Journal of Open Humanities Data

Are We There Yet? Notes Towards Benchmarking an Experimental AI-Assisted Workflow for Humanities Data Cleaning and Reconciliation

COLLECTION: BENCHMARKING IN DIGITAL HUMANITIES

DISCUSSION PAPER

ERIN A. MCCARTHY

ubiquity press

ABSTRACT

This paper introduces a novel AI-assisted pipeline developed to prepare data for the European Research Council-funded project "STEMMA: Systems of Transmitting Early Modern Manuscript Verse, 1475–1700." Now approaching its midpoint, STEMMA develops and applies a data-driven approach to provide the first comprehensive study of the circulation of early modern English poetry in manuscript. The project began by aggregating and reconciling five of the largest and most authoritative existing datasets about early modern verse circulation. The sheer volume of data, along with the need to preserve early modern English spelling and scribal idiosyncrasies for later analyses, meant that off-the-shelf data cleaning tools like OpenRefine were not fit for purpose. To that end, our software developer created a staged pipeline to aid the removal of duplicates, creation of authorities, reconciliation, and assignment of unique identifiers.

The rapid and pragmatic way that this process was developed and deployed means that we did not take the time to benchmark it, nor is it feasible to do so retrospectively. However, this discussion paper records observations from this process and reflects on challenges and bottlenecks as well as opportunities. It points the way toward future benchmarks that are increasingly needed for novel applications of computational methods in the digital humanities. It also briefly considers the relationship between technical benchmarking and research project management.

CORRESPONDING AUTHOR:
Erin A. McCarthy
Discipline of English, School of English, Media, and Creative Arts, University of Galway, Galway, Ireland
erin.mccarthy@universityofgalway.ie

KEYWORDS:
data cleaning; data reconciliation; data harmonization; LLMs; benchmarks; project management

TO CITE THIS ARTICLE:
McCarthy, E. A. (2026). Are We There Yet? Notes Towards Benchmarking an Experimental AI-Assisted Workflow for Humanities Data Cleaning and Reconciliation. *Journal of Open Humanities Data*, 12: 46, pp. 1–14. DOI: https://doi.org/10.5334/johd.490

DOI: 10.5334/johd.490

(1) CONTEXT AND MOTIVATION

The European Research Council-funded project "STEMMA: Systems of Transmitting Early Modern Manuscript Verse, 1475–1700" (https://stemma.universityofgalway.ie) maps and models the movement of English poetry through early modern social networks. It focuses on manuscripts written and used between the introduction of printing with moveable type in England in 1475 and the end of the seventeenth century, by which time the Restoration had ushered in rapid changes in literary taste and publishing norms. The project includes manuscripts circulating in England and anywhere else English was spoken and read, including Ireland, the North American colonies, and overseas exile communities. Scholars have tended to treat manuscripts as case studies, but while they have identified discrete groups in which manuscript poems circulated, the paths texts took between these groups are less well known. Thus, while parts of manuscript culture have been described in great detail, it remains impossible for scholars to see how these separate parts fit together. By applying insights from network analysis and graph theory, the project aims to go beyond surveying and visualizing the evidence that has survived to make theoretical and mathematical inferences about what has not. Ultimately, the project will show that the emerging audience for English poetry to was more varied and more geographically dispersed, and yet also more interconnected, than traditional methods of literary historical research have allowed us to see.

Manuscript was an inherently social medium: unless manuscript copyists composed their own original works, they had to get copy text from someone else, and they often shared what they had copied onward. Scholars have, by turns, described the social groups that facilitated the creation and distribution of literary manuscripts as circles, clubs, communities, coteries, fellowships, fraternities, peer groups, spheres, and tribes; these groups coalesced around relationships characterized as affiliation, alliance, co-education, friendship, family, faction, patronage, program, sect, and 'commercial sociability' (Marotti, 1995; Scott-Warren, 2000; Smith, 2014; Millstone, 2016). May and Marotti have acknowledged that "most scribal culture… must have been produced outside the centers manuscript scholars have tended to credit with virtually all of it" (2014, p. 7), but they also lament that manuscripts that circulated outside these centers are much more likely to have been lost (see also May, 2004; May and Wolfe, 2010). STEMMA moves beyond case studies to investigate and computationally model how these smaller communities linked, overlapped, and shared literary texts, facilitating macro-level studies like those carried out on the early modern book trade and highlighting promising new areas for close material and textual analysis.

Early modernist scholars have long used the language of networks to describe how texts circulated within defined communities (Levy, 1982; Cust, 1986; Marotti, 1995; Woudhuysen, 1996; Beal, 1998; Ezell, 1999; Scott-Warren, 2000; de Groot, 2006; May & Wolfe, 2010). Recent studies have investigated the transmission of verse (de Groot, 2006), news (Raymond & Moxham, 2016), intelligence (Daybell, 2011; Akkerman, 2018), controversial pamphlets (Millstone, 2016), recipes (Strocchia, 2014), religious texts (Crawford, 2010), correspondence (Daybell, 2012), and even an individual's own notebooks (Vine, 2019). There has also been some consideration of provincial (May & Marotti, 2014), national (Verweij, 2016), sectarian (Hackett, 2012), and colonial (Wilcox, 2012) networks.

