

1 Open Data Adoption Across the Humanities

Insights from the Journal of Open Humanities Data

*Barbara McGillivray, Daniele Borkowski,
Andrea Farina, Simon Mahony, and
Xiaoguang Wang*

1.1 The Rise of Open Data in Digital Humanities

Openness and sharing have been critical in defining Digital Humanities (DH), even as regional and linguistic differences continue to shape issues of equality, diversity, and inclusion (Mahony, 2018; Mahony & Gao, 2019). What distinguishes DH from many traditional disciplines is its embrace of openness, which has enabled unprecedented access to data and resources (Mayer-Schönberger & Cukier, 2013; Alaimo et al., 2020). This shift has accelerated the dissemination of knowledge and made it easier for researchers to build upon previous work. DH originated in large-scale digitisation efforts, such as the creation of manuscript and full-text databases, which were initially often proprietary but have become increasingly open due to funder mandates (e.g., Wellcome Trust, UKRI, Gates Foundation).

The nature of DH has encouraged the development of online resources and a growing interest in data-driven research (Jones, 2014; Terras, 2018; 2022). Its contributions include not only the creation of digital tools and methods but also the transformation of scholarly communication and engagement practices (Arthur and Hearn, 2021). According to the Open Definition, Open Data refers to data that is freely available for anyone to use, modify, and share, often extending its use well beyond the original research purpose. In the humanities, data can vary widely across disciplines and may include primary and secondary sources (such as texts and documents) alongside ancillary materials like code (Owens, 2011; Farina et al., 2025). This encompasses both born-digital content (e.g., digital art, social media) and digitised material (e.g., scanned manuscripts). However, such formats can pose preservation challenges, including issues of versioning, software dependencies, and long-term accessibility. Crucially, data should be recognised as a research output in its own right, deserving of appropriate academic credit. Often, those who compile, clean, or code datasets are not the same individuals who publish articles based on them, which makes it essential to value these contributions as core components of the research process.

Despite growing acceptance of openness, significant barriers remain to the adoption of open data in the humanities. Many researchers, including early-career

scholars, lack awareness of the value and responsibilities associated with sharing outputs. Preparing datasets for reuse can be time-consuming and is often unsupported at the institutional level. Moreover, concerns over others using the same data, along with legal, ethical, and technical challenges (e.g., copyright restrictions), can discourage openness (Candela et al., 2015). Nevertheless, funder requirements are helping to shift the norm. Research by the Wellcome Trust, for instance, highlights such mandates as key drivers for sharing full or partial datasets. Publishing datasets openly is increasingly seen as good scholarly practice, encouraging collaboration while enabling validation and replication (Van den Eynden et al., 2016; Colavizza et al., 2020). Although fears of data misuse persist, available evidence points to limited negative outcomes and significant benefits, such as new partnerships and increased citations (Van den Eynden et al., 2016). Still, the cost of data sharing, especially in resource-constrained humanities fields, remains a considerable hurdle.

Research practices in DH continue to evolve in response to emerging best practices, disciplinary standards, and shifting priorities. Open Access is becoming a default publication mode, and it brings tangible benefits: in addition to increasing discoverability and circulation, it leads to measurable rises in downloads, citations, and broader scholarly impact across fields (MacCallum & Parthasarathy, 2006; McGillivray & Astell, 2019; Maddi & Sapinho, 2022; Dorta-González & Dorta-González, 2023; Huang et al., 2024). Moreover, when openly shared datasets are linked using shared standards, creating linked open or smart data, they support new types of research questions that cross disciplinary boundaries (Mahony, 2022). Building on this potential, DH scholars are moving toward publishing open and smart data with built-in mechanisms for data exchange and interoperability, such as TEI XML, RDF, persistent URIs, and robust metadata standards. These practices allow research datasets to interact and form part of a wider digital ecosystem that further expands scholarly possibilities (e.g., Passarotti et al., 2024).

Achieving true openness requires addressing challenges across the entire research lifecycle, not just at the point of publication. Transparency in data collection, processing, and analysis not only supports reproducibility and verifiability but also enables the generation of new knowledge. Global initiatives (such as those by UKRI, LERU, the European Commission, and UNESCO) show the importance of long-term curation and preservation for the sustainability of research. DH plays a central role in these efforts, advocating for openness while engaging diverse audiences through collaborative and inclusive practices (Longley Arthur & Hearn, 2021).

1.2 The Journal of Open Humanities Data

The Journal of Open Humanities Data (JOHD),¹ established in 2015 and published by Ubiquity Press, provides a platform for publishing data papers, publications that describe datasets, including their structure, collection methods, and potential reuse. The journal aligns with the broader open research movement, promoting transparency, collaboration, and the FAIR (Findable, Accessible, Interoperable, and Reusable) principles for data management (Wilkinson et al., 2016). Since its

beginnings, JOHD has been committed to fostering a culture of open data sharing in the humanities. Unlike traditional research journals, JOHD's focus on data papers bridges a crucial gap in humanities research, where data management has historically been fragmented and less standardised compared to STEM fields. JOHD provides a dedicated venue for humanities researchers to share and document their data and encourages best practices in data curation, fostering a collaborative research environment. The journal publishes both short data papers of approximately 1,500 words and longer discussion papers that explore methodological and theoretical issues in data-driven humanities research. A requirement for publication of data papers in JOHD is that datasets are deposited in open-access repositories, to ensure long-term accessibility and reusability.

Research Data Journal for the Humanities and Social Sciences (RDJ)² was the other multi-disciplinary data journal specifically dedicated to the Humanities until early 2025, when it was discontinued. Established by Data Archiving and Networked Services (DANS) in 2016 and published by Brill, it featured medium-length articles (up to 2,500 words) describing datasets and their research context. Unlike JOHD, RDJ's scope also covers the social sciences. Both journals were founded as fully Open Access, a decision that aimed to lower barriers to access, promote transparency and reuse, and encourage a wider adoption of open data principles, particularly in the humanities (and social sciences), where data sharing has traditionally been less common than in the sciences.

Until 2019, JOHD and RDJ followed comparable trajectories. Since 2019, however, JOHD has undergone significant expansion (Figure 1.1), fuelled by the appointment of a new editor-in-chief (Barbara McGillivray), who has led a growing editorial team and established a strong presence in the academic community, via invited talks, events, academic publications by the editorial team, and social media activities (Wigdorowitz et al., 2024). Strengthening its engagement with the digital humanities communities, the journal has seen a steady rise in submissions, citations, and readership, reflecting the growing recognition of data papers as a valuable scholarly output (Marongiu et al., 2025).

The impact of data papers extends beyond publication metrics, as the journal team's research work confirms. As illustrated in McGillivray et al. (2022), datasets accompanied by data papers experience higher levels of engagement, citation, and reuse, reinforcing their value within the scholarly ecosystem. The study also finds that data papers enhance dataset discoverability, allowing researchers across different disciplines to utilise humanities data in innovative ways. Additionally, social media engagement, particularly through platforms like Twitter (X), has played a pivotal role in amplifying the visibility of data papers, leading to increased downloads and broader dissemination of open research.

JOHD serves a broad spectrum of disciplines within the humanities, including linguistics, history, archaeology, literary studies, media studies, and cultural studies. It includes a wide array of methodological approaches such as computational humanities, geospatial analysis, and network analysis, bridging qualitative and quantitative traditions in humanities research and fostering interdisciplinary collaboration. One of JOHD's core contributions is its role in providing the diversity of data types

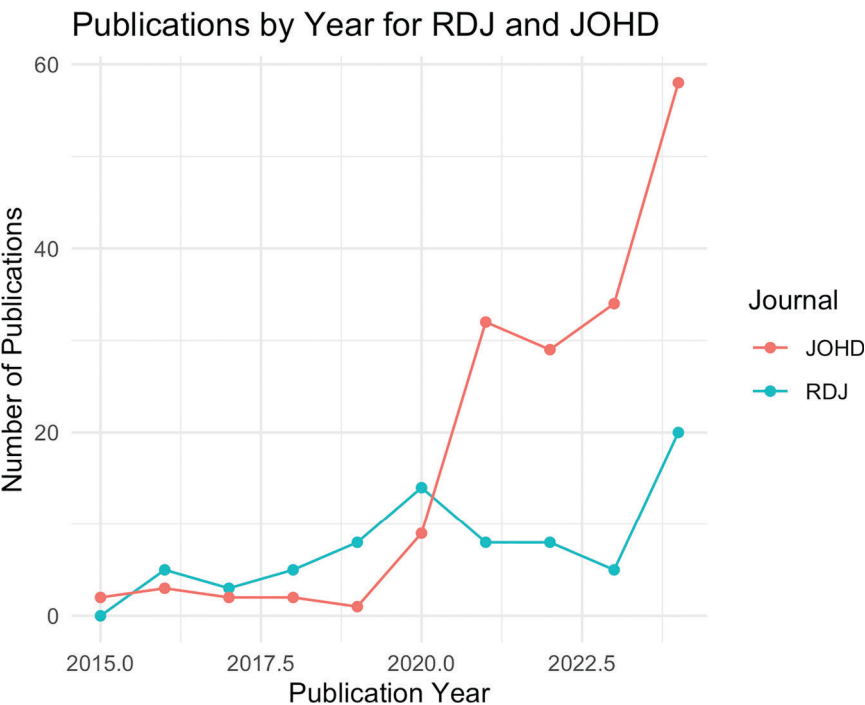


Figure 1.1 Yearly number of articles published in the Journal of Open Humanities Data (JOHD) and in Research Data Journal for the Humanities and Social Sciences (RDJ) from 2015 to 2024.

and practices in the humanities. Unlike STEM disciplines, where data is often highly structured, humanities data can take many forms (Farina et al., 2025, p. 5–6), including textual data such as digitised manuscripts, transcribed interviews, and literary corpora; visual data such as historical maps, scanned artworks, and film archives; geospatial data such as GIS datasets for urban history studies; and multi-modal data combining text, images, and sound in cultural analysis. This diversity necessitates flexible data management approaches, and JOHD provides a platform for scholars to explore and define best practices in handling these varied data types. In the analysis presented in the next section we will investigate this diversity in more detail. We focused our analysis on JOHD because JOHD publishes a larger volume of humanities-focused data papers compared to RDJ, particularly since 2019, providing a more substantial and consistent dataset for analysis.

1.3 An Analysis of Data Papers in DH

We analysed the full set of data papers published in JOHD between 2015 and the end of 2024. In this section we present the framework we developed to manually

annotate the data and the results of our quantitative and qualitative investigation of the data. Our analysis encompasses various dimensions, including publication dates, macro-fields, dataset sizes, dataset types, methods of generation, annotation levels, and sources and the dataset and code are available on Figshare.³

Of the original 109 data papers, 106 contained sufficient information about their associated dataset to be further analysed.⁴ We manually annotated the dataset, refining the guidelines through iterative discussion and testing to ensure consistency. Existing classification schemes were not well suited to the evolving nature of data practices in the humanities, so we developed custom annotation categories guided by the study's research questions. We first describe our annotation of key dataset features and analyse each in turn, then examine relationships between them to reveal patterns in open data practices across humanities disciplines and their implications for sharing, reuse, and interoperability.

Approximately half of the data papers in the dataset were published after 2022 (Figure 1.2). In general, the number of data papers in the dataset had a noticeable rise in 2020, followed by a steep upward trend. This may reflect a broader shift in scholarly communication practices within DH, including increasing recognition of data as a first-class research output and growing support from institutions and journals for open science and data sharing. The rise is also likely linked to targeted efforts by JOHD to promote data publication, including a series of community outreach initiatives that began around 2020 and actively encouraged researchers to submit data papers.

