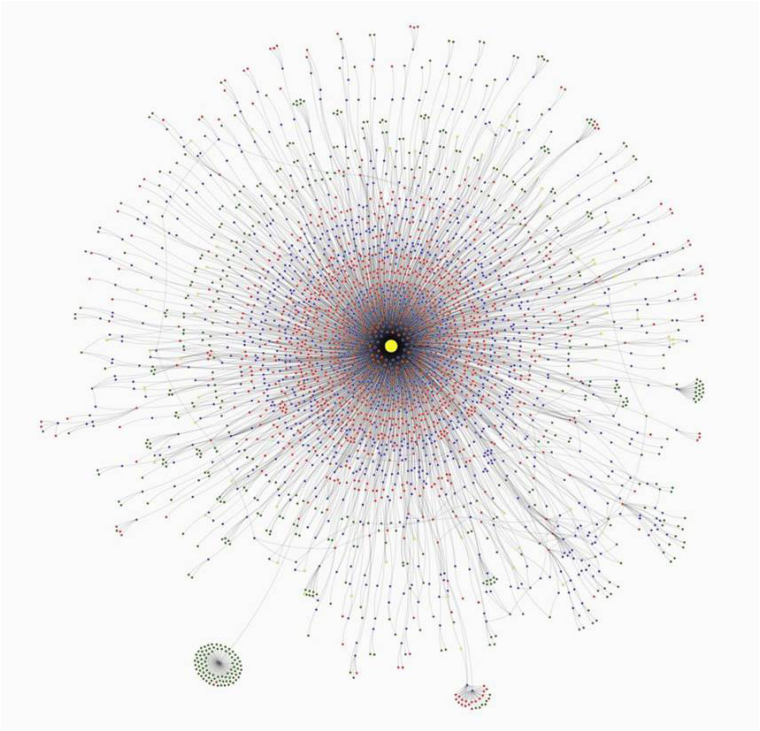


Figure 5 The social networks of imperial slaves and freedpersons

Our first example illustrates how digital mapping can radically change the interpretation of even well-known sources. Manumission inscriptions constitute the most abundant source of evidence for Greek freed persons (Vlassopoulos 2019). It is normally assumed that the purpose of manumission inscriptions was to achieve the widest possible publicity for the act of manumission and thus to safeguard freed persons from seizure and re-enslavement. Manumissions were always witnessed so that in the future there would be persons capable of verifying the status of the liberated slave; by inscribing the manumission record in publicly accessible places, like temples and agoras, knowledge of the manumission would be continuously publicized to a much greater audience than the few witnesses of the act. The theory sounds plausible, until we examine which Greek communities developed the habit of inscribing manumission acts **[fig. 6]**.

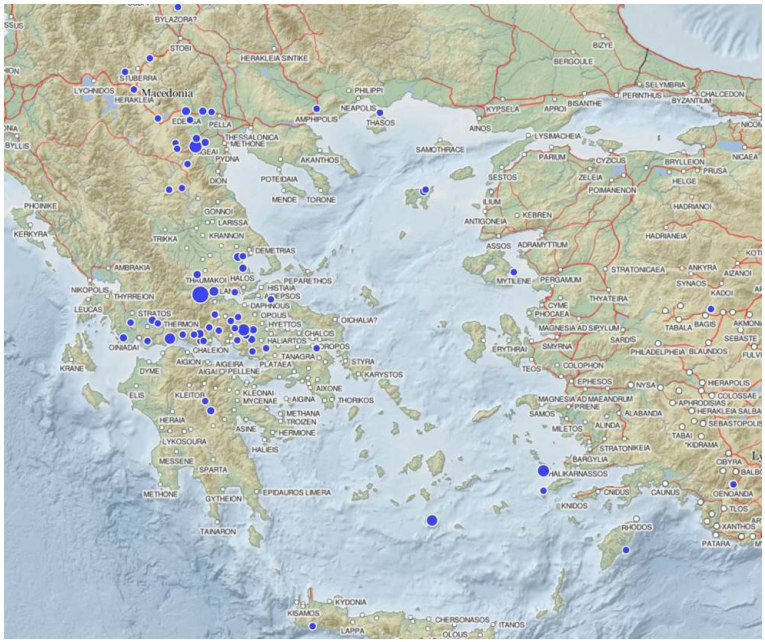


Figure 6 Manumission inscriptions in the Greek world

The vast majority of manumission inscriptions occur in central and northern Greece; there are very few manumission inscriptions from the Peloponnese, the Aegean islands and Asia Minor. Furthermore, one would have expected that most manumission inscriptions would be erected in large urban communities, where people would not know each other, and the need to publicize manumissions to a wider audience would be stronger. Surprisingly, the evidence points the other way round. We have no manumission inscriptions from large urban centres like Athens, Ephesus and Miletus, or large Aegean islands like Rhodes and Chios, where we know that thousands of slaves were employed. Instead, manumission inscriptions crop up in small island communities like Thera and Calymnos and relatively small rural communities, like Chyretiai and Leukopetra. The need to publicize manumission acts cannot therefore sufficiently account for manumission inscriptions; any account of manumission inscriptions must explain why they are overwhelmingly absent from large urban communities with strong and diversified epigraphic habits, where the problems of publicity would be particularly acute, and why they are present where they are. In other words, we need to understand the epigraphic habit of manumission, as well as the social dynamics of those communities that set up manumission inscriptions (Hewitt 2023).

The second example demonstrates another curious pattern of the epigraphic habit. If manumission inscriptions are reswtricted to certain communities, epitaphs and dedications constitute two epigraphic genres that were effectively universal across the eastern Mediterranean world. Given this, one would assume that the distribution of epitaphs and dedications that were erected by enslaved and freed persons would be determined by the size of ancient communities and the significance of slavery in them; the bigger the community and the number of slaves in it, the larger the number of epitaphs and dedications attested. But this assumption is highly misleading. There is a very wide dispersal of enslaved and freed epitaphs across Asia Minor and Macedonia, almost exclusively dating from the early imperial period; on the contrary, in mainland Greece there are very few epitaphs by enslaved and freed persons attested in any period **[fig. 7]**.

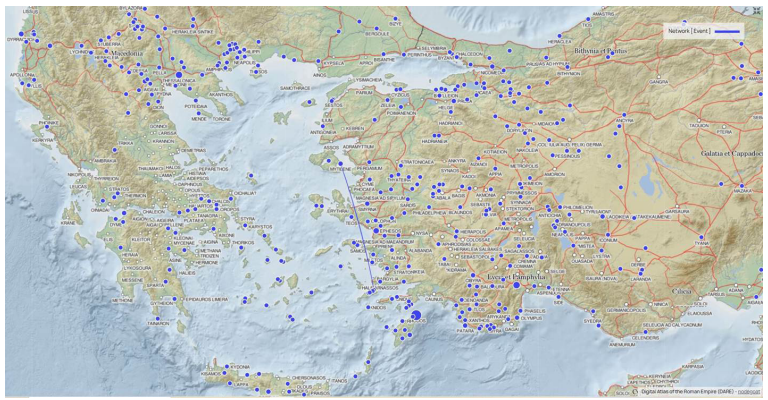


Figure 7 Epitaphs by enslaved and freed persons in the eastern Mediterranean

This pattern becomes even more pronounced when we examine dedications; with the exception of Delos, dedications by enslaved and freed persons are almost exclusively attested in Asia Minor and Macedonia **[fig. 8]**.