More recently, humanities researchers have begun to incorporate insights from network science into their analyses of early modern literature and culture. Malte Rehbein has applied network science in the editing of medieval manuscripts with multiple layers of text and commentary (Rehbein, 2014). Ruth and Sebastian Ahnert's studies of correspondence networks (including, *inter alia*, Ahnert & Ahnert, 2014; Ahnert & Ahnert, 2015) inspired sophisticated work by Evan Bourke on the Hartlib Circle (Bourke, 2017) and John R. Ladd on dedications in printed books (Ladd, 2021). Around the same time, the *Six Degrees of Francis Bacon* project used data from the *Oxford Dictionary of National Biography* to model the early modern social network, including statistically inferred relationships (Warren et al., 2016). Trends in scholarly publishing suggest that network analysis has reached an inflection point: the Université du Luxembourg began publishing *The Journal of Historical Network Research* in 2017; Blaine Greteman makes extensive use of data from the English Short-Title Catalogue to analyze networks of authors and stationers in the book trade (Greteman, 2021); and four leading scholars collaborated on *The Network Turn*, which argues for humanities researchers' deeper engagement with the mathematical and theoretical dimensions of networks (Ahnert et al., 2020).

However, as humanities researchers have begun to incorporate methods and insights from network science into their analyses of early modern literature and culture, the manuscript circulation of early modern verse has resisted this kind of study. (One exception to this is a paper by Greg Kneidel, Brent Nelson, and Kyle Dase at the 2021 Canadian Society of Digital Humanities conference 'Making the Network'; this paper featured pilot visualizations of data from DigitalDonne.) Challenges including the lack of a single standard finding list comparable to the STC or Wing catalogues of printed books, the idiosyncratic nature of individual manuscripts, and their wide dispersal have delayed large-scale quantitative research. Perhaps more significantly, though, humanities network analysis was initially used to study the actions of identifiable human agents, like the senders and receivers of letters, or the printers, publishers, and booksellers involved in the trade in printed books. Manuscript verse miscellanies present distinct methodological challenges because of the relative paucity of information about their contents' authorship, attribution, transcription, and provenance.

STEMMA overcomes these intellectual and logistical challenges by combining, augmenting, enriching, and returning the most comprehensive existing resources as structured, open data; applying insights from graph theory and social network analysis to model the circulation of English poetry in manuscript; and producing new knowledge about the settings in which English verse was read. In doing so, the project aims to overturn longstanding assumptions about verse manuscripts as products made by and for discrete, privileged groups.

(2) DATASET DESCRIPTION

In accordance with the ERC's recommendations that project data be "as open as possible, as closed as necessary," an open version of the full project dataset will be deposited when it is fully cleaned and ready for analysis (likely summer 2026) and updated throughout the remainder of the funding period. A Readme file, including sample records from the source datasets, has been deposited on Zenodo.

When designing the project, I was fortunate to be given permission to incorporate six important datasets about early modern English manuscript culture into STEMMA's dataset (see Table 1).

Table 1 Data sources for the STEMMA project.

SOURCE	CONTACT	FORMAT	DATA TYPE	DATA VOLUME
Catalogue of English Literary Manuscripts	John Lavagnino, Kings College London	XML	Bibliographic dataset	103.9 MB (979 XML files)
DigitalDonne	Brent Nelson, University of Saskatchewan	CSV	Bibliographic dataset	1.2 MB (1 table, 4,240 lines)
Index of Selected English Poetry Manuscripts, 1590–1660	Joshua Eckhardt, Virginia Commonwealth University	CSV	Bibliographic dataset	1.8 MB (1 table, 9,178 lines)
Perdita Project: A Database for Early Modern Women's Manuscript Compilations	Victoria Burke, University of Ottawa	HTML	Bibliographic dataset	25 MB (500 entries)
RECIRC: The Reception and Circulation of Early Modern Women's Writing 1550–1700	Marie-Louise Coolahan, University of Galway	CSV	Bibliographic dataset	100 MB (170 tables)
Union First-Line Index of English Verse	Eric Johnson, Folger Shakespeare Library	CSV	Bibliographic dataset	247.3 MB (1 table, 704,321 rows)

Together, these resources included nearly one million records related to the circulation of English verse. The collation of this material is not the primary aim of the STEMMA project; however, because these records will eventually be represented as nodes and edges in our network model, systematic data preparation is an essential prerequisite for the project's advanced literary and quantitative analyses.

REPOSITORY LOCATION

https://doi.org/10.5281/zenodo.18247735

REPOSITORY NAME

Zenodo

OBJECT NAME

STEMMA_DATASET_README v1.docx

CREATION DATES

2023-09-01 to present

DATASET CREATORS

Erin A. McCarthy (University of Galway): conceptualization, methodology, funding acquisition, project administration, supervision, data curation, writing – original draft

Jan Putzan (Ember Ltd.): software

Caitlin Burge (University of Galway): data curation

Douglas Clark (University of Galway): data curation

Kyle Dase (University of Galway): data curation

Meghan Kern (University of Galway): data curation

Millie Randall (University of Galway): data curation

Leah Veronese (University of Galway): data curation

LANGUAGE

English

LICENSE

ODC Open Database License v1.0

PUBLICATION DATE

2026-01-14

(3) METHOD

The project began by importing most of the material described above into a single PostgreSQL database. The exception is the Perdita data, which is preserved only in a frame-based website built in 1997. Although we had consent to import data from that project, we found that we could not scrape it programmatically. It will therefore be added to the STEMMA database manually once all other cleaning and reconciliation is complete.

The remainder of the data has been mapped to a uniform set of fields to facilitate the removal of duplicates, creation of authorities (where none yet existed), reconciliation, and assignment of unique identifiers. Cleaning has generally focused on three tasks:

1) Identifying copies of the same work to support the creation of a work-level entity
2) Removing duplicate and out-of-scope records
3) Standardizing metadata (e.g., naming/numbering conventions) but *not* textual data

It is imperative that these tasks are completed before we try to represent our data as a network.