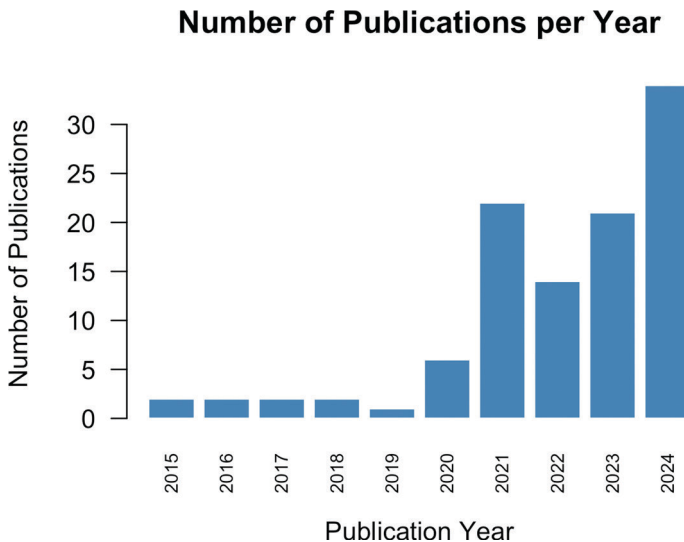


Figure 1.2 Bar chart showing the number of publications per year from 2015 to 2024 in the dataset.

Table 1.1 Frequency of disciplines represented in the dataset. Each data paper was associated with one main discipline

<i>Discipline</i>	<i>Number of data papers</i>
Linguistics	21
Literary studies	15
History	14
Classics	13
Media studies	13
Assyriology	6
Biblical studies	4
Information studies	4
Archaeology	3
Natural Language Processing	3
Bibliometrics	2
Ethnography	2
Criminology	1
Education Research	1
Heritage	1
Musicology	1
Publishing	1
Sociology	1

Table 1.2 Frequency of macro-disciplines represented in the dataset

<i>Macro-discipline</i>	<i>Number of data papers</i>
Literary & Cultural Studies	46
History & Heritage	24
Linguistics & NLP	24
Information & Bibliometrics	7
Social Sciences	5

Each paper was categorised according to the disciplines represented in it; we later selected the discipline most strongly associated with each paper (Table 1.1). The dataset primarily represents Literary Studies (15), Linguistics (21), History (14), Classics (13), and Media Studies (13), reflecting a strong focus on textual and historical analysis. Less frequent disciplines, such as Assyriology (6), Biblical Studies (4), and Information Studies (4), contribute to the dataset to a lesser extent, while fields like Criminology, Heritage, and Musicology appear only once. The disciplines were further grouped to create custom macro-field labels (Table 1.2) for easy comparison and analysis. Grouping disciplines into macro-fields highlights Literary & Cultural Studies as the most represented label (46), followed by History & Heritage and Linguistics & NLP (24 each). Information & Bibliometrics (7) and Social Sciences (5) are less prevalent. This

distribution shows the dataset's emphasis on language, literature, and history, with limited engagement from social sciences and information-related fields.

JOHD publishes a general issue and several so-called "special collections", guest-edited collections dedicated to specific themes. In the dataset, 76% of the data papers belong to the general issue, while the rest are distributed among the special collections ("Data and workflows for multilingual digital humanities", "Computational Humanities Research Data", "Humanities Data in the time of COVID-19", and "Representing the Ancient World through Data").

The data papers sometimes include some information regarding the size of the dataset they describe, often expressed in terms of the number of data points (e.g., entries, records, rows in a table, or images in a collection). Any information included in the data paper was recorded and later on the papers were grouped into the following five categories:

- Extra large-scale datasets exceeding 10 million tokens, records, or documents.
- Large-scale datasets ranging from 100,000 to 10 million records.
- Medium-scale datasets containing 10,000 to 100,000 tokens or items.
- Small-scale datasets including fewer than 10,000 records.
- N/A or Unspecified Size covers datasets without explicit size information.

When clear numerical thresholds were absent from the data paper, their category was inferred based on metadata, dataset descriptions, and comparisons to similar resources because. The most common categories are Medium and Large-scale, as shown in Table 1.3.

We looked for a possible association between dataset size and disciplines. The results of a Pearson's Chi-squared test with simulated p-value (based on 10,000 replicates) indicate that there is a statistically significant association between the dataset size and disciplines ($p < 0.05$). Figure 1.3 shows differences in dataset size among disciplines. For example, Information Studies and Linguistics have notable proportions of Extra Large-scale datasets, while fields like Ethnography and Biblical Studies rely predominantly on Small-scale and Medium-scale datasets.

Table 1.3 Frequency of the main categories of data size based on the content of the data papers

<i>Data size</i>	<i>Number of data papers</i>
Extra Large-Scale	12
Large-Scale	28
Medium-Scale	37
Small-Scale	22
Unspecified/Custom	2
N/A	5

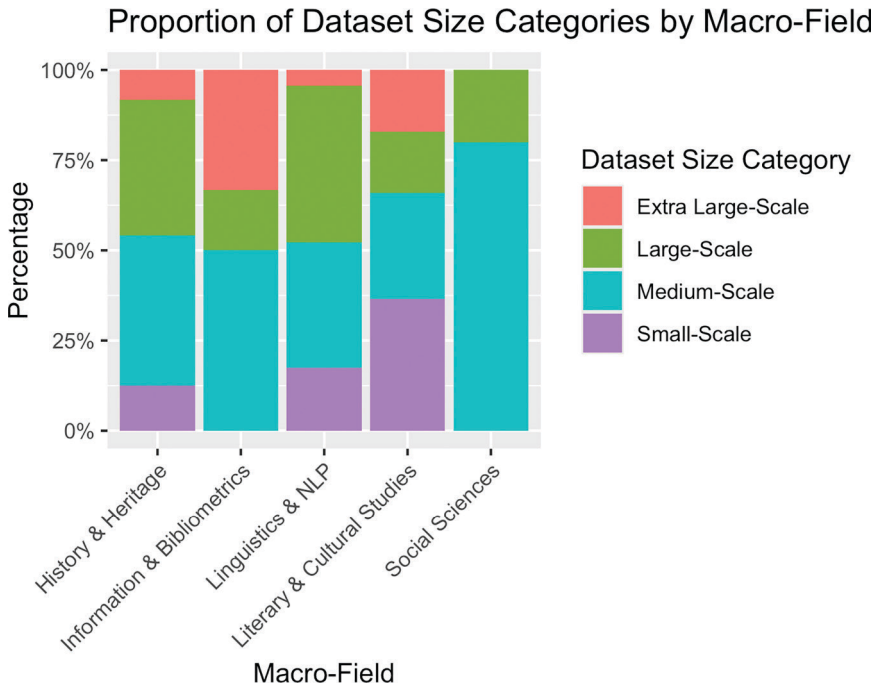


Figure 1.3 Bar chart showing the proportion of dataset size categories by macro-field.

The size of the dataset can also be assessed based on the size of the files composing the datasets. Whenever possible, we retrieved this information from the size of the files in the repositories where the datasets were deposited. We should note that repository practices and technical limitations can influence file size reporting and that in some cases datasets are split across multiple files, archived in compressed formats, or stored in repositories that do not readily display total file sizes, making it difficult to obtain accurate measurements. This additional metric, however, helped us infer the approximate scale of datasets when explicit quantitative details were lacking in the data papers. The way datasets are described in terms of size tends to be associated with their actual file sizes, as confirmed by the results of a Pearson’s Chi-square test with a simulation with 10,000 replicates ($p < 0.05$). However, a closer look at the distribution of the dataset size categories vs. the macro-categories based on the data paper’s text (Figure 1.4) shows that file size alone is not always a reliable proxy for dataset complexity or analytical utility. In other words, the narrative descriptions of dataset size do not always align with measurable file size categories (for example, Medium-scale datasets include a notable proportion of very small file sizes), which might reflect differences in how size is conceptualised or presented in the data paper. An explanation of this may be that authors sometimes emphasise the significance or richness of a dataset based on

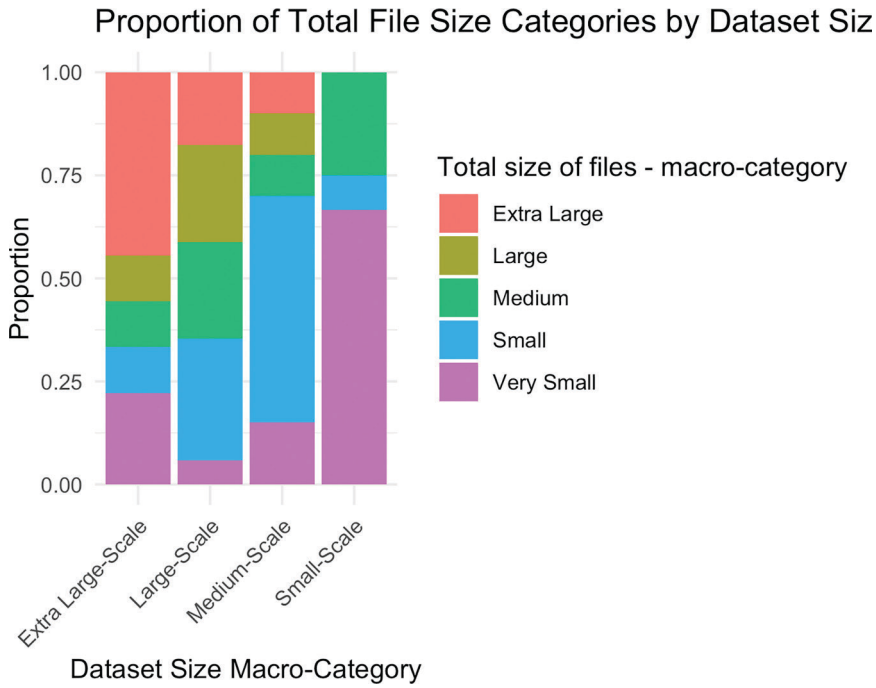


Figure 1.4 Proportion of total file size by dataset size macro-category based on the data papers' text.

qualitative factors, such as the uniqueness or granularity of the data, rather than its sheer volume or file size.

In order to better understand the nature and composition of the datasets, we annotated the technical structure or encoding of the dataset and categorised the datasets described in the data papers as follows:⁵ Textual (written language, such as corpora, transcriptions, or documents: 93 instances), Structured (tabular forms or databases, typically containing numeric, categorical, or relational data: 40 instances), Visual (images, illustrations, or other visual media: 4 instances), Geospatial (location-based data, often linked to maps or spatial analysis: 2 instances), Multimodal (explicitly combining two or more formats: 4 instances), and Auditory (sound recordings, including speech, music, or environmental audio: 2 instances). The same record may be associated with more of these categories. Textual and Structured data types are the most commonly co-occurring formats. In contrast, Visual, Geospatial, Multimodal, and Auditory datasets are far less common, both on their own and in combination with others. This suggests a dominance of text- and structure-based data in the analysed data papers, with less frequent inclusion of other media types.

Looking at the relation between dataset types and disciplines (Figure 1.5), we notice that Textual datasets are the most prevalent across almost all macro-fields,

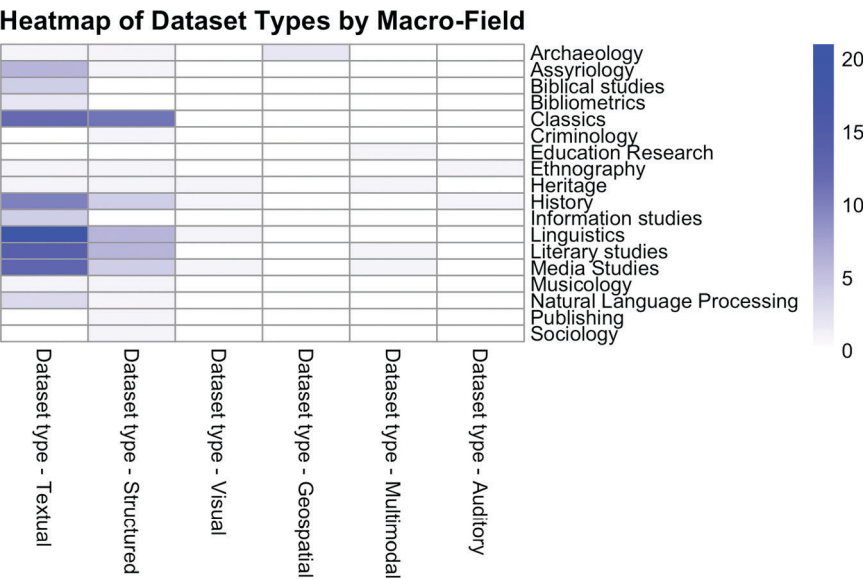


Figure 1.5 Distribution of dataset types across different academic disciplines represented in the analysed data papers.

with Linguistics, Literary studies, and Classics showing the highest counts. Structured datasets are common in Classics, Literary studies, and Linguistics. Geospatial datasets are uniquely found in Archaeology, which aligns with the nature of spatial data in this field; note that the dataset does not contain any papers in geography. Multimodal datasets appear in Media studies, Heritage, Literary studies, and Education Research, fields that often involve complex data types. Auditory datasets are rare but are present in Ethnography and History.