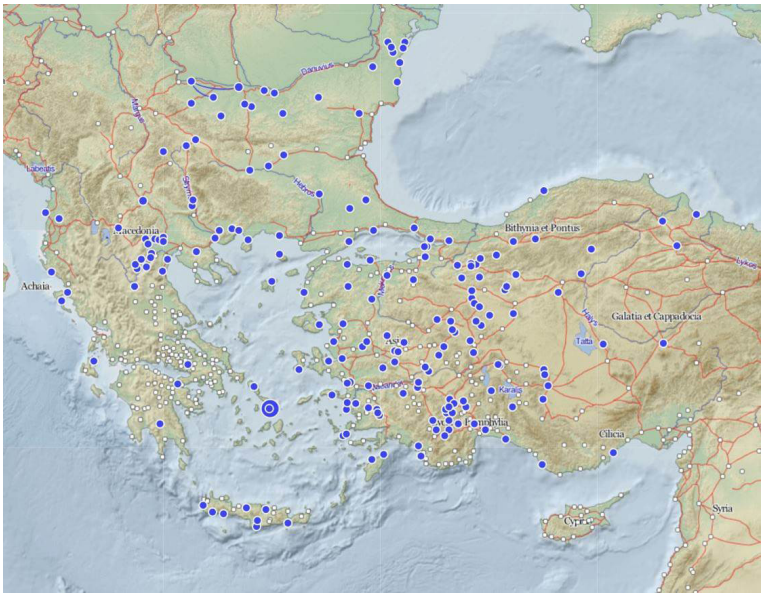


Figure 8 Dedications by slaves and freedpersons in the eastern Mediterranean

It is highly unlikely that sizeable slave populations only existed in Asia Minor and Macedonia (Vlassopoulos 2025); it is equally unlikely that enslaved and freed persons in mainland Greece did not erect epitaphs and dedications. What is more probable, is that enslaved and freed persons in mainland Greece chose not to advertise explicitly their legal status, and thus are invisible in the existing documents, while large numbers of enslaved and freed persons in Asia Minor and Macedonia made precisely the opposite choice. How should we explain these very divergent choices made by enslaved and freed people even during the same temporal period?

The third example concerns epigraphic attestations of the work identities of enslaved and freed persons (Joshel 1992; Tran 2013), and more specifically of the identities of estate managers and business agents (*institores*, *vilici* and *negotiatores* in Latin; *oikonomoi* and *pragmateutai* in Greek), recorded in epitaphs and dedications (Aubert 1994; Carlsen 1995). Adopting a Mediterranean-wide vista has some very surprising results **[fig. 9]**.

Figure 9 The epigraphic habit of *vilici* and *negotiatores*

Asia Minor shows again a very remarkable dispersal of evidence, accompanied by an equally significant number of attestations from the Danubian provinces. What is truly remarkable in this respect is the evidence from the Iberian provinces. Our digital prosopography includes over 3,000 enslaved and freed persons from the Iberian peninsula, which is one of the highest frequencies of attested slaves outside Italy; the equivalent number for the whole of Asia Minor is 2,000 enslaved and freed persons. Notwithstanding the high numbers from Iberia, it is fairly evident that the recording of occupational identities was very rarely adopted by enslaved and freed persons in Iberia. This clearly cannot be attributed to a supposed insignificant role of slaves and freed persons in the economic processes of Roman Iberia: the voluminous evidence of Iberian *instrumentum domesticum* leaves little doubt about the significance of enslaved and freed managers and business agents (Olesti Vila, Carreras Monfort 2013). Why did enslaved and freed managers and business agents in Iberia choose so rarely to record their occupational identity in epitaphs and dedications, and why did the same people in Asia Minor or the Danubian provinces make such a different choice? This is even more remarkable when we take into account the fact that recordings of occupational attestations in the Latin inscriptions of the Western Mediterranean are substantially more common than those in Greek inscriptions from the Eastern Mediterranean (Varga 2020).

Our final example concerns the epigraphic attestation of another occupational identity of enslaved and freed persons, that of gladiators. Although of course by the imperial period significant numbers of

gladiators were free, there is no doubt that slaves always constituted the most substantial group among the gladiatorial population. Thousands of Latin inscriptions from the Western Mediterranean concern the amphitheaters and the various games and activities that took place in them, prime among which were the gladiatorial shows (Sabbatini Tumolesi et al. 1988-2017).[3] Although the old idea that gladiatorial shows were shunned in the Greek-speaking Eastern Mediterranean has long been laid to rest by careful scholarly work (Robert 1940; Carter 1999), there is no doubt that the gladiatorial phenomenon had its origins in the Western Mediterranean and a very deep presence there. It would be natural to assume, accordingly, that epigraphic attestations of gladiators, usually in the form of epitaphs, would be primarily a Western Mediterranean phenomenon. But the opposite is rather the case; outside of Italy,[4] most of the epigraphic references to gladiators come from Greek funerary inscriptions from the Eastern Mediterranean **[fig. 10]**.[5]

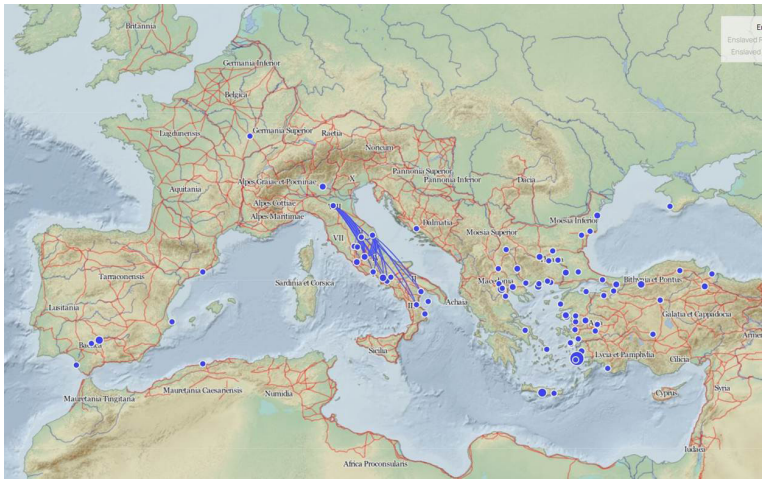


Figure 10 The epigraphic habit of gladiators

Why did enslaved and freed gladiators adopt the epigraphic habit of erecting epitaphs in the Eastern Mediterranean, but made very different choices in the Western Mediterranean?

3 See the digital database *Amphi-Theatrum*: `https://www.amphi-theatrum.de/home0.html`.

4 For the few tombstones of gladiators from Rome and Italy, see Hope 2000.

5 See the *Gladiators' Tombstones Database (GlaToDa)*: `http://www-v115.rz.uni-mannheim.de/index.php?page=home`.

The above examples have hopefully illustrated the substantial possibilities opened up by the digital epigraphy of ancient slavery offered by *SLaVEgents'* prosopography. The tools of digital humanities make possible the collection of big data on the agency of enslaved persons and its historical interpretation. For the first time it becomes possible to plot the evidence using spatial and temporal parameters, thus enabling the study of spatial diversity and temporal change. But these data are patterned by the diverse epigraphic habits of different groups and communities. The various patterns of epigraphic habits that we have discussed above raise fascinating questions about the historical agency of enslaved and freed persons and the various processes that lie behind them. The short space of this article forbids any detailed discussion; but we have hopefully convinced readers that the digital epigraphy of ancient slavery has a very bright future ahead.

Bibliography

Andreau, J.; Descat, R. (2006). *Esclave en Grèce et à Rome*. Paris: Hachette.

Aubert, J.-J. (1994). *Business Managers in Ancient Rome: A Social and Economic Study of Institores, 200 BC-AD 250*. Leiden: Brill.

Bagnall, R.S.; Heath, S. (2018). "Roman Studies and Digital Resources". *Journal of Roman Studies*, 108, 171-89. https://doi.org/1010.1017/S0075435818000874.

Bodard, G.; Cayless, H.; Depauw, M.; Isaksen, L.; Lawrence, F.; Rahtz, S. (2017). "Standards for Networking Ancient Person Data: Digital Approaches to Problems in Prosopographical Space". *Digital Classics Online*, 3(2), 28-43. http://dx.doi.org/10.11588/dco.2017.0.37975.

Bradley, K.; Cartledge, P. (eds) (2011). *The Cambridge World History of Slavery*. Vol. 1, *The Ancient Mediterranean World*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CHOL9780521840668.

Carlsen, J. (1995). *Vilici and Roman Estate Managers until AD 284*. Rome: L'Erma di Bretschneider.

Carter, M.D.J. (1999). *The Presentation of Gladiatorial Spectacles in the Greek East: Roman Culture and Greek Identity* [PhD Dissertation]. Hamilton, Ontario: McMaster University.

Christensen, J.P. (2022). "Digital Classics". *Transactions of the American Philological Association* 152.1, 43-54. https://dx.doi.org/10.1353/apa.2022.0005.

Courrier, C.; Magalhães de Oliveira, J.C. (eds) (2021). *Ancient History from Below: Subaltern Experiences and Actions in Context*. London; New York: Routledge.

Finley, M.I. (1980). *Ancient Slavery and Modern Ideology*. London: Chatto & Windus.