The term “data cleaning” (also known as “data cleansing,” “data wrangling,” “data munging,” and even “data janitor work”) can refer to a range of necessary and important, but also difficult and often tedious, activities, including (but not limited to) disambiguation, duplicate detection, identification of near neighbors, reconciliation, restructuring, and fuzzy searching as well as removal of data (Kim et al, 2003; Leahey, 2008; Christen, 2012; Wickham, 2014; Leonelli, Rappert, and Davies, 2017; Hyvönen et al, 2019). These processes necessarily precede analysis

and interpretation, as they are "directed not towards eliciting meaning, so much as towards eliciting the right form" (Walford, 2020). As we move towards a more robust environment of Linked Open Data (LOD), digital humanities researchers will need to be able to process more data more quickly to cope with the increase in access (Lewis et al, 2019; Ryan et al, 2020). Recent estimates hold that this work can consume 50–80% of the total person-hours for a project (Lohr, 2014; Wickham, 2014).

Digital humanities scholars have occasionally argued against cleaning, usually on the grounds that doing so removes complexity or nuance from the dataset (Rawson & Muñoz, 2019). However, because STEMMA depends on the aggregation of disparate datasets, a certain amount of harmonization, reconciliation, and non-destructive cleaning was necessary. Although the project is not intended to produce an authoritative bibliographical resource in its own right, our analysis, visualizations, and conclusions will all be better if we can get the underlying data into the best shape we can. Were we to use the data in its received form, any network model would inevitably be distorted by the inclusion of duplicates (and, in some cases, triplicates and quadruplicates) as well as competing and sometimes conflicting records. Moreover, this cleaning process has highlighted areas where the data is incomplete and allowed us to make decisions as a team about where further research is required to augment the data. Curating the best possible data will also benefit users of the forthcoming open dataset, which will eventually be made linkable and interoperable.

The STEMMA dataset presents challenges related to its source material and methodological orientation. Our data sources differ in scope and emphasis: the Catalogue of English Literary Manuscripts (CELM, a digital supplement to Beal's landmark 1980 *Index of English Literary Manuscripts*) focuses exclusively on works by 237 canonical authors; the Perdita Project (Perdita) catalogues women's manuscripts; the Union First-Line Index (UFLI) includes English verse from manuscripts held in eight repositories; and RECIRC: The Reception and Circulation of Early Modern Women's Writing records evidence that women's writing was read. More specialized digital resources have been created to support specific projects: DigitalDonne preserves work products created by the editors of *The Variorum Edition of the Poetry of John Donne*, while Joshua Eckhardt's Index of Selected English Poetry Manuscripts, 1590–1660 informed his monograph *Manuscript Verse Collectors and the Politics of Anti-Courtly Love Poetry*. The gaps, overlaps, and differences in data structure between these resources, along with unique researcher and institutional practices, complicate the process of mapping one source neatly onto another. Sample records showing the representations of a single transcription across all five source datasets are available in the Readme file described above.

Features of early modern English literary culture generally and manuscript transcription specifically further confound efforts at computational study. First, the English language started to look and sound the way it does today during the years of our study—and, as Arja Nurmi reminds us, "English around 1500 was very different from English around 1700" (2010, p. 15). At the beginning of the period, there was no standard orthography or spelling, even for one's own name (Nurmi, 2010). London English had not yet supplanted regional dialects, and writers in further-flung provinces and newly founded colonies produced texts that were markedly different from those written by their counterparts in the capital. The language was more recognizably modern by 1700. Moreover, early modern scribes altered their texts in both intentional and unintentional ways, making the question of what constitutes a copy of the same "work" less than straightforward (Marotti, 1995). Many of these idiosyncrasies provide valuable historical evidence, so while modernizing or otherwise regularizing the data may have made the computational dimensions of the project more straightforward (and indeed, would have been standard practice on similar projects until recently), this approach was a non-starter.

Although off-the-shelf tools proved useful in a pilot project, their tendency to overwrite "bad" or "dirty" data meant that they were not fit for purpose, either for STEMMA or for future users of our open dataset. OpenRefine is perhaps the best-known data cleaning application, but while it is graphically navigable and intuitive (Miller and Vielfaure, 2022), it is meant to standardize data. Furthermore, the matching methods available in OpenRefine would offer limited purchase on our dataset: Levenshtein distance (sometimes called "edit distance" because it measures "the number of operations [mismatch, insertion, deletion] needed to transform a string into another one") and phonetic matching can account for small variations and scribal changes, but these methods would likely miss more extensive adaptations and interpolations (Marçais et al,

2019, i128). Open Data Editor facilitates the exploration and cleaning of tabular data, but it focuses on the correction of "errors" (Open Knowledge Foundation). Easy Data Transform has many these functions and also allows users to change the format of data files (Oryx Digital, 2025). Flookup Data Wrangler and Alteryx (formerly Trifacta) are perhaps most similar to the solution developed by the STEMMA team as they use no-/low-code AI for fuzzy matching and reconciliation, but both still ultimately standardize records (Apell, 2025; Alteryx, 2025)—a method that may be desirable for some applications but that is at odds with most humanities projects, where variability in the representation of data may be as important as the data itself.

Data processing tools developed specifically for humanities researchers tend to focus on the creation of Linked Data. Consequently, they would not be comparators for STEMMA's process. These include Karma and its successor SAND, both developed by the Center on Knowledge Graphs at the University of Southern California, and the University of Mannheim's Silk: The Linked Data Integration Framework (Volz et al, 2009). All three facilitate the creation of ontologies and knowledge graphs and offer basic data cleaning functionality (SAND, n.d.). Recon, developed for the Cultures of Knowledge project, checks for string similarity between spreadsheets and authority lists (Hyvönen et al, 2019). Although we intend to move towards LOD nearer the project's conclusion, these tools were not useful at the project's beginning.