The distribution of dataset sizes varies considerably across different dataset types and macro-fields. As we can see in Figure 1.6, structured and textual datasets are the most widely used types across the macro-fields, with a dominance of medium-scale datasets, especially in fields like Archaeology, Assyriology, and Classics. However, there is greater diversity in dataset sizes within fields such as Information Studies, Linguistics, and Media Studies, where large-scale and extra large-scale structured and textual datasets are more common. Visual datasets tend to be predominantly large-scale or extra large-scale, with usage concentrated in fields like Heritage, History, and Publishing. Auditory and multimodal datasets appear infrequently and are associated mostly with medium-scale or large-scale sizes in specific fields. This pattern highlights how both the type and scale of datasets are shaped by the research needs and practices of each macro-field.

To contextualise the datasets within historical timelines and reveal disciplinary trends in temporal focus we reviewed each data paper to identify the primary

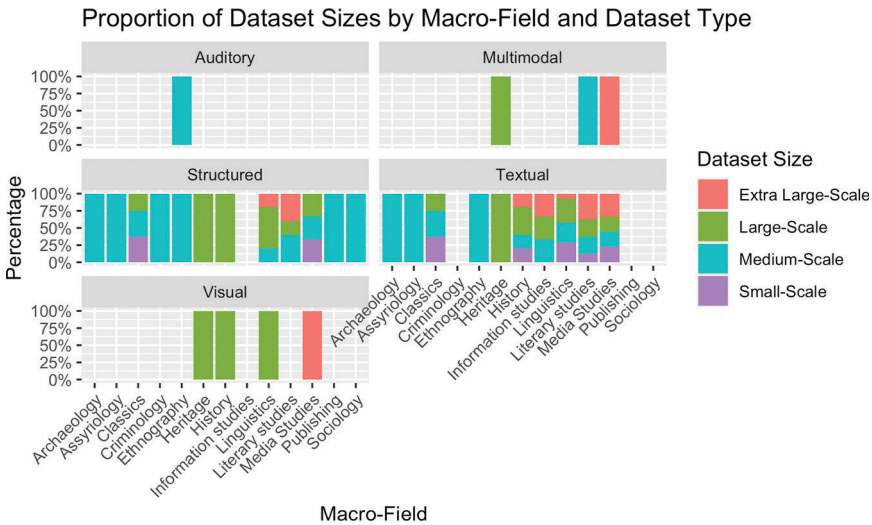


Figure 1.6 Proportion of dataset sizes by macro-discipline and dataset type.

temporal focus of the dataset. This was based on explicit information provided in the paper, such as the time period covered by the data, or inferred from contextual clues like the subject matter or the historical scope of the dataset. When a clear temporal period was stated, the dataset was assigned to one of the predefined macro-eras: Antiquity, Medieval Period, Early Modern Period, Modern Period, or Contemporary Period.⁶ If the dataset covered multiple distinct macro-eras or no clear timeframe was given, it was categorised as Varied or Unspecified, respectively. One case where no temporal information could be determined was marked as NA. Table 1.4 shows the distribution of this feature. The majority of datasets focus on the Contemporary period, with 34 instances, reflecting a strong interest in recent and current data. Antiquity is also well represented, with 27 datasets, highlighting the prominence of historical and archaeological research in the dataset collection, mainly due to the presence of the special collection “Representing the Ancient World through data”. The Early Modern period accounts for 23 datasets, while the Modern period includes 15 datasets. In contrast, the Medieval Period has significantly fewer datasets (3), suggesting a relative underrepresentation of this era.

Figure 1.7 illustrates how dataset sizes are distributed across different historical macro-eras. Datasets focused on the Medieval Period and Early Modern Period tend to be larger in scale, with a significant proportion categorised as large-scale and extra large-scale. In contrast, datasets associated with Antiquity and the Contemporary period display a more balanced distribution across all size categories, with a notable presence of small-scale and medium-scale datasets. This suggests that the temporal focus of datasets may influence their size, potentially

Table 1.4 Distribution of datasets by annotated macro-era

Macro-era	Number of data papers
Antiquity	27
Medieval Period	3
Early Modern Period	23
Modern Period	15
Contemporary	34
Unspecified	2
Varied	1
NA	1

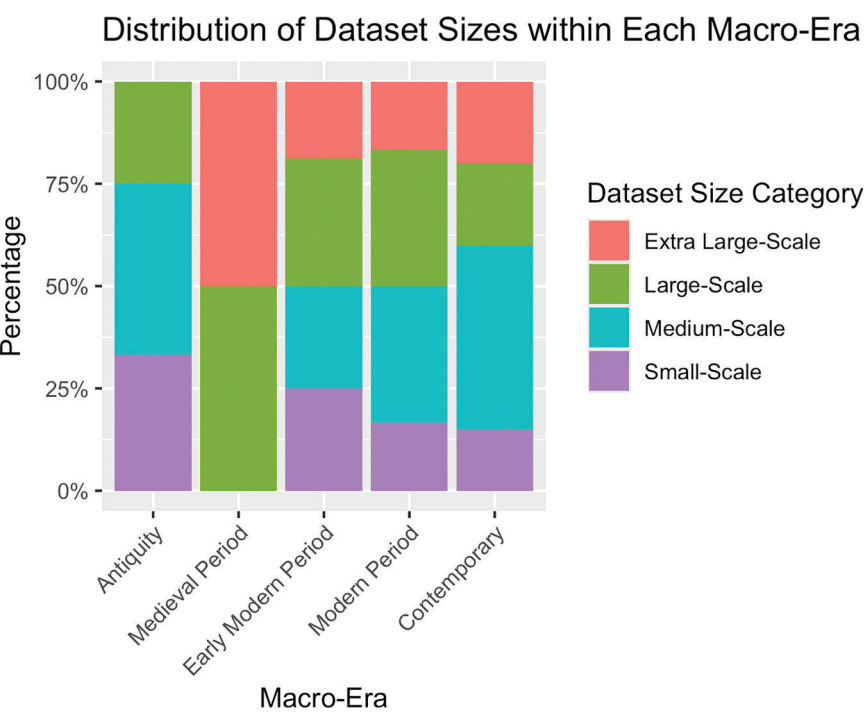


Figure 1.7 Distribution of dataset sizes across annotated macro-eras.

reflecting differences in data availability, research practices, or the scope of studies across historical periods.

To classify the datasets based on their geographic scope we analysed the geographical focus described in the data papers. If a dataset explicitly mentioned a specific region, it was classified accordingly. Datasets that spanned multiple

continents or did not specify a single primary region were categorised as Global. Europe accounts for the majority of datasets (49 out of the total). Global datasets are the second most represented category, with 18 datasets covering multiple or unspecified regions worldwide. Asia, the Americas, and the Middle East each have a modest number of datasets, ranging from seven to eight, while Africa has the fewest, with only five datasets. This distribution suggests a notable geographical imbalance in dataset coverage, with a clear overrepresentation of Europe compared to other regions. This overrepresentation can be partly attributed to the fact that JOHD's publisher is based in Europe, and JOHD tends to attract submissions from European scholars who, in turn, are more likely to work on European contexts or global phenomena framed through European academic perspectives. Moreover, this trend aligns with broader patterns observed in DH as a field. According to Syeda & Selva Babu (2024), academic output and influence in DH are heavily concentrated in specific regions (most notably the United States, the United Kingdom, and Italy), reflecting longstanding asymmetries in academic infrastructure and global research visibility. This centralisation is reinforced by strong institutional networks, preferential funding mechanisms, and the dominance of English-language publishing, all of which contribute to the marginalisation of contributions from underrepresented regions, and, possibly, of datasets comprising other regions of the world.

Our annotation of the datasets' method of generation aimed at providing insights into how the data were collected, processed, and curated, in order to better understand the level of human involvement and the technological approaches used in their creation. We categorised the datasets into three main types based on their described generation processes: Automated, where data was gathered or processed primarily through computational methods with minimal human intervention; Manual, where the data was collected, transcribed, or annotated entirely by humans, including crowdsourced efforts; and Hybrid, which combined automated processes with manual validation or refinement. In our sample manual methods of dataset generation are more prevalent than automated ones. Specifically, 93 datasets were generated using manual methods, compared to 75 datasets that involved automated methods. Additionally, many datasets use a combination of both manual and automated approaches, but a higher proportion of datasets relied on manual effort, either exclusively or in conjunction with automated techniques. This highlights the continued importance of human labour and expertise in dataset creation in humanities research, where judgment, validation, and complex annotation are required steps in the research process. It may also reflect disciplinary norms and constraints: in some fields, such as Classics, automatic methods are less commonly adopted or not easily applicable due to the nature of the sources and research questions (cf. Farina et al., 2025, p. 7–8). This also suggests that automation, while increasingly adopted, has not entirely replaced manual processes.

Looking at the temporal trends in the methods of dataset generation (Figures 1.8 and 1.9), we can see a clear upward trend in the adoption of automatic methods over time. In the earlier years of the timeline (2015 to 2018), the use of automatic generation methods was either limited or entirely absent. However, from 2019 onwards,

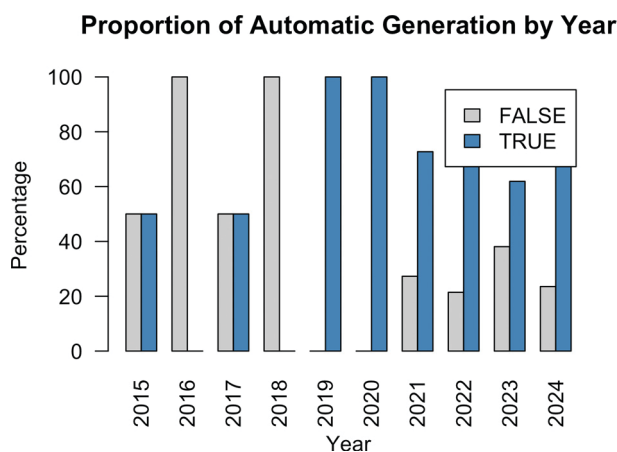


Figure 1.8 Number of datasets generated automatically by year.