Fragiadakis, C. (1988). *Die attischen Sklavennamen von der spätarchaischen Epoche bis in die römische Kaiserzeit. Eine historische und soziologische Untersuchung*. Athens: n.p.

Gartland, S.D.; Tandy, D.W. (eds) (2024). *Voiceless, Invisible, and Countless in Ancient Greece: The Experience of Subordinates, 700-300 BCE*. Oxford: Oxford University Press. https://doi.org/10.1093/9780191995514.001.0001.

Hewitt, M. (2023). *Inscribing Manumission in the Hellenistic World* [PhD Dissertation]. Oxford: University of Oxford.

Hope, V. (2000). "Fighting for Identity: The Funerary Commemoration of Italian Gladiators". *Bulletin of the Institute of Classical Studies, Supplement* 73, 93-113. https://doi.org/10.1111/j.2041–5370.2000.tb01940.x.

Hunt, P. (2018). *Ancient Greek and Roman Slavery*. Malden (MA): Wiley Blackwell.

Johnson, W. (2003). "On Agency". *Journal of Social History*, 37, 113-24. https://doi.org/10.1353/jsh.2003.0143.

Joshel, S.R. (1992). *Work, Identity and Legal Status at Rome: A Study of the Occupational Inscriptions*. Norman (OK); London: University of Oklahoma Press.

Middle, S. (2024). "Linked Ancient World Data: Implementation, Advantages, and Barriers". *Digital Classics Online*, 10(1), 16-49. https://doi.org/10.11588/dco.2024.10.104105.

Olesti Vila, O.; Carreras Monfort, C. (2013). "Le paysage social de la production vitivinicole dans l'ager Barcinonensis: esclaves, affranchis et institores". *Dialogues d'Histoire Ancienne*, 392(2), 147-89. https://doi.org/10.3917/dha.392.0147.

Nawotka, K. (ed.) (2020). *Epigraphic Culture in the Eastern Mediterranean in Antiquity*. London; New York: Routledge.

Robert, L. (1940). *Les gladiateurs dans l'Orient grec*. Paris: Champion.

Schiel, J.; Schürch, I.; Steinbrecher, A. (2017). "Von Sklaven, Pferden und Hunden: Trialog über den Nutzen aktueller Agency-Debatten für die Sozialgeschichte". *Schweizerisches Jahrbuch für Wirtschafts-und Sozialgeschichte*, 32, 17-48. https://doi.org/10.5169/seals–842463.

Schumacher, L. (2001). *Sklaverei in der Antike: Alltag und Schicksal der Unfreien*. Munich: Beck.

Solin, H. (1996). *Die stadtrömischen Sklavennamen: ein Namenbuch, I-III*. Stuttgart: Franz Steiner Verlag.

Varga, R. (2020). *Carving a Professional Identity: The Occupational Epigraphy of the Roman Latin West*. Oxford: Archaeopress. https://digital.casalini.it/9781789694659.

Sabbatini Tumolesi, P. et al. (1988-2017). *Epigrafia anfiteatrale dell'occidente romano*, I-IX. Roma: Quasar.

Taylor, C.; Vlassopoulos, K. (eds) (2015). *Communities and Networks in the Ancient Greek World*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198726494.001.0001.

Tran, N. (2013). "Les statuts de travail des esclaves et des affranchis dans les grands ports du monde romain (Ier siècle av. J.-C. - IIe siècle apr. J.-C.)". *Annales. Histoire, Sciences Sociales* 68, 999-1025. https://doi.org/10.1017/S0395264900015080.

van Bree, P.; Kessels, G. (2013). "Nodegoat: A Web-Based Data Management, Network Analysis & Visualization Environment". http://nodegoat.net from LAB1100, http://lab1100.com.

Vitale, V.; de Soto, P.; Simon, R.; Barker, E.; Isaksen, L.; Kahn, R. (2021). "Pelagios-Connecting Histories of Place. Part I: Methods and Tools". *International Journal of Humanities and Arts Computing: Pelagios,* Special Issue**,** 15(1-2), 5-32.

Vlassopoulos, K. (2019). "The End of Enslavement, 'Greek Style'". Hodkinson, S.; Kleijwegt, M.; Vlassopoulos, K. (eds) *The Oxford Handbook of Greek and Roman Slaveries*. Oxford: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199575251.013.39

Vlassopoulos, K. (2021). *Historicizing Ancient Slavery*. Edinburgh: Edinburgh University Press. https://www.jstor.org/stable/10.3366/j.ctv1vtz81x

Vlassopoulos, K. (2022). *The Multiple Identities of Enslaved Persons in Antiquity*. Berlin: EB Verlag. https://doi.org/10.53179/9783868934304.

Vlassopoulos, K. (2025). "Greek Slave Systems in the Hellenistic and Early Imperial Periods". Hodkinson, S.; Kleijwegt, M.; Vlassopoulos, K. (eds) *The Oxford Handbook of Greek and Roman Slaveries*. Oxford: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199575251.013.43.

Vlassopoulos, K. (2026). "What is Slave Agency? Enslaved Persons and the Making of Ancient Societies". *Journal of Global Slavery*, 11, 1-31.

Zelnick-Abramovitz, R. (2018). "Greek and Roman Terminologies of Slavery". Hodkinson, S.; Kleijwegt, M.; Vlassopoulos, K. (eds) *The Oxford Handbook of Greek and Roman Slaveries*. Oxford: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199575251.013.41.

e-ISSN 2724-3923

magazén
Vol. 6 — Num. 2 — Dicembre 2025

LLM-Mining Pre-Stemmatological Philological Literature

Armin Hoenen
Goethe Universität Frankfurt am Main, Germany

Abstract The current article outlines a new research avenue for the analyses of literature from the time before the advent of the stemmatic method in the nineteenth century using large collections of digitized images and texts of historical philological works. The main aim is to understand the dynamics behind the processes leading to the invention of said method. The proposed steps are object recognition (image analysis) with textual clues and relation extraction (text mining). Proof-of-concept-level experiments demonstrate the applicability.

Keywords Stemmatology. Object recognition. Llms. Yolo. Text mining.

Edizioni Ca'Foscari

Peer review

Submitted 2025-09-30
Accepted 2025-11-28
Published 2026-01-07

Open access

Citation Hoenen, A. (2025). "LLM-Mining Pre-Stemmatological Philological Literature". *magazén*, 6(2), 215-232.

DOI 10.30687/mag/2724-3923/2025/02/005

1 Introduction

The current article has two main objectives: demonstrate applications of Large Language Models to stemmatological research (digital [humanities] objective) and outline a research avenue for multimodal (image, text) analytic distant reading of large corpora ([digital] humanities objective). Large collections of recently digitised written sources could be used to explore philological literature from all ages. In this article, proof-of-concept (poc) level experiments demonstrate the feasibility of object recognition and text mining with the objective to explore, quantify and by these tokens improve our understanding of the prehistory of the stemmatic method.

Stemmatology is a subfield of philology occupied with textual evolution. It aims at a visual representation of the history of textual variants. The stemmatological methodology which may include text reconstruction and which is often connected with the name of Karl Lachmann (Trovato 2017) is especially useful for texts which originated in the chirographic age where it sometimes emends towards a more original or authentic text form. This in turn is closely tied to editing. Today, computational methods can be applied and are partly shared with the sister disciplines of phylogenetics/systematics (biology) and historical linguistics (Hoenen 2020).

The relationship and mutual influences of these fields date back much further than to the age of computation. These mutual influences have been variously analysed. Interestingly, all three witness their oldest trees and the onset of tree-drawing in about the same time period, the early-mid nineteenth century:

- Biology: after Darwin published a tree in *The origin of species* 1859 there was a "great burst of tree-making" (O'Hara 1996, 85);
- Linguistics: "The first genuine tree diagram of the history of Indo-European was apparently published around 1800 (Auroux 1990), but linguistic trees of history didn't really become widespread until the 1850s" (O'Hara 1996, 84);
- Philology: Collin and Schlyter (1827) were the first to publish a stemma. Shortly thereafter "Carl Zumpt published a genealogy of the known copies of *Ciceros Verrine Orations* in 1831, and Zumpt's stemma was followed by stemmata drawn by Friedrich Ritschl in 1832, and by J.N. Madvig in 1833" (O'Hara 1996, 85).