It also seemed impractical, if not impossible, to clean and reconcile the entire dataset manually. The team and subcontractors did some manual harmonization work during the data import process, including tidying dates and shelfmarks and disambiguating named people. This was generally focused on ensuring uniformity and precision rather than establishing relationships between the datasets, which began after the import was complete and the first version of the web application had been deployed. However, as Andres Karjus observes, "human time is a bottleneck" (2024, p. 17), as is attention and cognitive capacity (Messeri & Crockett, 2024). Instead of attempting to work directly with the data itself, we have explored using machine learning techniques that have allowed us to preserve important paleographical and orthographical evidence. Working with these problems computationally has saved us months, if not years, of manual data-checking by allowing us to target our work.

Some key terms from our database that might be helpful for the discussion that follows are "Manuscripts," the entity for recording data about a specific material document, and "Manuscript Items," which represent transcriptions within these documents. A key entity that was missing in most of our inherited data was what we initially called the "work-level entity," now just "Works," which captures distinct literary works in the Tansellean sense—that is, their texts can vary from document to document, but we generally still agree that they are meant to represent the same artistic creation (Tanselle, 1989). And it is here that AI has been most useful: in setting aside what textual editors, following W.W. Greg, would call both "accidentals" and "substantives" to try to locate all relevant witnesses, at scale (1950, p. 21).

It may be worth noting from the outset that this process emerged organically from our research. Because we did not seek to develop a new method, we did not document our intermediate steps as rigorously as we might have, nor did we compare our methods to others systemically (though it would be useful to do so in a future project). Moreover, because this method is entirely bespoke, a meaningful benchmarking effort would require testing it on at least one other "neutral" dataset. Nevertheless, the reflections below attempt to reconstruct the development of this method and point towards the necessity of new ways to validate experimental methods in DH research.

Below, I briefly outline a staged AI pipeline developed specifically for work on the STEMMA dataset. It was necessary to work in this manner because no single method worked well across the disparate sources and records. Different methods performed differently across the datasets not only because of their inherent strengths and limitations but also because our understanding of the methods and their application to our data improved over time. The overall AI architecture was developed rapidly and iteratively between March and November 2025 by Jan Putzan, a Senior Software Engineer and subcontractor from Ember Ltd., in response to requirements and feedback presented by the STEMMA team (see Figure 1). Each method was selected based on Putzan's practical experience, ease of integration with STEMMA's existing PostgreSQLdatabase, and performance on noisy, short text; cost and efficiency were also considered. First, locality sensitive hashing (LSH) was used to identify candidate matches across the dataset, where

spelling variation and incomplete fields made exact matching unreliable or impossible. Next, vector embeddings were used to capture semantic similarity where surface string overlap was weak or variable. Both general and domain-adapted models were used to reduce any bias from modern-language training data. Finally, an agentic model combining several imperfect signals was used to resolve borderline cases that could not be resolved on the basis of a single similarity score alone.

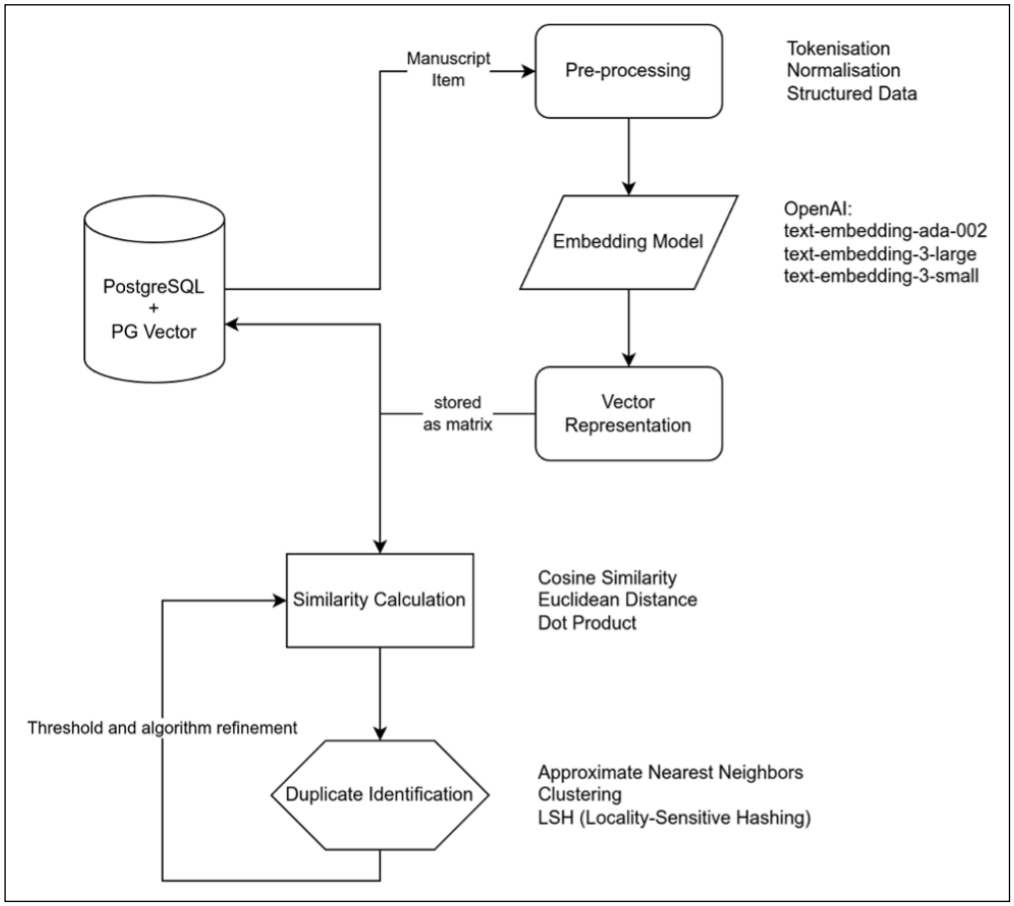

Figure 1 The first diagram of the STEMMA deduplication pipeline, created by Jan Putzan (Ember Ltd.).