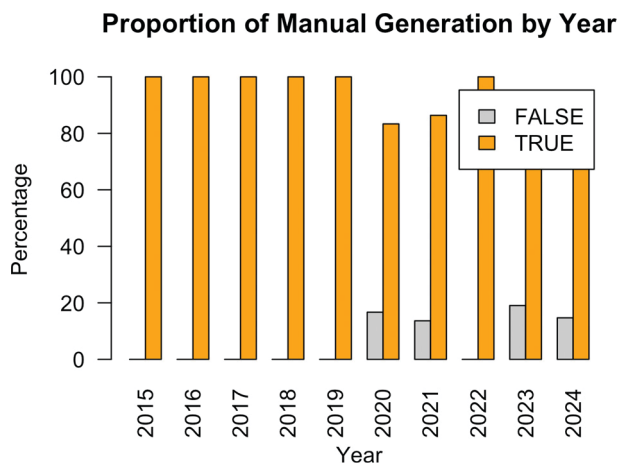


Figure 1.9 Number of datasets generated manually by year.

there is a marked increase in datasets created through automatic methods. By 2024, 26 datasets were generated automatically compared to just eight that were not, demonstrating a substantial shift toward automation. In percentage terms, the share of automatically generated datasets rose to nearly 75% in the most recent years, highlighting a growing reliance on automated techniques for dataset creation. At the same time, manual generation methods, including crowdsourcing and manual validation, remain consistently prevalent throughout the entire period. In most years, nearly all datasets involve some form of manual generation. Manual methods continue to play a

significant role in dataset generation, often in conjunction with automated processes, reflecting a trend toward hybrid approaches that combine the scalability of automation with the oversight and refinement provided by manual curation.

We analysed the distribution of data generation methods over time via two logistic regression models, one for the automatic generation method and one for the manual generation method. The models revealed no statistically significant temporal trends in the use of automatic ($p = 0.103$) or manual ($p = 0.263$) data generation methods across publication years. Although not significant, the odds of automatic generation decreased by approximately 15% per year, while the odds of manual generation increased by about 23% per year, suggesting a possible stabilisation or balancing of both approaches over time.

As expected, manual generation is dominant, particularly in small and medium-scale datasets, where it exceeds 85%. In contrast, extra-large and large-scale datasets have a higher proportion of automatically generated data compared to smaller datasets (Figures 1.10 and 1.11).

There are clear disciplinary differences in data generation practices. For automatic generation (Figure 1.12), the data shows that fields such as Literary & Cultural Studies and History & Heritage have the highest absolute counts of automatically generated datasets, with 31 and 18, respectively. However, when considering the proportions, Linguistics & NLP, along with Information & Bibliometrics, exhibit the greatest reliance on automatic methods. In Linguistics & NLP, approximately 75% of the datasets are automatically generated, reflecting the field's strong integration of computational approaches. Information & Bibliometrics also demonstrates

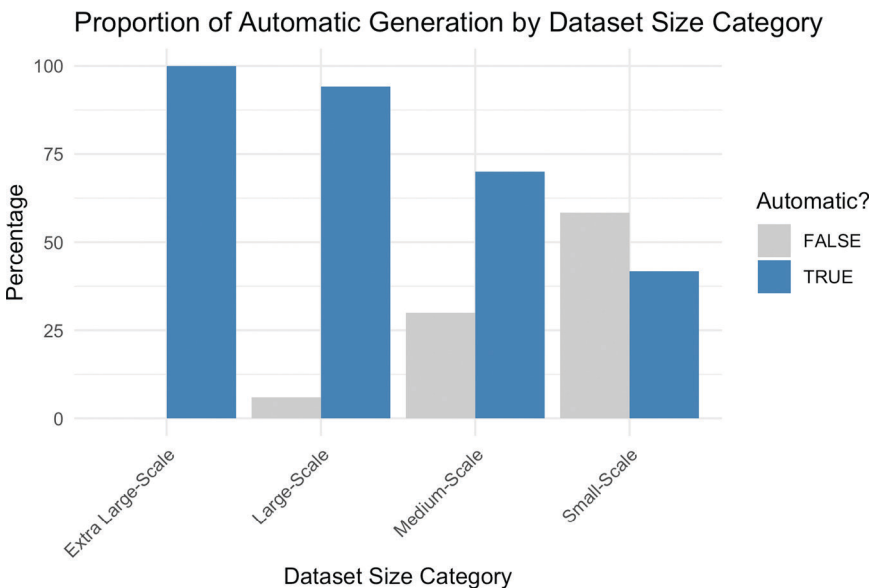


Figure 1.10 Proportion of automatic methods for data generation by dataset size.

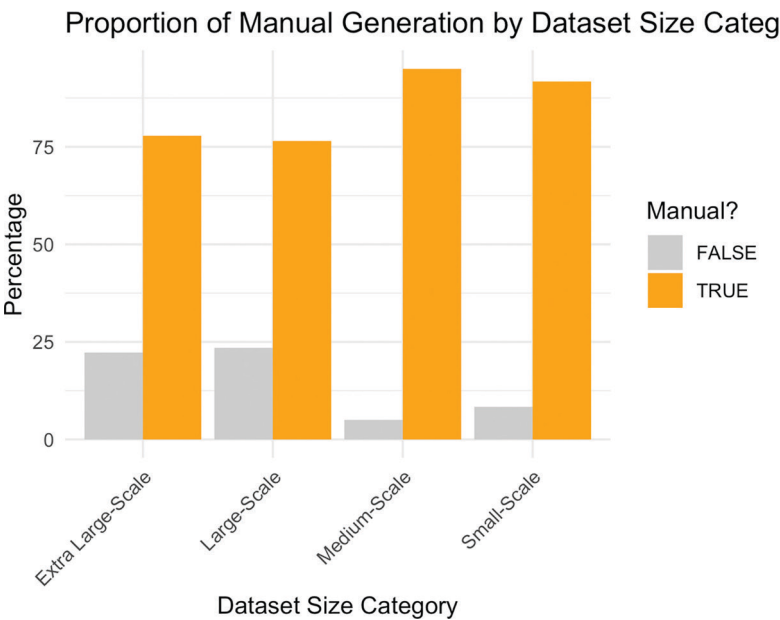


Figure 1.11 Proportion of manual methods for data generation by dataset size.

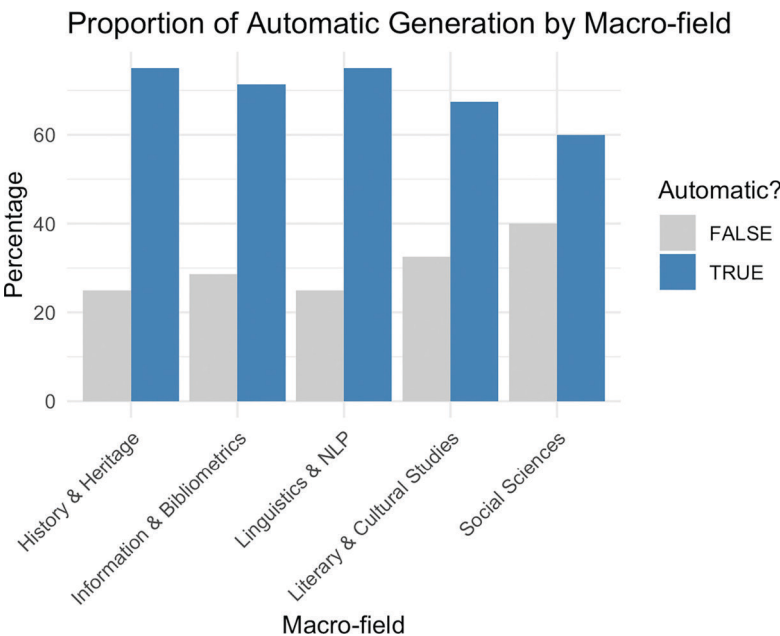


Figure 1.12 Proportion of automatic methods for data generation by macro-discipline.

a relatively high proportion of automatic generation, consistent with its data-intensive and quantitative research practices. On the other hand, Social Sciences show a more balanced distribution, with about 60% of the datasets generated automatically, suggesting a more cautious or selective use of automation.

The patterns are different when it comes to manual generation (Figure 1.13). Literary & Cultural Studies show a dominant preference for manual methods, with 42 manually generated datasets out of 46. This accounts for more than 90% of the datasets in that field, highlighting its traditional reliance on qualitative, interpretive methodologies. History & Heritage follows a similar trend, with 22 out of 24 datasets manually generated, again suggesting a preference for hands-on, manual approaches to data curation and annotation. Linguistics & NLP and Information & Bibliometrics present a more mixed picture but still demonstrate a significant reliance on manual methods in absolute terms. In Social Sciences, the pattern suggests a blend of both manual and automatic methods, though manual generation is still more common, representing about 80% of the datasets.

We annotated the degree and complexity of labels assigned to the datasets through the feature “Annotation level”. Basic annotation includes minimal meta-data such as file names, brief descriptions, or simple categorical labels, providing foundational information but lacking structured classification. Intermediate annotation incorporates additional metadata such as thematic keywords, category

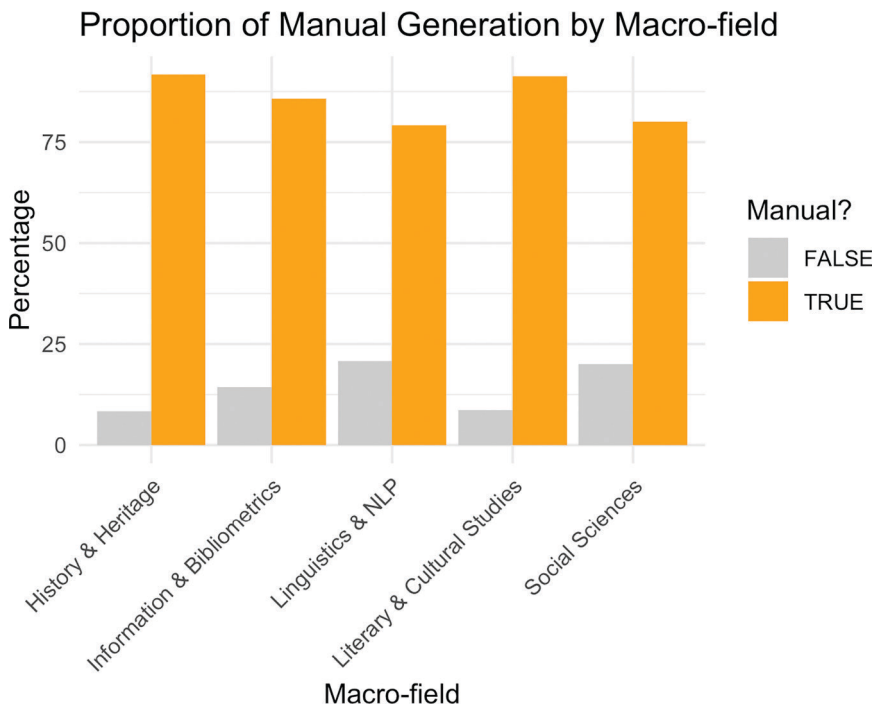


Figure 1.13 Proportion of manual methods for data generation by macro-discipline.

assignments, or general subject tags, offering a more organised dataset. Advanced annotation involves detailed linguistic tagging, semantic enrichment, structured ontologies, or deeper metadata layers, making the dataset highly structured and machine-readable.

The proportion of automatically generated annotations rises with annotation complexity, as shown in Figures 1.14 and 1.15. At the basic level, there is a more balanced contribution from both generation methods. However, as the annotation level increases, both manual and automatic generation become more prominent. Figure 1.15 shows a clear upward trend in the proportion of manual generation with higher annotation levels, indicating a growing reliance on human input for complex tasks. Figure 1.14 reflects a similar increase in automatic generation as annotation complexity rises. This suggests that while automation is increasingly employed for complex annotations, manual generation is not supplanted; rather, both approaches are utilised more extensively, with manual input playing a particularly critical role in ensuring the quality and accuracy of the most complex annotations.