The similar time range is somewhat striking, all the more since collaborations between these disciplines, despite existent, are usually not understood as the main driving-force of the epochal changes. What is rather uncontroversial though is the benefit tree-drawing meant.

2 The Enormous Benefit of a Visualisation

Whilst language forces us to express concepts one by one forcing our reasoning into one sequence, the visual domain of a stemma is not that restricted. Describing the relationships between witnesses in words is thus more inherently ambiguous than displaying a clear and simple stemma. And this goes for any tree. Taking a simple example such as the tree **[fig. 1]**, one could describe the relationships in words and find many different narrative sequences. An example could begin as 'From a now believed to be lost archetype, only two copies were made. The descendent of one, A, is now at the royal library of Sweden, …'. At this point alone, language would force the author to decide whether to first mention the other copies of root (breadth-first) or to describe the descendants of A had it had some (depth first). The same would go for every witness node in the stemma and naturally one could jump back and forth between breadth-first and depth-first, between bottom-up and top-down or even jump wildly. The point is that many possible narratives map to the same stemma. If now, in addition to this, authors write about the same tradition with different views of relationships or get iffy, it will become much more difficult to compare two narratives than it is to juxtapose and compare two trees.

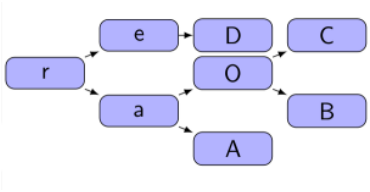


Figure 1
A stemma-like structure with descriptii, capitals for surviving witnesses

The complex genealogical information is much more digestible if displayed as a visualisation rather than if presented as a narrative. The vast success of the tree is in part due to this effect of allowing a simple overview over complicated relations. Scientific exchange is fostered hereby. We shall call the effect visual simplification effect (VSE).

Given the VSE, a valid question is why a tree-like structure for the analysis of mutating units has not been invented before, especially in philology, since staggering amounts of textual variation were known much earlier. Collations, that is side-by-side representations of texts are known very early on, for instance in China where textual criticism is traced back to Liu Xiang (first century BCE) (Fölster, Staack 2021). Although writing system evolution complicated emendation in China, woodblock printing appeared already in the Tang era, around the seventh or eighth century (Barrett 2001) and

would have made stemmata useful for editors. In Western antiquity, the library of Alexandria is known to have hosted many versions of the Homeric epics, compare Nagy (2004). These texts were rather invented orally and may not have one clearly defined original in the same way as born-written works would. Alas, no stemma is known from Alexandria.

In holy texts, variation was present already early on, compare the Qumran manuscripts for instance (Tov 2018). One reaction was that the importance of strict copying for copyists in the Tannaim group (Wegner, 2006, 73) was emphasized, but again, no stemma is known.

Also colophons came into being, recording the local copy histories of single witnesses, but colophons were also copied, sometimes omitted etc. and did not contain any stemmata. Yet, trees as analytic structures have been used since antiquity in a plethora of ways after all, comprising so diverse subjects as the depiction of the descent of aristocratic families and trees of virtues connecting desirable personality traits (Lima 2014). It could thus have been a small step for an early author to transfer this structure.

Being far from exhaustive, this at least shows that there were many possible places where an earlier stemma or graph would already have been useful in order to exploit the VSE. Similar to devices such as the steam engine described by Heron of Alexandria (Roby 2023) or the Baghdad battery (Keyser 1993) there might have been a graph-like structure for text versions ahead of its time which then remained isolated.

Which experiments or research could we conduct to find such a graph, if it had been overlooked? Before diving into this, let us look at the first stemma itself and analyse the invention and the prerequisites.

3 The Invention of the First Stemma

What was the actual reason for Collin and Schlyter to invent their stemma, which they put only into an appendix? It appears, references of philological discourse were rather scarce in their work, but they had a source for the texts they analysed which displayed some family trees of Old Nordic aristocratic families: Fant (1818). Their calling their stemma "et slags stamtre" (Swe. 'a kind of inheritance tree') points to these depictions as their primary inspiration. Their struggle with the terminology, as they called the stemma *schema cognationis* in Latin, corroborates the hypothesis that they had not seen a stemma like graphical depiction for text evolution whatsoever before or something similar. In their case, the coincidental appearance of family trees in one of the secondary sources of their research for the compilation of an edition was just enough to cause the idea of a stemma.

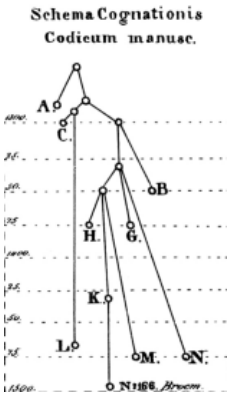


Figure 2
The probably first recorded stemma codicum by Collin, Schlyter 1827

Collin and Schlyter did probably not influence others. What an irony: the mostly tree-shaped stemma has apparently evolved more than once and the visualisation of the genealogy of evolutionary trees in science would thus have multiple roots and be no tree or DAG in a strict sense.

Their invention builds upon previous works and a meticulous collection and analyses of catalogued manuscripts, but in principle, similar conditions could have existed much earlier at least for some works.

4 Was there a More Ancient Stemma or Graph-Like Visualisation?

With the advent of large collections of globally available digitised ancient sources, textual such as the Patrologia Latina (Migne 1993, 1998) or including images and standardized access (IIIF) such as via the VeDPH at Ca' Foscari and on the other hand an impressive increase in technological image recognition capabilities, the time seems right to approach this question with the help of image technology.

LLMs combined with vision encoders could be used without fine-tuning for stemma object recognition. A more conservative approach would be to train an object recognition model for stemmata. Both methods could be used to scan large digital collections for early stemmata. The image recognition could additionally be combined with text extraction. In order to explore if this technology could work, we conduct some poc-level experiments in order to determine whether such an endeavour could be feasible and which technologies seem most promising.

4.1 Dataset and LLM

A small dataset was created, containing:

- one family tree from Fant (1818)
- 125 pages from Collin and Schlyter (1827) with only one page containing the first stemma
- 50 synthetic pages with stemmata and pseudo-text
- Additionally, thanks to the project Open Stemmata (Camps et al. 2021) a corpus of publicly available papers from Persée was available for the experiments containing 61 papers featuring 66 Stemmata and 81 other stemma-like diagrams (false positives)

In the Python programming language, the library *graphviz* was used to generate random trees. These were then placed at a random position onto an artificially generated page with pseudo-text. These pages are not exactly like the ones which occur in the target philological literature, furthermore, text is quadratic and has ragged ends but the appearance is not entirely dissimilar to target structures **[fig. 3]**. For now, if already the poc on synthetic data alone fails, the endeavour would probably be not worth the time and effort. The dataset was forwarded to *GPT4o-mini* via the API from *openAI* for recognition. The prompt combined a role, and some information on how to combine textual and visual evidence.

In a second run, the text of the corresponding page of the image was combined with the image for the Persee dataset where each page was saved as a separate png and each text per page correspondingly (roughly 2500 instances). However, 9 images of graphs had to be excluded due to their being so blurred that even human eyes were not able to distinguish, what kind of diagram that might be.

4.2 Object Recognition with YOLO

LLMs tend to be slow and demanding in hosting. It might be, that one instead want to train one's own object recognition model. Aouinti et al. (2021) used the predominant Object recognition technology YOLO for the detection of illumination. The next poc technique was thus training a YOLO model for stemma recognition. Synthetic training data was generated in the same way as above (1,000 train, 100 val set instances). Additionally, various data augmentation techniques inherent in Yolo were tested, such as rotating by a random angle, overlay, and so forth. In the best condition however training without the augmented examples performed best. In the step placing the trees onto the random text pages label files were generated contemporaneously, indicating the stemma object position by rectangle coordinates. This was enough to train a model with

yolov5. This model was then used to predict all pages of Collin and Schlyter (1827) and the Handbook of stemmatology (Roelli 2020), and the Persee dataset.

4.3 Results

The LLM (GPT4o) recognized trees and distinguished between an ordinary family tree and stemmata (synthetic and real). On the Persee data, the LLM was able to achieve a recall of 0.98, but since almost as many false positives were identified as stemmata, the precision was at a mere 0.44. Given that some of the data were graphically blurred, before the more exact distinction between stemmata and other graphs can be made, the dataset must be improved. The high recall together with the huge number of more than 1400 pages of true negatives which were without any error detected, shows that the LLM is able to recognize graphs. In so far, the poc supports the claim, that a more thorough project set-up will achieve suitable recognition ratios using LLMs.