(3.1) LOCALITY SENSITIVE HASHING

As a first step, we used locality-sensitive hashing to group similar first lines into "buckets" that seem to be copies of the same poetic work. This technique reduces data dimensionality by using a random hash (or sketch) function, then places the resulting signatures into "buckets" where each hash has a high probability of being similar to the other items. "Low dimensionality" is relative in this context—our signatures were 256 dimensions, generated algorithmically from textual features such as n-grams (shingles) and random projections. The number of dimensions was chosen, after some trial and error, to balance recall, collision rate, and performance. The aim was to preserve similarity between records rather than to represent meaning explicitly.

This process clustered approximately 198,000 Manuscript Items then in the combined dataset into 14,500 buckets for manual review. Duplicate records for individual Manuscript Items were merged, and matching Manuscript Items were also assigned to a new Work entity. Because this was a probabilistic method, it seemed wise to check its work, which we did in a bespoke terminal application. At this stage, the expert review was conducted using a local instance of our database and a command-line application installed on a single laptop physically passed between members of the research team (see Figure 2). The terminal application did not offer any metrics about the number of items merged/removed, but no items were added at this stage. Once this review was complete, approximately 38% of Manuscript Items were associated with a Work.

DOI: 10.5334/johd.490

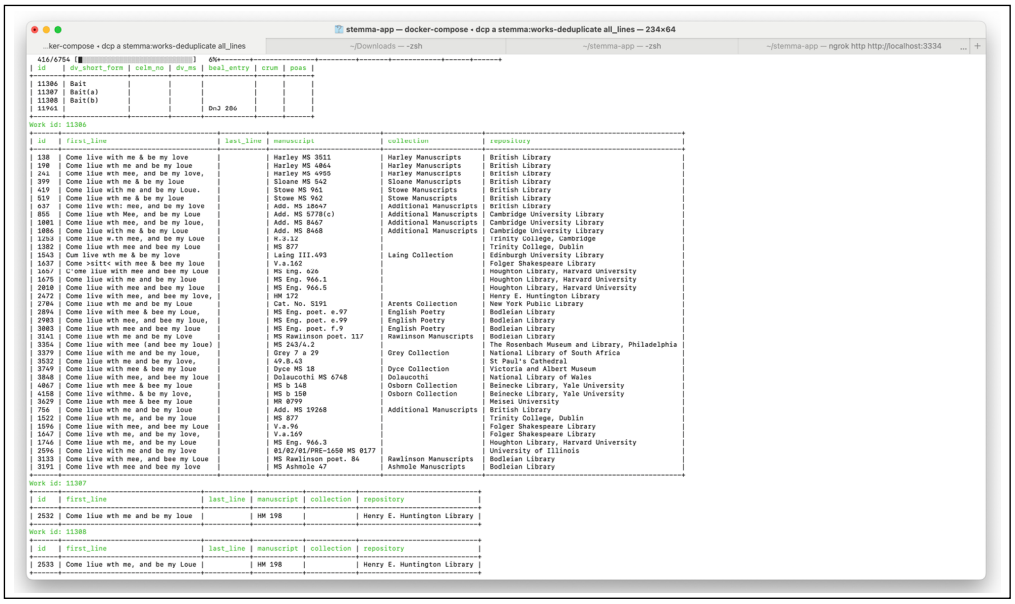

Figure 2 A screenshot of the Terminal application for reviewing "buckets" created by locality sensitive hashing.

(3.2) VECTOR EMBEDDINGS

The second method we used involved creating vector embeddings, then clustering items on the basis of Euclidean distance and cosine similarity. These buckets of potential proposed matches were again reviewed manually by members of the research team. This time, however, the matches were uploaded to our web application, allowing the team to work simultaneously.

In the first instance, vectors were created with OpenAI's text-embedding-3-small model, because our developer has found that it performs reliably on short, noisy text like manuscript catalogue data. The model is also efficient (and affordable) enough to support repeated experiments and reindexing, thus offering a stable baseline for semantic similarity. The similarity search was carried out with PG Vector, largely because it integrates directly with the existing relational database, allowing vector similarity search to be combined with structured metadata such as dates, repositories, and Work identifications. This also had the benefit of making the system easier to iterate on and easier to reproduce.

On the first pass, we used two different scopes: one in which all lines were vectorized, and one in which titles and headings were vectorized along with lines. This yielded 11,184 clusters comprising 32,323 distinct Manuscript Items. By the conclusion of this stage, 54% of Manuscript Items were associated with Works (a substantial increase from the 38% at the conclusion of the previous stage).

On a second pass, we supplemented the OpenAI embeddings with embeddings created with Enrique Manjavacas and Lauren Fonteyn's MacBERTh, a pre-trained model fine-tuned on an extensive corpus of early modern English texts (Manjavacas & Fonteyn, 2021; Manjavacas & Fonteyn, 2022). As expected, this model seemed to handle early modern spelling variation better than the OpenAI model. We also attempted to find dissimilar Items that were still unassociated with Works as a means of identifying unique Items; however, this method did not give useful results.

This was the longest and most intense phase of our work to date.

(3.3) INTERMEZZO: SQL AND REGEX

Between the first and second LLM passes described above, we did some more manual checking using more traditional processes, including regular expressions and SQL queries. This was partly to address issues in the data that had become obvious, but it also gave the team a brief respite from checking LLM results. There were also still numerous duplicate Manuscript Items to remove. A simple SQL query identified approximately 11,000 such items, and a further LLM-informed process highlighted 3,209.