Looking at the annotation level by discipline (Figure 1.16) we see that the complexity of annotation work is not uniform across disciplines, highlighting the potential influence of field-specific research questions and data characteristics on annotation practices. There is noticeable variation in the prevalence of different

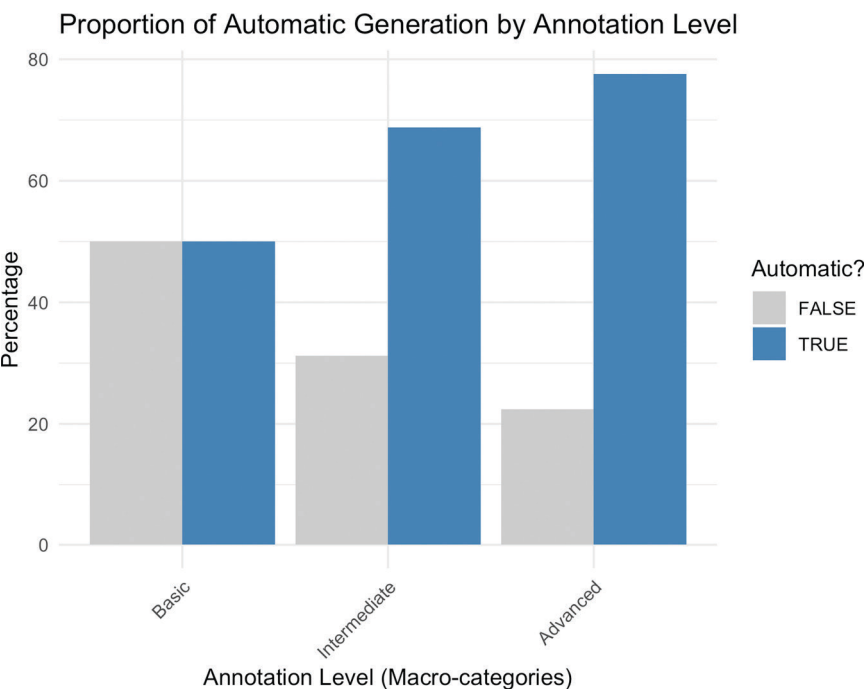


Figure 1.14 Proportion of automatic generation methods across different annotation levels.

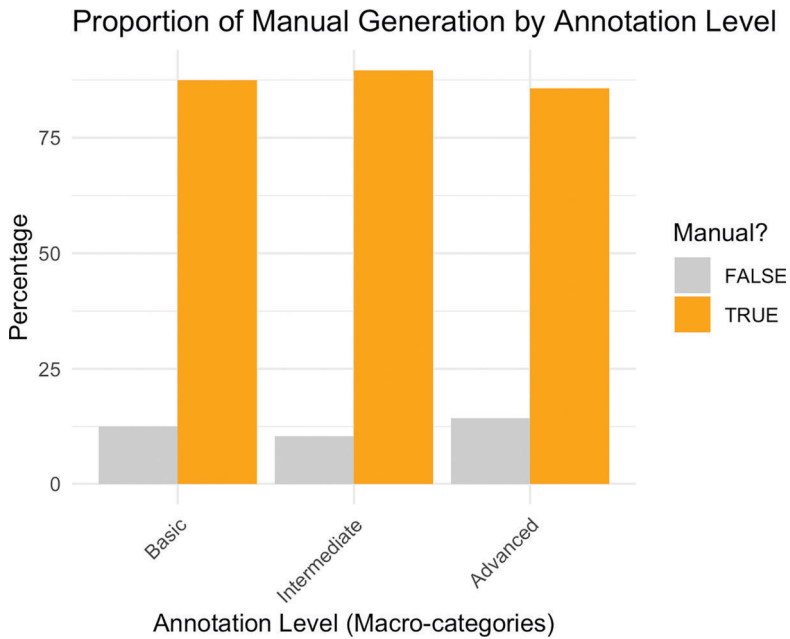


Figure 1.15 Proportion of manual generation methods across different annotation levels.

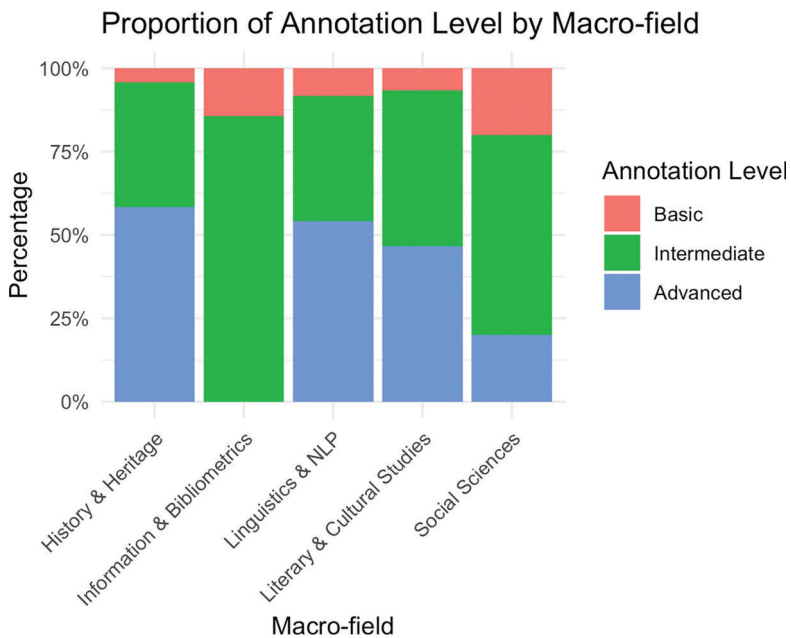


Figure 1.16 Proportion of annotation level by macro-field.

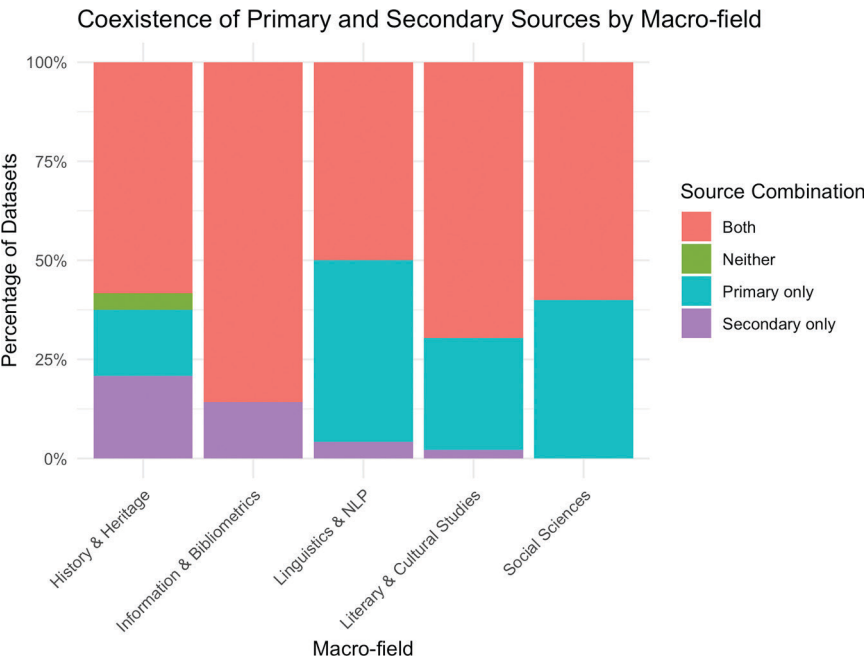


Figure 1.17 Coexistence of primary and secondary sources by macro-field.

annotation levels depending on the macro-field. Disciplines like Linguistics & NLP, and Information & Bibliometrics exhibit a high proportion of advanced annotations. Conversely, fields in the Social Sciences show a higher proportion of basic annotations.

We further annotated whether the datasets were generated from primary sources only, secondary sources only, or both. The way these source types are combined varies depending on the specific disciplines (Figure 1.17). Generally speaking, there is a common practice of integrating different source types across disciplines. However, History & Heritage exhibits a substantial proportion of datasets relying solely on primary sources, while Linguistics & NLP shows a higher concentration of datasets using both source types.

Finally, we analysed the datasets’ reusability for further research or applications, interoperability, and their potential for integration with smart data practices in terms of the use of advanced techniques to make the datasets more intelligent or interoperable. Reusability is a key factor in determining the value and impact of a dataset within the academic community. As expected, most datasets are openly available under licenses such as CC BY 4.0, CC0 1.0, and CC BY-NC-SA 4.0. These licenses ensure that the data can be freely used, shared, and adapted, which is essential for supporting collaborative research and innovation. Furthermore, several datasets highlight a commitment to interoperability, often achieved through standardised data formats (like CSV, XML-TEI, SQL), linked open data (LOD) principles, and

integration with established platforms and resources (e.g., GeoNames, Wikidata, Sketch Engine, Trismegistos). For example, many datasets are explicitly designed for training and evaluating NLP models, particularly in tasks like named entity recognition (NER), sentiment analysis, and historical language processing. Many datasets are also described as “machine learning-ready” or suitable for training and evaluating models.

1.4 Case Studies in Research on the Ancient World and Assyriology

Building on our quantitative analysis of data papers in JOHD, we now turn to practical examples from fields that have traditionally been more distant from smart data and AI practices. This section presents case studies demonstrating how such datasets are increasingly shaping research in the Ancient World and Assyriology, highlighting the integration of AI-driven methods and smart data approaches in these domains. We highlight smart data practices such as knowledge graphs, linked data, and predictive modelling, with case studies from the Ancient World. AI’s potential to address discipline-specific challenges, like automating annotation for large datasets, will be discussed, demonstrating its impact across various fields.

1.4.1 Data Dealing with the Past: Open Data, AI, and Ancient World Studies

As we discussed earlier in Section 1.1, in recent years, the humanities have increasingly embraced data-driven methodologies, leading to a growing interest in computational and quantitative approaches. This shift has also significantly influenced Ancient World (AW) studies, where structured datasets now play a crucial role in research. Scholars have long relied on curated data, including textual corpora, field reports, and collections of inscriptions and artifacts. However, the increasing availability of digital tools and computational resources has demonstrated the importance of ensuring not only high-quality research, but also open data sharing to facilitate reuse and reproducibility (Farina et al., 2024; 2025).

Unlike modern datasets that can be expanded through direct observation or experimentation, AW data is inherently fragmentary and fixed, making interoperability and standardisation crucial (Farina et al., 2024; 2025). In response, AI tools have emerged as powerful assets in addressing some of the field’s most persistent challenges, such as data scarcity and standardisation. Machine learning algorithms have been employed, for instance, to reconstruct damaged inscriptions, complete missing text or fragmented inscriptions, and enhance the decipherment of complex writing systems (e.g., Parker et al., 2019; Sommerschild, 2020; Chapman et al., 2021a; 2021b; Brusuelas, 2021; Assael et al., 2022; Vesuvius Challenge, 2024). These techniques not only increase the accuracy of textual reconstructions, but also reduce the manual effort required for labour-intensive philological work. Beyond text-based research, AI-powered tools, such as Pythia, a deep neural network trained to restore damaged Greek inscriptions (Assael et al., 2019; 2022), and DeepScribe, which applies computer vision to read cuneiform tablets (Williams et al., 2025), are revolutionising the study of ancient writing systems. At the same

time, computer vision enables large-scale artifact classification (Palladino & Bodard, 2023) and geospatial models help reconstruct historical landscapes and migrations (Holleran, 2021).

Despite the growing recognition of data-driven and AI-enhanced research in AW studies, formal venues for publishing AW datasets remain scarce. This lack of recognition extends to crucial work, such as corpus design, data curation, and the development of AI-driven models, mirroring, once again, ongoing discussions in STEM disciplines about the proper avenues for crediting datasets (see Section 1.2). In this evolving landscape, data papers have emerged as a crucial means of documenting dataset creation, methodology, and reuse potential (see Section 1.3). They ensure transparency and reproducibility and contribute to a more open and collaborative research environment. Notably, *Representing the Ancient World through Data* (2023–2024) has been JOHD’s most successful special collection so far, with 18 published papers focused on AW studies and over 19,000 views and downloads (Farina et al., 2024). This increasing engagement with data papers reflects a broader shift toward recognising the value of structured datasets, which are either already built upon or lay the foundations for AI-assisted methodologies. Despite this success, some gaps remain. Key datasets related to lesser-studied regions, certain material culture types, and lesser-documented inscriptions continue to be underrepresented. This is likely due to the limited availability of primary sources (a common challenge in AW studies), as well as some skepticism among more traditional scholars in the field (Farina et al., 2025). Furthermore, the relatively small number of published AW data papers underscores the need to encourage broader participation and promote interdisciplinary collaboration.