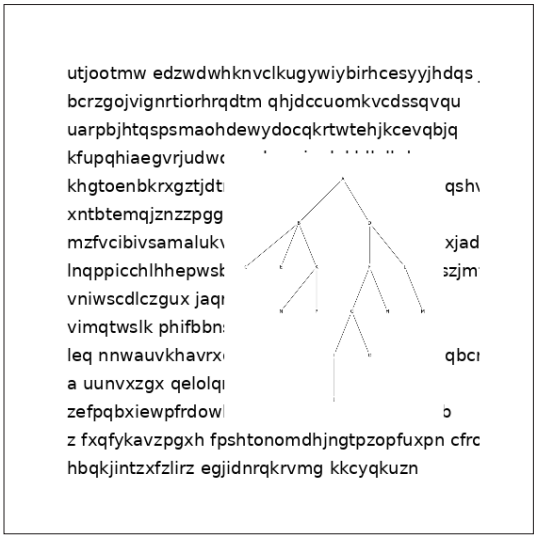
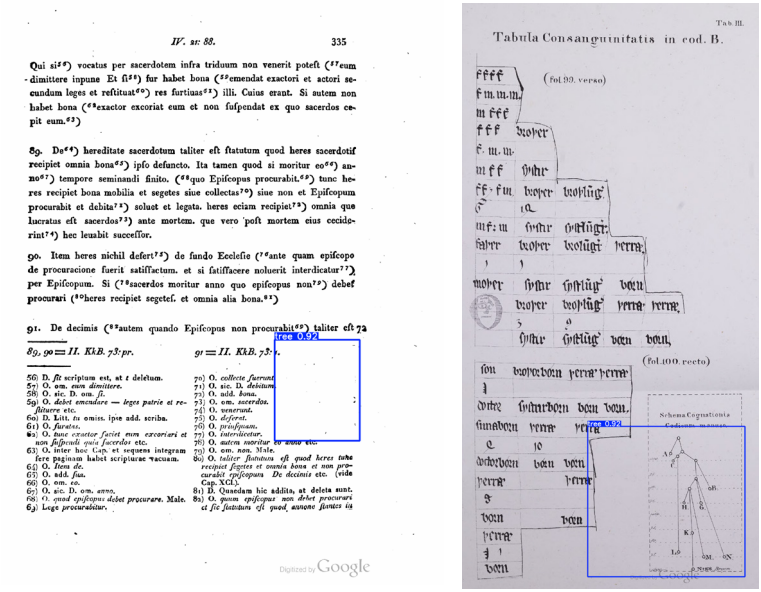


Figure 3 A synthetically generated stemma codicum on a pseudo text page



Figures 4a-b Yolo object recognition, left, an artifact (p. 441) and right, the original true stemma (p. 703)

The YOLO model on the other hand had learned some artefacts of the synthetic data leading to the recognition of rectangles in the margins. This was partly because the graphs with a white box background had been placed onto the texts leaving some text to all sides, which is not matched in authentic texts. Yolo computes the probability of its objects and it is both easy to filter for the margins by the coordinates of the recognized objects and by the probability. Excluding empty pages, objects in the margins and objects below 0.9 probability, the model recognized the true stemma, see Figure 4 plus 3 false positives. The recognition boundaries of the original stemma were a bit distorted and include the unusual lines of the adjacent table. Double checking recognition on all 694 pages from the Handbook of stemmatology, which contains many different kinds of stemmata and graphs, the model does truly recognize stemmatic structures even if not strictly trees and such which are not extremely similar to its input, see Figure 5. At 0.9 probability threshold, on the Persee dataset, the performance dropped to 0.35 precision and a mere 0.1 recall. Many stemmata were not recognized, false positives outnumbered true positives, but true negatives were correctly matched. The model itself being only trained with synthetic data has of course utter limitations, especially since almost none of the persee stemmata looked like the ones, the model had seen during training. The results still point to an applicability because Yolo is known to be a powerful model and

e-ISSN 2724-3923

because it distinguished true negatives well. The same problem as with the LLM might occur, namely that the distinction from other diagrams must be well trained towards and even a combination with text might not be working. This is a challenge if one targets somewhat creative stemmata of the past, the appearance of which is unclear, if they exist.

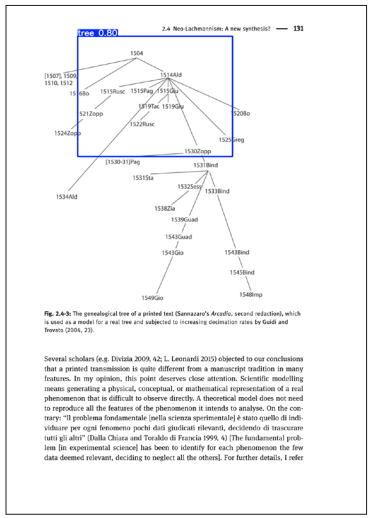
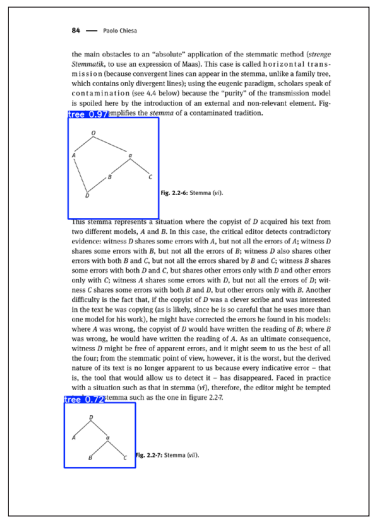


Fig. 2.4-3: The genealogical tree of a printed text (Sannazaro's *Arcadia*, second redaction), which is used as a model for a real tree and subjected to increasing decimation rates by Guidi and Trovato (2004, 23).

Several scholars (e.g. Divizia 2009, 42; L. Leonardi 2015) objected to our conclusions that a printed transmission is quite different from a manuscript tradition in many features. In my opinion, this point deserves close attention. Scientific modelling means generating a physical, conceptual, or mathematical representation of a real phenomenon that is difficult to observe directly. A theoretical model does not need to reproduce all the features of the phenomenon it intends to analyse. On the contrary: "Il problema fondamentale [nella scienza sperimentale] è stato quello di individuare per ogni fenomeno pochi dati giudicati rilevanti, decidendo di trascurare tutti gli altri" (Dalla Chiara and Toraldo di Francia 1999, 4) [The fundamental problem [in experimental science] has been to identify for each phenomenon the few data deemed relevant, deciding to neglect all the others]. For further details, I refer



the main obstacles to an "absolute" application of the stemmatic method (*stronge Stemmatik*, to use an expression of Maas). This case is called horizontal transmission (because convergent lines can appear in the stemma, unlike a family tree, which contains only divergent lines); using the eugenic paradigm, scholars speak of contamination (see 4.6 below) because the "purity" of the transmission model is spoiled here by the introduction of an external and non-relevant element. Figure 2.2-6 simplifies the *stemma* of a contaminated tradition.

This stemma represents a situation where the copyist of *D* acquired his text from two different models, *A* and *B*. In this case, the critical editor detects contradictory evidence: witness *D* shares some errors with *A*, but not all the errors of *A*; witness *D* shares some errors with *B*, but not all the errors of *B*; witness *D* also shares other errors with both *B* and *C*, but not all the errors shared by *B* and *C*; witness *B* shares some errors with both *D* and *C*, but shares other errors only with *D* and other errors only with *C*; witness *A* shares some errors with *D*, but not all the errors of *D*; witness *C* shares some errors with both *B* and *D*, but other errors only with *B*. Another difficulty is the fact that, if the copyist of *D* was a clever scribe and was interested in the text he was copying (as is likely, since he is so careful that he uses more than one model for his work), he might have corrected the errors he found in his models: where *A* was wrong, the copyist of *D* would have written the reading of *B*; where *B* was wrong, he would have written the reading of *A*. As an ultimate consequence, witness *D* might be free of apparent errors, and it might seem to us the best of all the four; from the stemmatic point of view, however, it is the worst, but the derived nature of its text is no longer apparent to us because every indicative error – that is, the tool that would allow us to detect it – has disappeared. Faced in practice with a situation such as that in stemma (vi), therefore, the editor might be tempted to adopt a stemma such as the one in figure 2.2-7.