Two SQL queries were particularly useful at this stage. The first addressed the 8,729 Manuscript Items that had only entered the database from a single data source, on the principle that these might be more likely to be unique. The second isolated Manuscript Items found in Manuscripts

where some, but not all, items had been merged. The idea was that these "Multiple source manuscripts" would identify Manuscript Items that the research team would see as duplicates but that the LLMs had missed. That there were only 551 such groups out of (at the time) around 100,000 Manuscript Items suggests that the LLM pipelines were quite effective. An additional 48,596 items were removed from the dataset at this time because they were outside the upper limit of the project's chronological scope.

(3.4) PRISM AGENTIC MODEL

For the final stages of automated cleaning, we experimented with a lightweight agentic model. Perhaps in a nod to our funder's fondness for acronyms, our developer dubbed this pipeline "PRISM: Pipeline for Retrieval Integrated Similarity Mapping."

PRISM begins by calculating two new metrics:

> **Item fingerprints**: embeddings (shingles, tokens (especially rare ones), metaphone tokens, hash signatures) from both the OpenAI and MacBERTh models, number of lines and average length
>
> **Work prototypes**: represent a work but from the point of view of one representative item (more weight for dates, extra lines)

Each of these is then rendered as a single centroid vector, now calculated with Facebook AI Similarity Search (Meta). FAISS is faster than PG Vector, and although it does not integrate directly with the STEMMA database, the additional processing steps in this pipeline made this less of a consideration. The model compares the Item fingerprint for each unassociated Manuscript Item with the Work prototypes to propose the likeliest matches, generating 2.5 million candidate pairs. These are ranked into bands (Auto-Link, Review, or New Work) with a LightGBM reranker using thresholds and training pairs based on our manual cleaning work (Microsoft Corporation, 2025). The entire pipeline is diagrammed in Figure 3.

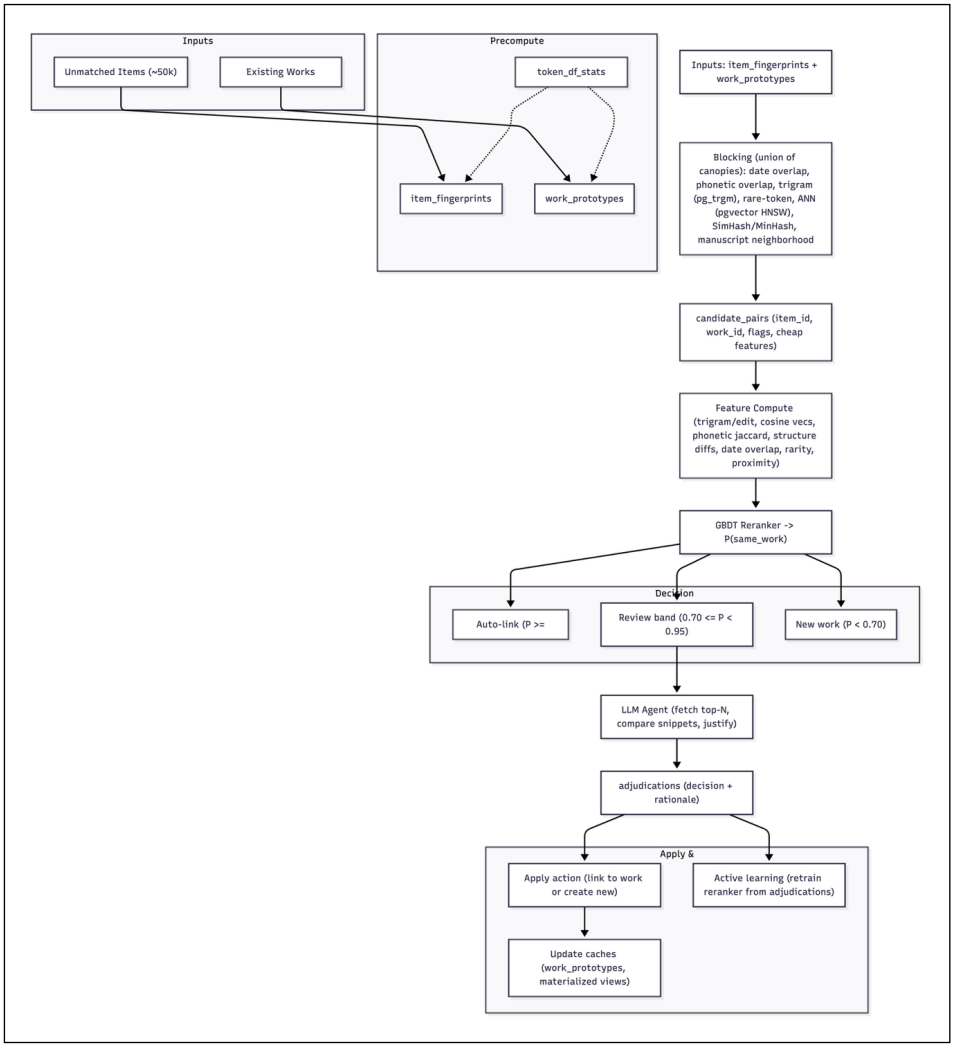

Figure 3 A diagram of the PRISM method created by Jan Putzan (Ember Ltd).

We considered, but ultimately decided against, allowing the agent to implement the decisions without a level of human review. Instead, the results were again uploaded to the web application as a new table with a label indicating the suggested action, which team members could approve, modify, or reject entirely. Recent deployments have also displayed reasoning narratives, which are useful insofar as they highlight strong signals despite the documented limitations of such "explanations" (see Levy et al, 2025). At this stage, we once again became excited about the idea of "Negative pairs"—that is, identifying the items that were *least* like each other—but abandoned this approach when we found that there were ten times as many negative pairs as positive pairs.