Nonetheless, as AI-driven methodologies continue to evolve, they are fundamentally transforming the study of ancient texts, allowing for more efficient and scalable analysis. Data papers in JOHD (Gorman, 2020; Ribary, 2020; Vierros & Henriksson, 2021; Mambrini & Passarotti, 2023; Palladino et al., 2023; Tresoldi, 2023; Dexter et al., 2024; Romanello & Najem-Meyer, 2024; Fendel, 2024) have demonstrated how Natural Language Processing (NLP) tools, syntactic tagging, and machine learning models are used or can be used to significantly enhance our understanding of ancient languages and literary corpora. These studies showcase how AI technologies are or can be applied to automate tasks that were once labour-intensive, such as lemmatisation, parsing, and translation alignment. For instance, Ribary (2020) presents a relational database that organises the text of Justinian’s *Digest* (533 CE) into a machine-readable format. The project integrates multiple layers of legal data: the compositional structure of the *Digest*, the network of legal concepts, and the logical framework of legal rules. Through Python scripts and SQL-based relational models, the study transforms raw texts into an interconnected database that enhances legal texts analysis. The database further facilitates AI-assisted legal research as it allows automated keyword searches, concept mapping, and term frequency analysis. It supports multilingual exploration, linking Latin and Ancient Greek textual segments with English metadata. Similarly, Palladino et al. (2023) describe a structured dataset containing Ancient Greek texts manually aligned with translations in Latin, English, and Portuguese. This dataset serves as

a gold standard for evaluating automatic translation alignment models and training AI-based systems. The research employed the Ugarit Translation Alignment Editor to create high-quality word-level alignments, with an inter-annotator agreement above 80%. Farina (2023) further explores the role of structured data in AI applications, by compiling a manually curated dataset of Ancient Greek and Latin words related to the semantic field of SEA. This dataset can serve as a training set for Large Language Models (LLMs) or NLP pipelines, enabling automatic semantic disambiguation between literal, metaphorical, and metonymic meanings of nouns, adjectives, and verbs. Other studies focus on data standardisation for improved interoperability. Bilby et al. (2023) and Bilby and BeDuhn (2023) discuss how their datasets can be transformed into TEI-XML, a standardised format that enhances accessibility and ensures compatibility with other DH tools. Romanello and Najem-Meyer (2024) apply NLP techniques to historical scholarly texts. This project focuses on digitised commentaries on Sophocles' *Ajax*, written in English, German, and French, but also containing extensive Latin and Greek quotations. These commentaries present unique challenges for automated text processing due to their specialised terminology, complex bibliographic references, and frequent OCR errors. To address these issues, the researchers developed a Named Entity (NE) corpus, annotating both traditional entities (persons, locations, works) and bibliographic references. The dataset, containing over 111,000 tokens and 7,300 entity mentions, was designed to support NE Recognition (NER) and entity linking (EL) in historical texts. The study found that NER models performed well, with F1-scores exceeding 80% in multiple languages. However, entity linking remained a difficult task due to OCR errors, abbreviation-heavy writing styles, and ambiguities in historical references, showing that there is still significant progress to be made in fully automating entity linking in historical texts. The corpus was used in the HIPE-2022 shared task, where it served as a benchmark for testing NLP systems' ability to handle noisy, domain-specific multilingual texts.

Other opportunities for integrating datasets and methods across domains – such as text, legal analysis, and historical geography – are emerging. These advances in AI-driven text analysis not only enhance individual tasks like translation alignment and entity recognition, but also pave the way for a more integrated approach in AW studies. The combination of datasets from different fields allows AI technologies to enable a more holistic understanding of the AW. A key development facilitating this interdisciplinary approach is the growing integration of Linked Open Data (LOD) and knowledge graphs. These technologies, powered by persistent identifiers and geospatial frameworks, are crucial for ensuring interoperability and enriching analyses. Several data papers in JOHD (e.g., Gorman, 2020; Ribary, 2020; Vierros & Henriksson, 2021; Farina, 2023; Mambrini & Passarotti, 2023) highlight the power of linking datasets through persistent identifiers and geospatial frameworks, enabling a more interconnected and comprehensive understanding of the AW. For instance, Ribary (2020) explores the integration of Roman law datasets with external resources like Papyri.info (Sosin, 2010) and Trismegistos (Depauw & Ghedof, 2014), facilitating legal and historical analyses that span multiple domains. The dataset described in Farina (2023) is also designed for interoperability, allowing

seamless integration with other digital resources. It employs persistent identifiers, such as WordNet synsets, to enhance searchability and ensure precise semantic classification. The dataset's utility could be further expanded by linking it to geographic resources like Pelagios (Simon et al., 2012) and the World Historical Gazetteer (Manning & Mostern, 2015), which would provide spatial contextualisation and enable its integration into LOD frameworks. The *LiLa Lemma Bank*, developed by Mambrini and Passarotti (2023) further emphasises this push for standardised, interconnected datasets. The dataset is designed to standardise and interconnect Latin lexical resources using the principles of Linguistic LOD. Developed within the ERC project *LiLa: Linking Latin* (Passarotti et al., 2020), the dataset provides a structured repository of over 215,000 canonical forms (lemmas), each assigned a stable Uniform Resource Identifier (URI). Through ontologies such as OntoLex, the *LiLa Lemma Bank* facilitates interoperability across diverse linguistic datasets, making it a crucial resource for projects engaged in Latin lemmatisation, lexicography, and NLP applications. The dataset's foundation is the morphological analyser *LEMLAT 3.0*, which compiles Latin lexical items from Classical and Medieval sources, including dictionaries, glossaries, and onomasticons. The lemmas undergo continuous refinement through manual revision and automated linking with external corpora, ensuring accuracy and completeness. Stored in both RDF and SQL formats, the dataset is accessible via Zenodo and GitHub, making it available for reuse across different computational frameworks.

The integration of AI and data-driven methodologies into AW studies will not only improve efficiency, but also reshape the ways scholars engage with historical data. The increasing adoption of LOD and structured datasets signals a shift toward greater collaboration, standardisation, and accessibility in the field. The following section specifically deals with the adoption of structured data for an underrepresented field (cf. Section 1.3) in AW studies, i.e., Assyriology.

1.4.2 The Use of AI and Computational Methods in Assyriology

As shown in the previous section, over the past decade, AI has emerged as a transformative tool in various disciplines of Ancient World studies, and Assyriology is no exception. The study of the history, language, and culture of the civilisations that inhabited the Ancient Near East, such as Sumer, Akkad, Babylonia, and Assyria, has fully embraced computational methods and AI as a means to decipher, reconstruct, and analyse vast collections of cuneiform tablets. As a result, it is not surprising to see Assyriology as the second field in Ancient World studies – after Classics – to have published the most with the Journal of Open Humanities Data (JOHD). It is therefore worth exploring the reasons behind this successful adoption, the methodologies being used, and the tools that have been created so far.

Cuneiform, one of the earliest known writing systems, is extremely complex. It consists of wedge-shaped marks inscribed on mainly clay tablets, with hundreds of variations in sign forms across different periods and regions (Finkel & Taylor, 2015, p. 6). Beyond the initial challenge of identifying the correct signs, cuneiform tablets are often poorly preserved, making them even harder to read. Furthermore,

many of these tablets were excavated without modern archaeological methods, leaving little information about their original context. This lack of provenance makes their interpretation especially challenging for Assyriologists. To complicate matters further, it is still not known how many cuneiform tablets there are, and often related fragments are found in different institutions.

With all these challenges, it is no surprise that Assyriologists were quick to turn to digital storage and computational methods to better organise and analyse the vast collection of cuneiform tablets and fragments. From this initial phase of exploration come what are now foundational tools for the discipline, above all the Cuneiform Digital Library Initiative (CDLI) which has to this date collected 360,000 cuneiform inscriptions (CDLI contributors, 2025) of an estimated total of 500,000 cuneiform inscriptions housed in public and private collections worldwide (Streck 2010). This database serves as the first port of call for any Assyriologist, and provides them with basic information on their tablet of study (photo, publication number, provenance, date, etc.). As a natural complement to the CDLI, the Open Richly Annotated Cuneiform Corpus (ORACC) was created soon after. This open-access digital platform provides scholars with fully annotated cuneiform tablets (Tinney & Robson, 2019). Another key digital resource from this initial digital adoption was the Electronic Text Corpus of Sumerian Literature (ETCSL), a digital project that still provides access to a vast collection of Sumerian literary texts. To this day, it serves as an important tool for studying Sumerian language and literature, along with the Electronic Pennsylvania Sumerian Dictionary (EPSD) available on ORACC. Lastly, worth mentioning is also the Base de Datos de Textos Neosumerios (BDTNS), a specialised database dedicated to Neo-Sumerian administrative and economic texts, one of the largest cuneiform collections known and primarily dated the Ur III period (c. 2112–2004 BCE). It is one of the most important resources for studying the economic and bureaucratic systems of the Ancient Near East.

Following this initial phase of exploration, Assyriology has mainly focused on expanding these platforms. Since its initial launch, ORACC has expanded to include various sub-projects that focus on different aspects of cuneiform studies, such as lexical texts, royal inscriptions, divination, and administrative records. It provides transliterations, translations, grammatical analyses, and metadata for cuneiform tablets in multiple languages, including Sumerian, Akkadian, and others. For a while, an Akkadian literary repository equivalent to ETCSL did not exist, until the Sources of Early Akkadian Literature (SEAL) database was made available starting from 2007. This database has grown to include over 600 Akkadian compositions as of this year (Streck & Wasserman 2025).

The advent of the Electronic Babylonian library (eBL) in 2020 marked an important transition in the use of computational tools in Assyriology. On one hand, it followed past digital projects by providing an annotated repository of all known Akkadian compositions, similar to the SEAL project. On the other hand, it has harnessed the power of AI to create an interactive platform called the “Library”, which helps reconstruct cuneiform fragments. In essence, the “Library” is a comprehensive digital database of cuneiform fragments, which uses AI to match and align

fragments, facilitating the identification of overlapping sections among numerous manuscripts and allowing for the reconstruction of new tablets and compositions.

In many ways, the Assyriological articles published in JOHD reflect this transition from traditional digital collection methods to more innovative methods using AI to reconstruct tablets, such as eBL. In fact, the first results of the eBL project, i.e., a dataset containing around 25,000 tablets with over 350,000 lines of transliterated text were published with JOHD (see Cobanoglu et al., 2024). At the same time, JOHD has also published more traditional digital methodologies, such as Mohr's (2021) work on geolocating cuneiform tablet collections in the United States and Clark and Gordin's (2023) gazetteer of the Uruk Urbanscape using toponyms found on administrative texts mostly from the first millennium BCE archives of Uruk.

The shift to AI is slowly taking place in Assyriology, and other articles published in JOHD highlight this transition, for example, Ong and Gordin's (2024) study of metaphor construction in the Neo-Assyrian letters using an AI trained language model. Similarly, Ong (2024) has applied another language model to the Neo-Assyrian letters corpus and trained it to fully annotate the morphological and syntactical variations of 2600 letters. The shift to AI is also evidenced by two ongoing projects in the field of Assyriology, namely DeepScribe and Machine Translation and Automated Analysis of Cuneiform Languages (MTAAC). The first is a project that applies AI to facilitate the reading and transcription of cuneiform signs on tablets. Recently, a successful pilot study was conducted by the project on deciphering Elamite cuneiform tablets (Williams et al., 2025). The latter project is an initiative that applies natural language processing, machine translation, and computational linguistics to the study of Sumerian texts, specifically the administrative texts of the Ur III period. Once again, the objective of this project is to train an AI machine to read cuneiform signs and automatically translate the vast amount of cuneiform tablets present from this period.