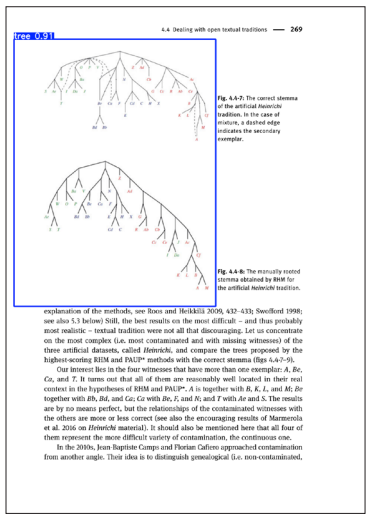


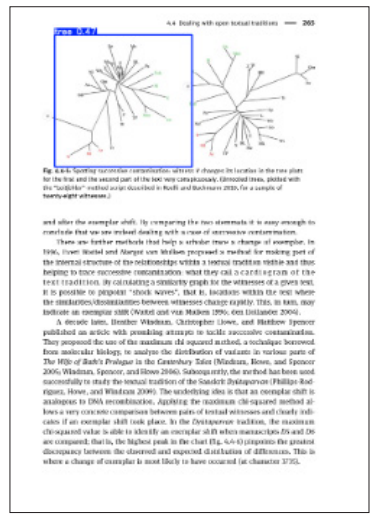
Fig. 4.4-7: The correct stemma of the artificial *Heinrichi* tradition. In the case of mixture, a dashed edge indicates the secondary exemplar.

Fig. 4.4-8: The manually rooted stemma obtained by RHM for the artificial *Heinrichi* tradition.

explanation of the methods, see Roos and Heikkilä 2009, 432–433; Swofford 1998; see also 5.3 below). Still, the best results on the most difficult – and thus probably most realistic – textual tradition were not all that discouraging. Let us concentrate on the most complex (i.e. most contaminated and with missing witnesses) of the three artificial datasets, called *Heinrichi*, and compare the trees proposed by the highest-scoring RHM and PAUP* methods with the correct stemma (figs 4.4-7–9).

Our interest lies in the four witnesses that have more than one exemplar: *A*, *Be*, *Co*, and *T*. It turns out that all of them are reasonably well located in their real context in the hypotheses of RHM and PAUP*. *A* is together with *B*, *K*, *L*, and *M*; *Be* together with *Bb*, *Bd*, and *Ge*; *Co* with *Be*, *F*, and *N*; and *T* with *Ae* and *S*. The results are by no means perfect, but the relationships of the contaminated witnesses with the others are more or less correct (see also the encouraging results of Marmerola et al. 2016 on *Heinrichi* material). It should also be mentioned here that all four of them represent the more difficult variety of contamination, the continuous one.

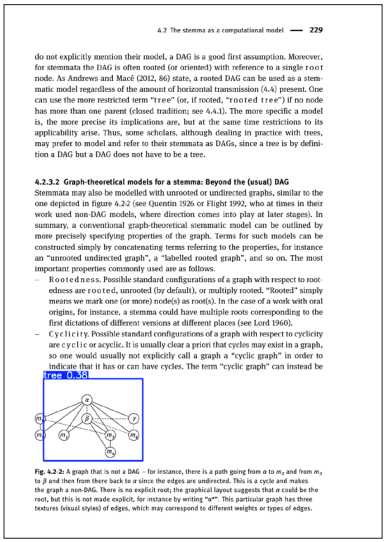
In the 2010s, Jean-Baptiste Camps and Florian Cafiero approached contamination from another angle. Their idea is to distinguish genealogical (i.e. non-contaminated,



and after the exemplar shift. By comparing the two stemmata it is easy enough to conclude that we are indeed dealing with a case of successive contamination.

There are further methods that help a scholar trace a change of exemplar. In 1936, Isaac Boekel and Margot van Mulken proposed a method for making part of the inverted structure of the relationships within a textual tradition visible and thus helping to trace successive contamination: what they call a cardiogram of the text tradition. By calculating a similarity graph for the witnesses of a given text, it is possible to pinpoint "shock waves", that is, locations within the text where the similarities/dissimilarities between witnesses change rapidly. This, in turn, may indicate an exemplar shift (Waitzel and van Mulken 1996, den Uitlander 2004).

A decade later, Heather Windram, Christopher Howe, and Matthew Spencer published an article with promising attempts to tackle successive contamination. They proposed the use of the maximum chi squared method, a technique borrowed from molecular biology, to analyse the distribution of variants in various parts of *The Wife of Bath's Prologue* in the Canterbury Tales (Windram, Howe, and Spencer 2006; Windram, Spencer, and Howe 2006). Subsequently, the method has been used successfully to study the textual tradition of the Sanskrit *Śyāmaparvan* (Phillips Rodriguez, Howe, and Windram 2008). The underlying idea is that an exemplar shift is analogous to DNA recombination. Applying the maximum chi-squared method allows a very concrete comparison between pairs of textual witnesses and clearly indicates if an exemplar shift took place. In the *Śyāmaparvan* tradition, the maximum chi-squared value is able to identify an exemplar shift when manuscripts D5 and D6 are compared; that is, the highest peak in the chart (fig. 4.4-6) pinpoints the greatest discrepancy between the observed and expected distribution of differences. This is where a change of exemplar is most likely to have occurred (at character 3735).



4.2 The stemma as a computational model — 229

do not explicitly mention their model, a DAG is a good first assumption. Moreover, for stemmata the DAG is often rooted (or oriented) with reference to a single `root` node. As Andrews and Macé (2012, 86) state, a rooted DAG can be used as a stemmatic model regardless of the amount of horizontal transmission (4.4) present. One can use the more restricted term "`tree`" (or, if rooted, "`rooted tree`") if no node has more than one parent (closed tradition; see 4.4.1). The more specific a model is, the more precise its implications are, but at the same time restrictions to its applicability arise. Thus, some scholars, although dealing in practice with trees, may prefer to model and refer to their stemmata as DAGs, since a tree is by definition a DAG but a DAG does not have to be a tree.

4.2.3.2 Graph-theoretical models for a stemma: Beyond the (usual) DAG

Stemmata may also be modelled with unrooted or undirected graphs, similar to the one depicted in figure 4.2-2 (see Quentin 1926 or Flight 1992, who at times in their work used non-DAG models, where direction comes into play at later stages). In summary, a conventional graph-theoretical stemmatic model can be outlined by more precisely specifying properties of the graph. Terms for such models can be constructed simply by concatenating terms referring to the properties, for instance an "unrooted undirected graph", a "labelled rooted graph", and so on. The most important properties commonly used are as follows.

- `Rootedness`. Possible standard configurations of a graph with respect to rootedness are `rooted`, unrooted (by default), or multiply rooted. "Rooted" simply means we mark one (or more) node(s) as root(s). In the case of a work with oral origins, for instance, a stemma could have multiple roots corresponding to the first dictations of different versions at different places (see Lord 1960).
- `Cyclicity`. Possible standard configurations of a graph with respect to cyclicity are `cyclic` or acyclic. It is usually clear a priori that cycles may exist in a graph, so one would usually not explicitly call a graph a "cyclic graph" in order to indicate that it has or can have cycles. The term "cyclic graph" can instead be

`tree (4.3)`

Fig. 4.2-2: A graph that is not a DAG – for instance, there is a path going from α to m_3 and from m_3 to β and then from there back to α since the edges are undirected. This is a cycle and makes the graph a non-DAG. There is no explicit root; the graphical layout suggests that α could be the root, but this is not made explicit, for instance by writing "α^{root}". This particular graph has three textures (visual styles) of edges, which may correspond to different weights or types of edges.

Figures 5a-e
Recognition of different types of stemmata from the Handbook of stemmatology through the model ranging from similar to the training to quite different

All in all, the poc has shown that a larger scale object recognition for larger collection seems feasible.

5 Recapitulation of Historical Processes towards Text Mining

What is similar among the three sciences using trees presumably is a larger overall increase in amounts of data that they had to analyse. The reasons for these increasing amounts of data were presumably at least partly different for the three. Colonialism expanded numbers of known species and the knowledge of languages considerably. For philology, however, the main reason for an increase in data should primarily have been the invention of the printing press.