At the conclusion of this stage, we had 93,662 Manuscript Items (47% of the total number of Manuscript Items at the beginning of the cleaning process), with 68% associated with Works.

(4) RESULTS AND DISCUSSION

(4.1) EVALUATING OUR STAGED PIPELINE

To be clear, we have not yet benchmarked our process, nor is it possible to do so in this discussion paper. Our immediate goals were pragmatic: to clean and reconcile project data as quickly as possible. It was only in the process of doing this work that we realized that our pipeline might be of benefit to others. By then, it was too late to capture full metrics about our initial steps. Moreover, because we iterated over the same dataset multiple times, real-time benchmarking results would not have been useful, as progressively more sophisticated methods were applied to incrementally smaller datasets each time. It likely would have appeared that the more powerful methods were less useful, when in fact there was less for them to find. We now realize that a thorough benchmarking exercise would be a useful next step in communicating and sharing this method for possible reuse and extension (see also Weber et al, 2019). Until that is feasible, however, the observations below are offered in hopes that they might point up new directions for similar work.

Throughout the project, we have used cardinality (the number of items in each bucket or cluster), the total number of buckets/clusters, and the percentage of Manuscript Items associated with Works as proxies for our progress. It would likely have been beneficial to have tracked these metrics more carefully. However, because there is no ground truth—in other words, we did not have a clear sense of what values might represent the conclusion of our work—the need to do so was not apparent in the project's early stages.

The vector embedding method excelled at finding variants with significant differences in diction and word order—precisely the kinds of variation characteristic of early modern scribal practice. However, it was somewhat disappointing when confronted with smaller variations, such as spelling (a more general problem in early modern studies). Using the domain-adapted model, MacBERTh, mitigated these challenges. The choice of scopes was also impactful at this stage, as certain early modern scribal practices introduced noise; for example, data stored in the "Heading" field tended not to be useful because it often contained vague, redundant labels like "Song," "Sonnet," "Epitaph," or "Another." We re-ran our pipelines at several different thresholds, but perhaps less systematically than might be wished.

Our more recent agentic process ultimately involves classification ("match"/"no match"). A simple confusion matrix might be a useful benchmark here, along with related calculations of accuracy, sensitivity, specificity, balanced accuracy, precision, and recall (Bauer et al, 2025).

At all stages, it would also have been worthwhile to compare systematically the performance of the different embedding models that we used, and future benchmarking efforts could include additional models, including more open ones in line with scholarly best practice (Palmer et al, 2024).

Finally, although we have not benchmarked our computational methods, users of the open dataset can be assured of its quality and reliability because we have undertaken and nearly completed a thorough, record-by-record manual validation process. That said, a systematic benchmarking process may give us sufficient confidence to avoid this labor-intensive work in future projects using this pipeline.

(4.2) CONSEQUENCES FOR PROJECT MANAGEMENT

Although benchmarks are distinct from milestones, decisions made about technology inevitably impact project timelines and work patterns. When the tools available to researchers change

even as the work is underway, it is not always feasible—or desirable—to think in rigid terms about timelines or dependencies. Without a baseline sense of how long this work *should* take, we have effectively begun to work in Agile sprints in tandem with our database developers. This more flexible, bottom-up approach to project planning may strike some humanities researchers as unusual, but the evolving nature of our methods, along with the data's scope and complexity, make it difficult to set rigid, waterfall-style timelines for this part of the work (Posner, 2022; Ahnert et al, 2023).

One remaining problem is that of unique Items. The percentage of Manuscript Items still not associated with Works has remained stuck near 30% for months. As a manuscript scholar, I have an impressionistic sense that there will be a long tail of the dataset that is unique, but I am not yet able to quantify it on any evidentiary ground. How, then, can we know whether 30% is too high—or right where it needs to be? Our current review of each Manuscript record should provide clarity.

(5) FUTURE DIRECTIONS

Our method has been developed specifically for our own data. This would necessarily confound any attempt to benchmark it at this stage, as one would naturally expect that it would work particularly well on the dataset for which it was expressly designed. However, were we to scale this method, several open datasets could prove useful. Within early modern studies, these might include corpora from the Early English Books Online – Text Creation Partnership (EEBO-TCP) and the Folger Shakespeare Library's Early Modern Manuscripts Online (EMMO) projects. One might also expand the effort to include a broader chronological range, a more diverse range of material written in English, or multilingual corpora. It is unclear whether any existing purpose-built benchmark datasets would be suitable for this method, as test data would need to include errors and conflicting data to be useful.

This is a bespoke method that has been refined on each iteration; therefore, it would only have been possible to compare successive versions of the method by re-running it over the full, uncleaned dataset, which would have been at odds with the objectives and agreed timelines of our current funding. However, the PI is currently seeking funding to build the pipelines described here into a freestanding application, and fuller benchmarking will be useful as this tool is developed and refined. This follow-on project, "ARCHIVE: Automated Reconciliation and Cleaning of Historical Information with Vector Embeddings," was recently awarded a Seal of Excellence in the most recent ERC Proof of Concept call. Comparing the efficacy of each of the steps outlined in Section 3 would be particularly useful.

A reviewer of an early draft of this essay helpfully suggested that an inter-rater reliability protocol could facilitate a quantitative assessment of these methods (Zou, 2012). The team could not implement such a process in time to obtain and analyze meaningful results for this essay, but it will be an important step towards extending this method.

It is clear that different, and likely novel, validation frameworks are necessary for iterative and exploratory digital humanities research. Existing methods from related fields, including machine learning, are ill-suited for the evaluation of new tools on highly specialized, and occasionally idiosyncratic, tasks.