As we have seen, the integration of AI is also happening in smaller fields of AW studies, such as Assyriology. That said, this integration comes with its own set of unique challenges. First and foremost is the issue of standardisation. Cuneiform tablets are often incomplete and incredibly varied and thus training AI models on such a dataset is extremely difficult. It is therefore not surprising that we see the first application of AI in this field on very standardised corpora, such as the Neo-Assyrian letters and Ur III administrative texts. As highlighted in the data papers of Ong (2024) and Ong and Gordin (2024), even on these very standardised texts, an element of human annotation and fact-checking is still required, suggesting that a fully automated AI approach to Assyriology is still a long way away. Second is the issue of access. Many cuneiform tablets reside in museums and private collections, making it difficult for AI models to access comprehensive datasets. For this reason, the foundational tool for cuneiform studies, the CDLI project, has worked to become more and more interoperable and accessible to other platforms. For example, it is now possible to access geo-location data on cuneiform tablets collected by the GLOW project (Rattenborg et al., 2021) on the CDLI database.

Other features the CDLI includes are hyperlinks to transliterations and translations of tablets on the ETCSL, eBL, and ORACC platforms.

Looking ahead, AI is expected to play an even more prominent role in Assyriology in the future. Although AI is not yet a fully autonomous tool in Assyriology, as machine learning models improve and datasets expand, AI-powered translation and transcription tools will become increasingly reliable. These tools will help resolve the main issues that lie at the basis of Assyriological inquiry: the location, reconstruction, and decipherment of cuneiform tablets. The continued use of AI promises to solve these issues, reshaping the study of the Ancient Near East.

1.5 Conclusion

Pulling these points together, the landscape of open data in DH reflects an evolving balance between traditional scholarly practices and technological innovation. In recent years, there has been increasing reliance on automated methods for data generation, signalling a shift toward more efficient, scalable, and systematic workflows. However, manual generation remains vital for creating and curating datasets, ensuring quality control, domain expertise, and contextual sensitivity central to humanities research. Automation complements rather than replaces manual methods, creating workflows that are both robust and responsive to humanities data complexities. Distinct disciplinary patterns reveal how data generation preferences differ across DH fields. Computational and data-driven disciplines adopt automatic methods more readily, while traditionally qualitative fields favour manual approaches. Even where automation is common, manual work remains essential, highlighting a hybrid workflow that combines automation's speed and scale with human expertise's precision and interpretive power. These trends reflect ongoing transitions as fields integrate new technologies while maintaining established scholarly practices, with leading examples in the Ancient World and Assyriology (Sections 1.4.1 and 1.4.2).

The increasing adoption of smart data practices further characterises this transition. These practices emphasise strategic use of both automated tools and human expertise to improve data quality, usability, and interoperability. They involve thoughtful decision-making at every data lifecycle stage, from collection and annotation to sharing and reuse, ensuring datasets are not just large or complex but meaningful, well-documented, and ethically managed. In DH, such practices are crucial for managing the sophisticated annotation processes that balance efficiency with interpretive depth. The presence of intermediate and advanced annotation levels across datasets shows a growing commitment to detailed, high-quality curation while allowing simpler processes where appropriate. The interdisciplinary nature of open data adoption in DH is evident in the dataset composition analysed here. Literary and Cultural Studies, Linguistics and NLP, and History and Heritage lead, reflecting the broad relevance of open data practices across fields. This diversity supports cross-disciplinary collaboration and methodological integration, enriching research. As researchers adopt smart data practices, they enhance not

only their own fields but also the broader DH ecosystem, enabling more connected, reusable, and impactful research.

Additional drivers for open data publication include funder mandates and recognition that open data increases downloads and citations. Open datasets also offer greater potential for unforeseen reuse. However, datasets cannot be reused unless findable. Without findability, other FAIR principles – accessibility, interoperability, and reusability – do not apply. Discovery metadata and common standards help, but more is needed. JOHD adds a further layer supporting the first FAIR principle, findability. JOHD requires authors to describe their dataset's construction, nature, reuse suggestions, and clear download and unpacking instructions. The dataset must have at least one fully open, non-proprietary version with an open licence and be deposited in a sustainable repository to ensure long-term accessibility and reusability. Ethical approval is required for human subject data: once published, the data paper openly describes and points to the dataset, directing traffic to it and increasing reuse potential. This enables scholarship to build on scholarship, with the publication model as a cornerstone of humanities knowledge production.

The predominance of medium-scale datasets suggests a balance between comprehensiveness and manageability, reflecting a practical approach to humanities data sharing. The widespread use of textual and structured datasets aligns with DH's historical text-analysis focus, while visual, geospatial, multimodal, and auditory datasets remain underrepresented, revealing opportunities for growth. Embracing smart data practices in these emerging areas would promote diversity and inclusivity, enabling richer analyses. DH data sourcing reflects a comprehensive approach combining primary and secondary sources, enriching datasets by integrating original collection with existing resources, and reinforcing smart data principles of transparency, provenance, and responsible reuse.

At the forefront of these developments, JOHD has emerged as a major driver of open humanities research. Over the last five years, JOHD has expanded readership, submissions, and impact, establishing itself as a leading venue for data papers that promote transparency, reproducibility, and collaboration. Supporting a scholarly ecosystem grounded in smart data practices, JOHD enhances visibility and accessibility of DH datasets, fostering a more open, ethical, and innovative research culture. Continued advocacy, editorial innovation, and community engagement will further cement JOHD's role as a cornerstone of open humanities scholarship.

Despite JOHD's progress, challenges remain. Datasets from some regions, languages, and media types, especially outside Europe and North America and non-textual formats, are underrepresented. Disciplinary asymmetries persist, with computationally intensive fields more prominent and contributions from less digitally mature disciplines limited. While FAIR principles and reusability are emphasised, more could be done to support articulation of research questions and scholarly value beyond data sharing. Editors and researchers recognise the importance of promoting best practices alongside identifying growth areas, such as strengthening support for underrepresented fields, clarifying data quality criteria, and encouraging reflexive engagement with the epistemologies of data work. Addressing

these openly will ensure data journals in the humanities remain a critical, inclusive platform advancing data-driven humanities scholarship.

Authors' Contributions

All authors read and approved the final version of the manuscript. BMcG conceived the study and was responsible for conceptualisation, methodology, data curation, drafting Sections 1.2, 1.3, and 1.5 of the original manuscript, reviewing and editing the text, and supervision; she also led the annotation process, annotated 72 data papers in the dataset, and conducted the quantitative analysis. DB annotated 12 data papers in the dataset and drafted Section 1.4.2. AF contributed to methodology, annotated 22 data papers in the dataset, drafted Section 1.4.1, and participated in reviewing and editing the manuscript. SM drafted Sections 1.1 and 1.5 and contributed to reviewing and editing. XW participated in reviewing and editing the manuscript.

Notes

- 1 Journal of Open Humanities Data, <https://openhumanitiesdata.metajnl.com/>.
- 2 Research Data Journal for the Humanities and Social Sciences, <https://brill.com/view/journals/rdj/rdj-overview.xml>.
- 3 Detailed analysis of data papers, <https://figshare.com/s/0e8b238ad584bec75e2d>.
- 4 While we did not systematically evaluate the persistence of dataset links, we acknowledge that this is a critical aspect of open data publication, particularly from a digital preservation perspective. Anecdotally, we observed that datasets hosted by recognised platforms such as Ubiquity Press and DANS generally remained accessible, while some links hosted on project websites or less formal repositories appeared to be broken or redirected, which supports the importance of depositing datasets in trusted, long-term archives that support persistent identifiers and robust preservation infrastructure.
- 5 These categories were developed inductively based on the characteristics of the datasets and our research aims. While existing metadata schemas such as Dublin Core or the DataCite schema include fields for resource type, we found them insufficiently granular or not well-suited to the diverse data formats in our dataset.
- 6 We have adhered to the following periodisation: Antiquity (from the beginning of recorded human history to the fall of the Western Roman Empire in 476 CE), Medieval Period (from 476 CE to the beginning of the Renaissance around the 14th century), Early Modern Period (from the end of the Medieval Period around 1500 CE to the late 18th century), Modern Period (from the late 18th century to the end of World War II in 1945) and Contemporary Period (from the end of World War II to the present day).

References

- Alaimo, C., Kallinikos, J., & Aaltonen, A. (2020). Data and value. In Nambisan, S., Lyytinen, K., & Yoo, Y., (Eds.), *Handbook of Digital Innovation*, 162–178. Cheltenham/Northampton: Edward Elgar Publishing.
- Assael, Y., Sommerschild, T., & Prag, J. (2019). Restoring ancient text using deep learning: A case study on Greek epigraphy. In *Proceedings of the 2019 Conference on*

- Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 6368–6375. Hong Kong, China: Association for Computational Linguistics.
- Assael, Y., Sommerschild, T., Shillingford, B., Bordbar, M., Pavlopoulos, J., Chatzipanagiotou, M., Androutsopoulos, I., Prag, J., & de Freitas, N. (2022). Restoring and attributing ancient texts using deep neural networks. *Nature (London)*, 603(7900), 280–283. <https://doi.org/10.1038/s41586-022-04448-z>
- Bilby, M. G., & BeDuhn, J. D. (2023). BeDuhn’s Greek reconstruction of Marcion’s Gospel. *Journal of Open Humanities Data*, 9, 25. <https://doi.org/10.5334/johd.126>
- Bilby, M. G., Bull, J., & Lotharp, K. L. (2023). Normalized datasets of Zahn’s, van Manen’s, and Harnack’s Greek reconstructions of Marcion’s Apostolos. *Journal of Open Humanities Data*, 9(1), 21. <https://doi.org/10.5334/johd.122>
- Brusuelas, J. H. (2021). Scholarly editing and AI: Machine predicted text and Herculaneum papyri. *Magazén*, 2(1), 45–70.
- Candela, L., Castelli, D., Manghi, P., & Tani, A. (2015). Data journals: A survey. *Journal of the Association for Information Science and Technology*, 66(9), 1747–1762. <https://doi.org/10.1002/asi.23358>
- CDLI contributors. (2025, March 19). *About CDLI*. Cuneiform Digital Library Initiative. <https://cdli.earth/about>
- Chapman, C. Y., Parker, C. S., Bertelsman, A., Gessel, K., Hatch, H., Seevers, K., Brusuelas, J. H., Parsons, S., & Seales, W. B. (2021a). The digital compilation and restoration of Herculaneum fragment PHerc.118. Manuscript studies. *A Journal of the Schoenberg Institute for Manuscript Studies*, 6(1), 1–32. <https://doi.org/10.1353/mns.2021.0000>
- Chapman, C. Y., Parker, C. S., Parsons, S., & Seales, W. B. (2021b). Using METS to express digital provenance for complex digital objects. *Metadata and Semantic Research*, 1355, 143–154. https://doi.org/10.1007/978-3-030-71903-6_15
- Clark, S., & Gordin, S. (2023). The Mesopotamian Ancient Place-Names Almanac (MAPA): A gazetteer of the Uruk urbanscape in the age of empires. *Journal of Open Humanities Data*, 9, 20. <https://doi.org/10.5334/johd.146>
- Cobanoglu, Y., Laasonen, J., Simonjetz, F., Khait, I., Cohen, S., Földi, Z., ... Jiménez, E. (2024). Transliterated cuneiform tablets of the Electronic Babylonian Library platform. *Journal of Open Humanities Data*, 10, 19. <https://doi.org/10.5334/johd.148>
- Colavizza G., Hrynaskiewicz I., Staden I., Whitaker K., & McGillivray B. (2020) The citation advantage of linking publications to research data. *PLoS ONE* 15(4), e0230416. <https://doi.org/10.1371/journal.pone.0230416>
- Depauw, M., & Gheldof, T. (2014). Trismegistos: An interdisciplinary platform for ancient world texts and related information. In Bolikowski, L., Casarosa, V., Goodale, P., Houssos, N., Manghi, P., & Schirrwagen, J. (Eds.), *Theory and Practice of Digital Libraries — TPDL 2013 Selected Workshops*. TPDL 2013. Communications in Computer and Information Science, vol 416. Cham: Springer. https://doi.org/10.1007/978-3-319-08425-1_5
- Dexter, J. P., Chaudhuri, P., Burns, P. J., Adams, E. D., Bolt, T. J., Cásarez, A., ... Shumway, S. (2024). A database of intertexts in Valerius Flaccus’ *Argonautica* 1: A benchmarking resource for the evaluation of computational intertextual search of Latin corpora. *Journal of Open Humanities Data*, 10(1), 14. <https://doi.org/10.5334/johd.153>
- Dorta-González, P., & Dorta-González, M.I. (2023). Citation differences across research funding and access modalities. *The Journal of Academic Librarianship*, 49(4), 102734, <https://doi.org/10.1016/j.acalib.2023.102734>