As soon as an editor had to print an ancient work, transmitted in handwriting, he could ask which version of the slightly differing versions at his disposal he should use. The key question and probably the birth helper of stemmatology. Readers would naturally prefer and read that edition which could plausibly offer the best possible version exerting a certain pressure on early print age editors. For choosing, of course, an editor would have to have access to multiple versions – prerequisite 1 – and the philological insight that some versions might be more authentic than others – prerequisite 2.

As for the availability of versions, it should have depended on many factors such as an increased mobility, superior cataloguing and less circulation of the manuscripts themselves. This was only

gradually happening after printing had been invented. In the early days of printing, it is logical that handwritten manuscripts remained in circulation and thus harder to access for printing and throughout the sixteenth century books were still rather rare as compared to today, consider Pettegree (2010). After all, one needed people able to read them and broader literacy through schooling only slowly but steadily advanced, compare for instance Eskelson (2021) on literacy development. At some point in time, printed books must then have become the norm for private and public reading, whilst handwritten manuscripts, codices etc. were becoming less common. The time frame for these processes can roughly be estimated to be during the seventeenth and eighteenth centuries. For an in depth analysis based on the outputs of printing presses, see Buringh, van Zanden 2009. Improved cataloguing in libraries (compare e.g. the first printed catalogue of the Bodleian in 1605; Bodleian 1605/1986) falls in the same range. Thus it is safe to assume that there was a steady increase in available sources for editors.

At the same time, the awareness of variation and how to deal with it in philological discourse increased and concepts such as the shared innovation or error were elaborated upon. In fact, many mechanisms of the stemmatic method and of emendation have been understood since antiquity, compare Haverling (2020). However, putting them together systematically into a rigorous method which is known, learned, practised and taught at least in part of the field ever since falls into the nineteenth century (Haugen 2020, 57) clearly coinciding with the development of the stemmatological method.

These prerequisites could be very different in linguistics and biology. Another important peculiarity for philology is that a single tradition, a text, is a rather isolatable unit. There is no such thing as a tree of all texts. In linguistics and biology however scientists engaging in the analysis of whatever evolutionary entities (clams, canines, felines, … or Indo-European languages etc.) would additionally have to deal with the question of how to accommodate their units into a tree of life or of all languages (if one believes in one language origin). The 'laboratories' of editors are smaller improving chances for an earlier holistic graphical approach.

If it were only for the awareness of change and the availability of versions, one could argue for holy texts evidence has been tantamount ever since, even before printing. However, in that case, two thoughts might help to explain why holy texts witnessed an independent development within philology/stemmatology. On the one hand many other aspects than only micro-variation on a linguistic level would play a role when going towards an urtext, exegesis, the implications of the text. Emendation would be difficult to explain. On the other, for holy texts there was so much evidence that despite the idea of a genealogical tree for the New Testament being expressed by

Bengel (1763) the setting was so complicated that it simply took much longer than in the classics. Here, available evidence was increasing just as much as that for certain works stemmatic relations became too complex to be easily comprehensible by words alone and not complex enough to refrain from attempts to approach the entirety of the evidence graphically.

Given these assumptions, the relatively similar time frame of occurrence of the tree-drawing branches of the three sciences of biology, linguistics and philology could be at least in part incidental. All three saw an increasing amount of data and discourse, whilst the reasons for the increases might have been different. In order to validate such hypotheses, quantitative analyses would be needed. One way could be to use text mining to measure the amounts of witnesses editors used over time and how their relations have been analysed.

6 Text Mining and Information Extraction

In this experiment, relation extraction from secondary literature, especially from philological literature for the time before the invention of stemmata is being investigated. First, a small artificial corpus of editorial descriptions of the same stemma-like structure is generated, then GPT4-o is used with an appropriate prompt. The task more precisely is, from differing textual descriptions of the same tree-like structure to retrieve that structure and display it in an unambiguous format. As a target format, we choose the Newick format.[1] Previous experiments to make the LLM generate a graph directly had led to less usable results.

First, we choose some stemma-like structure, seen in Figure 1. Then, we generate textual descriptions for this structure. We do this by first defining 4 base sentences for each parent. '*O has been copied twice, once into C, once into B*.' could be one such sentence. For each of the sentences, we manually create 5 or 7 alternative formulations. Each text shall feature a variant of each of the four sentences. In this way the entire structure is described. The sequence however may wildly differ, as is normal for human descriptive text. We generate 100 distinct sequence permutations of the four sentences and randomly fill each with a variant. To round up the text, we add some introductory and final text without structural implications. Finally, we insert so-called distractor sentences, that is sentences which do not bare information on the direct relationships, we insert them via

1 See the full definition here: https://phylipweb.github.io/phylip/newicktree.html.

a randomizer. Such a sentence may read *'O was not copied from e.'* and can be inserted anywhere in the text. 44 times such a distractor was inserted. A full example of a generated text would be:

> This text treats the tradition of Rabanus Testus Textus. The text has been transmitted in handwriting. We have located 5 extant copies in various libraries. e was copied, the copy is D. O and A are closely related, probably they have been copied from the same lost manuscript a. O was not copied from e. The archetype r was copied into e and a. O has been copied into B and C. The tradition is thus a limited one in size and scope but the relations are quite clear leading to a wonderful stemma albeit with descriptii and chains of hypothetical nodes.

The expected target structure should be 'r(e(D),a(O(C,B),A)' or any equivalent.

As for the prompt engineering, we chose different approaches. We started with a basic zero-shot approach asking GPT4-o to 'Extract the Newick tree from this text:'. Then we tried a one-shot prompt with a smaller example, we tried chain-of-thought (cot) prompting (Wei et al. 2022), a technique where the task is broken down into subsequent smaller steps, the LLM solves step-by-step. Here, we asked GPT4-o to first extract relations (edges) and then from the relations to build a stemma. Finally, we tried a two-shot scenario. We varied one-shot scenarios as to whether the example was given within the prompt or whether we simulated a conversation as user-assistant interaction (interactive). Finally, we prompted for non wordiness, that is prompted to provide only the tree, no explanation or anything else. For results see Table 1.

The results suggest that zero-shot and cot alone are not enough. This might be so, because given GPT's training data, the task is relatively unusual. However, more shots improve the result significantly consistent with state-of-the-art research on LLMs. With two shots more than 90% of texts lead to the entirely correct extraction. Given that we did not optimize prompt engineering as to wording etc., this is a very good result. Interestingly, the interactive one-shot scenario performed noticeably worse than the non-interactive example. Also cot alone achieved a better result than zero-shot, but adding an example, this was turned around. In order to investigate these effects more data would be required. The format requirement to provide only the tree, not a wordy answer was adhered to.

Table 1 Stemma extraction performance of different prompting approaches

Method	Hits	Misses	Accuracy (%)
zero-shot	6	94	0.06
one-shot	89	11	0.89
cot	9	81	0.09
one-shot interactive	67	33	0.67
two-shot	94	6	0.94
cot with one-shot	82	18	0.82

The distractors did not affect the results, there was no meaningful difference between the accuracies for texts with as opposed to texts without distractors. In a certain sense, the introductory and final phrases were also distractors. Their differing position however suggests that at least placement of such a distractor has few influence on extraction.

The implication of the experiment is that LLMs in stemmatology could be used in the future to compare older but also more recent philological literature and extract stemmata from texts even where the editorial approach may be opposed to stemmata. The number of nodes is the number of witnesses. Especially for historical descriptions of traditions which may also mention and describe witnesses which have later been lost and for older literature at scale this approach of information extraction could lead to new insights. The good results also would point to possibly related tasks such as a binary classification if two texts are equivalent in the tree-structure they describe or not and a task where from a tree, a textual description can be generated for instance for visually impaired readers. Technically, the experiment belongs to the field of text mining or more precisely information extraction. An overview of applications in biology can be found in Farrell et al. 2022. Fine-tuning, dataset simulation and so forth are ultimately other pathways for research in this direction.

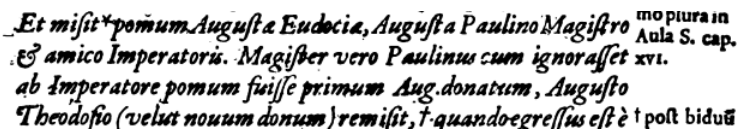
6.1 Application to Real Text

Finally, the previous experiment is again only operating on synthetic data and whilst image recognition already showed that with synthetic data alone, good results can be achieved, here applicability to true data should at least be tested. As a case study we use the *Chronicon Alexandrinum*, a Greek chronicle spanning Creation to Byzantine Emperor Heraclius published in Latin by Matthaeus Raderus (Munich, 1615). Although it is not perfect for our purposes as chonicles have transmission peculiarities, the time range is the right one. The meta-text includes sources like Eusebius, Africanus, Epiphanius, etc.

The images of the Google Book were loaded into GPT4o-mini

alongside a prompt: '*Correcting small OCR inconsistencies, analyze this text and extract from the text any relations between manuscripts or versions in a structured machine readable way.*' An example input page on which this operated, in order to demonstrate how distorted Latin OCR for these texts can be, which is rather the norm than the exception.

Original from p.742 in the pdf (excerpt):



Underlying OCR:

> Etmifit*pomumAugustaEudocia,AugustaPaulinoMagiftroAula S.cap.
> &amico Imperatoris. Magiſter vero Paulinus cum ignoraffet xv1.
> ab Imperatorepomumfuiſſeprimum Aug.donatum,Augusto Theodofio (velut nouumdonum)remifit,&quandoegreſſus estètpoſt biduű

Normalized Version by prompt (showing the underlying ability of the LLM to master such inconsistencies, the note in the margin was however not recognized as such):

> Et misit pomum Augusta Eudocia, Augusta Paulino Magistro Aulae S. cap.
> & amico Imperatoris. Magister vero Paulinus cum ignorasset XVI ab Imperatore pomum fuisse primum Aug. donatum, Augusto Theodosio (velut novum donum) remisit,, quando egressus est et post biduum

The experiment showed that LLMs of the size of GPT4o-mini are able to handle distorted OCR and Latin when extracting copy or citation relations which is an important addition to the first text extraction experiment based only on synthetic English.

7 Conclusion

Poc-level experiments have demonstrated the potential of LLMs and other Machine Learning Models in analysing large collections of digitized data in order to elucidate the pre-history of stemmatology. Object recognition could find and analyse graphical precursors and

possibly even earlier stemmata, whilst text mining methods such as stemmatic relation extraction could trace witness availability and amounts of discourse on witness relations. This could help understand the processes at work in the appearance of tree-drawing in philology across languages and time periods. Data, scripts including prompts and a yolo model have been released on https://github.com/ArminHoenen/prehistorical_stemmata.

Bibliography

Aouinti, F.; Eyharabide, V.; Fresquet, X.; Billiet, F. (2022). "Illumination detection in IIIF medieval manuscripts using deep learning". *Digital Scholarship in the Humanities*. https://academic.oup.com/dsh.

Auroux, S. (1990). "Representation and the Place of Linguistic Change Before Comparative Grammar". de Mauzo, T.; Formigari, L. (eds), *Leipzig, Humboldt, and the Origins of Comparativism*. Amsterdam: John Benjamins, 213-38.

Barrett, T.H. (2001). "Woodblock dyeing and printing technology in China, c. 700 A.D.: The innovations of Ms. Liu, and other evidence". *Bulletin of the School of Oriental and African Studies*, 64(1), 85-9.

Bengel, J.A. (1763) *D. Io. Alberti Bengelii Apparatus criticus ad Novum Testamentum*. 2nd ed. Tubingae (Tübingen): Sumtibus Io. Georgii Cottae.

Bodleian Library (1605/1986). *The First Printed Catalogue of the Bodleian Library, 1605: A Facsimile*. Comp. by T. James. Oxford: Clarendon Press.

Buringh, E.; van Zanden, J.L. (2009). "Charting the 'rise of the West': Manuscripts and Printed Books in Europe, a Long-term Perspective from the Sixth Through the Eighteenth Centuries". *Journal of Economic History*, 69(2), 409-45. https://doi.org/10.1017/S0022050709000837.

Camps, J.-B.; Gabay, S.; Fernández Riva, G. (2021). *Open Stemmata: A Digital Collection of Textual Genealogies = EADH2021: Interdisciplinary Perspectives on Data, 2nd International Conference of the European Association for Digital Humanities* (Krasnoyarsk, 21-25 September 2021).

Collin, H.S.; Schlyter, C.J. (1827). *Corpus Iuris Sueo-Gotorum Antiqui 1*. Stockholm: Z. Haggestrom.

Eskelson, T.C. (2021). "States, institutions, and literacy rates in early-modern Western Europe". *Journal of Education and Learning*, 10(2), 83-92.

Fant, E.M. (ed.) (1818). *Scriptores rerum Svecicarum medii aevi ex schedis praecipue Nordinianis collectos, dispositos ac emendatos*, Tomus I. Upsaliae: Zeipel et Palmblad (Reg. Acad. Typographi).

Farrell, M.J.; Brierley, L.; Willoughby, A.; Yates, A.; Mideo, N. (2022). "Past and Future Uses of Text Mining in Ecology and Evolution". *Proceedings of the Royal Society B*, 289(1975), 20212721.

Fölster, M.J.; Staack, T. (2021). "Collation in Early Imperial China: From Administrative Procedure to Philological Tool". Quenzer, J.B. (ed.), *Exploring Written Artefacts*, 889-912. Berlin: De Gruyter.

Haugen, O.E. (2020). "2 The genealogical method". Roelli, P. (ed.), *Handbook of Stemmatology: History, Methodology, Digital Approaches*. Berlin: De Gruyter, 57-138. https://doi.org/10.1515/9783110684384-003.

Haverling, G.V.M. (2020). "2.1 Background and early developments". Roelli, P. (ed.), *Handbook of Stemmatology: History, Methodology, Digital Approaches*. Berlin: De Gruyter, 59-80. https://doi.org/10.1515/9783110684384-003.

Hoenen, A. (2020). "8 Evolutionary models in other disciplines". Roelli, P. (ed.), *Handbook of Stemmatology: History, Methodology, Digital Approaches*. Berlin: De Gruyter, 534-86. https://doi.org/10.1515/9783110684384-009.

Jerome (ca. 384). *Praefatio Hieronymi in Quatuor Evangelia* [Latin text, Migne PL 29, col. 525-528]. Early Church Texts. https://earlychurchtexts.com/main/jerome/preface_to_four_gospels.shtml.

Keyser, P.T. (1993). "The Purpose of the Parthian Galvanic Cells: A First-Century A.D. Electric Battery Used for Analgesia". Journal of Near Eastern Studies, 52(2), 81-98.

Migne, J.-P. (ed.) (1844-65/1864-65). *Patrologiae Cursus Completus: Series Latina*. 221 vols. Paris: Migne; indices 1862-65. Repr.: Turnhout: Brepols, 1982-93.

Migne, J.-P. (ed.) (1857-66). *Patrologiae Cursus Completus: Series Graeca*. 161 vols. Paris: Migne. Repr.: Athens: Centre for Patristic Publications, 1997-98.

Nagy, G. (2004). *Homer's Text and Language*. Urbana: University of Illinois Press.

O'Hara, R.J. (1996). "Trees of history in systematics and philology". *Memorie della Società Italiana di Scienze Naturali e del Museo Civico di Storia Naturale di Milano,* 27(1), 81-8.

Pettegree, A. (2010). *The Book in the Renaissance*. New Haven: Yale University Press.

Rader, Ma. (ed., trans.) (1615). *Chronicon Alexandrinum idemque astronomicum et ecclesiasticum (vulgò Siculum seu Fasti Siculi)*. Monachii: Ex formis Annae Bergiae viduae.

Roby, C.A. (2023). *The Mechanical Tradition of Hero of Alexandria*. Cambridge: Cambridge University Press.

Tov, E. (2018). *The Essence and History of the Masoretic Text* (lecture paper). Jerusalem: Hebrew University of Jerusalem, 6-8.

Trovato, P. (2017). *Everything You Always Wanted to Know about Lachmann's Method*. Limena (PD): Libreriauniversitaria edizioni.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E.; Le, Q.; Zhou, D. (2022). "Chain-of-thought Prompting Elicits Reasoning in Large Language Models". *Advances in Neural Information Processing Systems*, 35, 24824-37.

Semestral journal

Venice Centre
for Digital and Public Humanities
Università Ca' Foscari Venezia

Università
Ca'Foscari
Venezia