- Farina, A. (2023). Lost at sea: A dataset of 25+ SEA words morpho-semantically annotated in Ancient Greek and Latin. *Journal of Open Humanities Data*, 9(24), 1–7. <https://doi.org/10.5334/johd.139>
- Farina, A., Marongiu, P., Bru, M., & Borkowski, D. (2025). When data meets the past: Data collection, sharing, and reuse in Ancient World studies. *Open Information Science*, 9, 1–20. <https://doi.org/10.1515/opis-2025-0014>
- Farina, A., Marongiu, P., & Rodda, M. A. (2024). Editorial: Representing the ancient world through data. *Journal of Open Humanities Data*, 10(57), 1–6. <https://doi.org/10.5334/johd.245>
- Fendel, V. B. (2024). The I. Sicily Sketch engine corpus. *Journal of Open Humanities Data*, 10, 56. <https://doi.org/10.5334/johd.258>
- Finkel, I. L., & Taylor, J. (2015). *Cuneiform*. London: The British Museum.
- Gorman, V. B. (2020). Dependency treebanks of Ancient Greek prose. *Journal of Open Humanities Data*, 6(1), 1. <https://doi.org/10.5334/johd.13>
- Holleran, C. (2021). *Mapping Migration in Roman Iberia*. <https://mappingromanmigration.exeter.ac.uk/index.html>
- Huang, C. K., Neylon, C., Montgomery, L. et al. (2024). Open access research outputs receive more diverse citations. *Scientometrics*, 129, 825–845. <https://doi.org/10.1007/s11192-023-04894-0>
- Jones, S. E. (2014). *The Emergence of Digital Humanities*. New York/London: Routledge.
- Longley Arthur, P., & Hearn, L. (2021). Toward open research: A narrative review of the challenges and opportunities for open humanities. *Journal of Communication*, 71(5), 827–853. <https://doi.org/10.1093/joc/jqab028>
- MacCallum, C. J., & Parthasarathy, H. (2006). Open access increases citation rate. *PLOS Biology*, 4(5), e176.
- Maddi, A., & Sapinho, D. (2022). Does open access really increase impact? A large-scale randomized analysis. *26th International Conference on Science and Technology Indicators IST 2022*, Granada (Spain), 7–9 September 2022. <https://arxiv.org/abs/2206.06874>
- Mahony, S. (2018). Cultural diversity and the digital humanities. *Fudan Journal of the Humanities and Social Sciences*, 11, 1–18. <https://doi.org/10.1007/s40647-018-0216-0>
- Mahony, S. (2022). Toward openness and transparency to better facilitate knowledge creation. *Journal of the Association for Information Science and Technology*, 73(10), 1–15. <https://doi.org/10.1002/asi.24652>
- Mahony, S., & Gao, J. (2019). Linguistic and cultural hegemony in the digital humanities. In *Proceedings of the Digital Humanities Congress 2018*, Digital Humanities Institute, University of Sheffield. www.dhi.ac.uk/openbook/chapter/dhc2018-mahony
- Mambrini, F., & Passarotti, M. C. (2023). The LiLa lemma bank: A knowledge base of Latin canonical forms. *Journal of Open Humanities Data*, 9(1), 28. <https://doi.org/10.5334/johd.145>
- Manning, P., & Mostern, R. (2015). World-Historical Gazetteer. <https://doi.org/10.17613/M6GQ0Q>
- Marongiu, P., Pedrazzini, N., Ribary, M., & McGillivray, B. (2025). Le Journal of Open Humanities Data (JOHD): enjeux et défis dans la publication de data papers pour les sciences humaines et sociales (SHS). In Schopf, J. & Kosmopoulos, C. (Eds.), *Publier, partager, réutiliser les données de la recherche: les data papers et leurs enjeux*. Lille: Presses Universitaires du Septentrion.
- Mayer-Schönberger, V., & Cukier, K. (2013). *Big Data: A Revolution that Will Transform How We Live, Work, and Think*. Boston (MA): Houghton Mifflin Harcourt.

- McGillivray, B., & Astell, M. (2019). The relationship between usage and citations in an open access mega-journal. *Scientometrics*, 121, 817–838. <https://doi.org/10.1007/s11192-019-03228-3>
- McGillivray, B., Marongiu, P., Pedrazzini, N., Ribary, M., Wigdorowitz, M., & Zordan, E. (2022). Deep impact: A study on the impact of data papers and datasets in the humanities and social sciences. *Publications*, 10(4), 39.
- Mohr, S. (2021). A bibliography of cuneiform tablet editions in United States colleges and universities through 2020. *Journal of Open Humanities Data*, 9, 3. <https://doi.org/10.5334/johd.27>
- Ong, M. (2024). A full morphosyntactic annotation of the State Archives of Assyria letter corpus. *Journal of Open Humanities Data*, 10, 30. <https://doi.org/10.5334/johd.202>
- Ong, M., & Gordin, S. (2024). A survey of body part construction metaphors in the Neo-Assyrian letter corpus. *Journal of Open Humanities Data*, 10, 10. <https://doi.org/10.5334/johd.142>
- Owens, T. (2011). Defining data for humanists: Text, artifact, information or evidence? *Journal of Digital Humanities*, 1(1).
- Palladino, C., & Bodard, G. (Eds.). (2023). *Can't Touch This. Digital Approaches to Materiality in Cultural Heritage*. London: Ubiquity Press.
- Palladino, C., Shamsian, F., Yousef, T., Wright, D. J., d'Orange Ferreira, A., & dos Reis, M. F. (2023). Translation alignment for Ancient Greek: Annotation guidelines and gold standards. *Journal of Open Humanities Data*, 9(1), 1–22. <https://doi.org/10.5334/johd.131>
- Parker, C. S., Parsons, S., Bandy, J., Chapman, C., Coppens, F., & Seales, W. B. (2019). From invisibility to readability: Recovering the ink of Herculaneum. *Public Library of Science, PLOS ONE*, 14(5). <https://doi.org/10.1371/journal.pone.0215775>
- Passarotti, M., Mambrini, F., Franzini, G., Cecchini, F. M., Litta, E., Moretti, G., Ruffolo, P., & Sprugnoli, R. (2020). Interlinking through lemmas. The lexical collection of the LiLa knowledge base of linguistic resources for Latin. *Current Approaches in Latin Lemmatization, Studi e Saggi Linguistici, LVIII*(1), 177–212. Passarotti M. (Ed.). <https://doi.org/10.4454/ssl.v58i1.277>
- Passarotti, M., Mambrini, F., & Moretti, G. (2024). The services of the LiLa knowledge base of interoperable linguistic resources for Latin. In *Proceedings of the 9th Workshop on Linked Data in Linguistics @ LREC-COLING 2024*, 75–83, Torino, Italia: ELRA and ICCL.
- Rattenborg, R., Johannsson, C., Smidt, G. R., Andersson, J., & Nett, S. (2021). An open access index for the geographical distribution of the cuneiform corpus. *Cuneiform Digital Library Journal*, 2021(1), 1–12. <https://cdli.mpiwg-berlin.mpg.de/articles/cdlj/2021-1>
- Ribary, M. (2020). A relational database of Roman law based on Justinian's Digest. *Journal of Open Humanities Data*, 6(1), 5. <https://doi.org/10.5334/johd.17>
- Romanello, M., & Najem-Meyer, S. (2024). A named entity-annotated corpus of 19th century classical commentaries. *Journal of Open Humanities Data*, 10(1), 1. <https://doi.org/10.5334/johd.150>
- Simon, R., Barker, E., & Isaksen, L. (2012). Exploring Pelagios: A visual browser for geo-tagged datasets. *International Workshop on Supporting Users' Exploration of Digital Libraries*, Paphos (Cyprus), pp. 1–6.
- Sommerschield, T. (2020). Restoring ancient texts using Machine Learning: A case-study on Greek and Latin epigraphy. *Papers of the British School at Rome*, 88, 387–388.
- Sosin, J. (ed.) (2010). Papyri.info. <http://papyri.info>
- Streck, M. P. (2010). Großes Fach Altorientalistik: Der Umfang des keilschriftlichen Textkorpus. *Mitteilungen der Deutschen Orient-Gesellschaft*, 142, 35–58.

- Streck, M. P., & Wasserman, N. (2025). *Sources of Early Akkadian Literature*: <https://seal.huji.ac.il/>
- Syeda, N., & Selva Babu, K. (2024). Exploring the role of digital humanities in the centralization of knowledge production: Clusters, networks, or echo chambers. *The Scientific Temper*, 15(4), 3358–3368. <https://doi.org/10.58414/SCIENTIFICTEMPER.2024.15.4.46>
- Terras, M. (2018). Digitization and digital resources in the humanities. In Warwick C., Terras M., & Nyhan J. (Eds.), *Digital Humanities in Practice*, 47–70. Cambridge: Cambridge University Press.
- Terras, M. (2022). Digital humanities and digitised cultural heritage. In O’Sullivan, J. (Ed.), *The Bloomsbury Handbook to the Digital Humanities*, 255–266. Bloomsbury: Bloomsbury Handbooks. <https://doi.org/10.5040/9781350232143.ch-24>
- Tinney, S., & Robson E. (2019). *About Oracc*. The Open Richly Annotated Cuneiform Corpus. <https://oracc.museum.upenn.edu/doc/about/aboutoracc/>
- Tresoldi, T. (2023). A Global Lexical Database (GLED) for computational historical linguistics. *Journal of Open Humanities Data*, 9(1), 2. <https://doi.org/10.5334/johd.96>
- Van den Eynden, V., Knight, G., Vlad, A., Radler, B., Tenopir, C., Leon, D., Manista, F., Whitworth, J., & Corti, L. (2016). *Survey of Wellcome Researchers and their Attitudes to Open Research*. <https://doi.org/10.6084/m9.figshare.4055448.v1>
- Vesuvius Challenge. (2024). *AI Reads Ancient Scroll Buried by Vesuvius Eruption*. <https://phys.org/news/2024-02-ai-ancient-scroll-vesuvius-eruption.html>
- Vierros, M., & Henriksson, E. (2021). PapyGreek treebanks: A dataset of linguistically annotated Greek documentary papyri. *Journal of Open Humanities Data*, 7(26), 1–6. <https://doi.org/10.5334/johd.55>
- Wigdorowitz, M., Ribary, M., Farina, A., Lima, E., Borkowski, D., Marongiu, P., Sorensen, A. H., Timis, C., & McGillivray, B. (2024). It takes a village! Editorship, advocacy, and research in running an open access data journal. *Publications*, 12(3), 24. <https://doi.org/10.3390/publications12030024>
- Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ... Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3(1), 160018. <https://doi.org/10.1038/sdata.2016.18>
- Williams, E. C., Su, G., Schloen, S. R., Prosser, M. C., Paulus, S., & Krishnan, S. (2025). DeepScribe: Localization and classification of Elamite cuneiform signs via deep learning. *arXiv:2306.01268*.