(6) IMPLICATIONS/APPLICATIONS

This discussion paper has considered the challenges associated with benchmarking a novel application of LLMs that emerged organically from a large-scale data cleaning and reconciliation process. Just as Ziems et al observe that LLMs can assist humans with text analysis and annotation (2024), the STEMMA team has found LLM-assisted pipelines invaluable in preparing our literary historical dataset. Although it is no longer possible to evaluate the efficacy of each iteration fully on this dataset, it is my hope that these reflections can offer useful insights for extensions of this method and suggest best practices for future projects addressing similar issues. It therefore responds to recent calls for benchmarks tailored to humanities and social science research (Kang et al, 2025) and LLM-based methods (Ziems et al, 2024). As computational humanities research becomes increasingly important, our methods will become more robust. It seems likely that our methods will be informed to some extent by our collaborators in other disciplines and industries, but there are also promising signs that the influence may run both ways (Messeri and Crockett, 2024).

ACKNOWLEDGEMENTS

Thanks to the creators of our original datasets for permission to reuse and adapt their foundational work; to Jan Putzan for developing the infrastructure and responding to our many queries; and to Jen Smith for her expert advice on open practice.

FUNDING STATEMENT

This research is funded by the European Union (ERC grant agreement no. 101088497). Views and opinions expressed are however those of the author only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

COMPETING INTERESTS

The author has no competing interests to declare.

AUTHOR AFFILIATIONS

Erin A. McCarthy orcid.org/0000-0002-5080-7051
Discipline of English, School of English, Media, and Creative Arts, University of Galway, Galway, Ireland

REFERENCES

Ahnert, R., & **Ahnert, S. E.** (2014). A community under attack: Protestant letter networks in the reign of Mary I. *Leonardo*, *47*, 275. https://doi.org/10.1162/LEON_a_00778

Ahnert, R., & **Ahnert, S. E.** (2015). Protestant letter networks in the reign of Mary I: A quantitative approach. *English literary history*, *82*, 1–33. https://doi.org/10.1353/elh.2015.0000

Ahnert, R., Ahnert, S. E., Coleman, C. N., & **Weingart, S. B.** (2020). *The network turn: changing perspectives in the humanities*. Cambridge University Press. https://doi.org/10.1017/9781108866804

Ahnert, R., Griffin, E., Ridge, M., & **Tolfo, G.** (2023). *Collaborative historical research in the age of big data: Lessons from an interdisciplinary project*. Cambridge University Press. https://doi.org/10.1017/9781009175548

Akkerman, N. (2018). *Invisible agents: Women and espionage in seventeenth-century Britain*. Oxford University Press.

Alteryx. (2025). *Data preparation, blending, and enrichment tools*. Retrieved February 27, 2026, from https://www.alteryx.com/products/capabilities/data-preparation-tools.

Apell, A. (2025). *Our incredible journey, so far*. https://www.getflookup.com/about-us/

Bauer, A., Züfle, M., Grohmann, J., & **Kounev, S.** (2025). Machine learning and artificial intelligence. In S. Kounev, K. D. Lange, & J. von Kistowski (Eds.), *Systems benchmarking* (pp. 323–346). Springer. https://doi.org/10.1007/978-3-031-85634-1_16

Beal, P. (1998). *In praise of scribes: Manuscripts and their makers in seventeenth-century England*. Clarendon Press. https://doi.org/10.1093/oso/9780198184713.001.0001

Bourke, E. (2017). Female involvement, membership, and centrality: A social network analysis of the Hartlib Circle. *Literature Compass*, *14*(4). https://doi.org/10.1111/lic3.12388

Christen, P. (2012). *Data matching: Concepts and techniques for record linkage, entity resolution, and duplicate detection*. Springer Nature. https://doi.org/10.1007/978-3-642-31164-2

Crawford, J. (2010). Literary circles and communities. In C. Bicks & J. Summit (Eds.), *The history of British women's writing, 1500–1610* (vol. 2, pp. 147–164). Palgrave Macmillan.

Cust, R. (1986). News and politics in early seventeenth-century England. *Past & present*, *112*, 60–90. https://doi.org/10.1093/past/112.1.60

Daybell, J. (2011). Gender, politics and diplomacy: Women, news and intelligence networks in Elizabethan England. In R. Adams & R. Cox (Eds.), *Diplomacy in early modern culture*, pp. 101–19. Palgrave Macmillan. https://doi.org/10.1057/9780230298125_7

Daybell, J. (2012). *The material letter: Manuscript letters and the culture and practices of letter-writing in early modern England*. Palgrave. https://doi.org/10.1057/9781137006066

De Groot, J. (2006). Coteries, complications and the question of female agency. In I. Atherton & J. Sanders (Eds.), *The 1630s: Interdisciplinary essays on culture and politics in the Caroline era* (pp. 189–209). Manchester University Press. https://doi.org/10.1515/9781503627994

Ezell, M. J. M. (1999). *Social authorship and the advent of print*. Johns Hopkins University Press.

Greg, W. W. (1950–51). The rationale of copy-text. *Studies in bibliography*, *3*, 19–36.

Greteman, B. (2021). *Networking print in Shakespeare's England: Influence, agency, and revolutionary change*. Text technologies. Stanford University Press.

Hackett, H. (2012). Women and Catholic manuscript networks in seventeenth-century England: New research on Constance Aston Fowler's miscellany of sacred and secular verse. *Renaissance quarterly*, *65*, 1094–24. https://doi.org/10.1086/669346

Hyvönen, E., Ahnert, R., Ahnert, S. E., Touminen, J., Mäkelä, E., Lewis, M., & **Filarski, G.** (2019). Reconciling metadata. In H. Hotson & T. Wallnig (Eds.), *Reassembling the Republic of Letters* (pp. 223–236). Göttingen University Press.

Kang, Z., Gong, J., Yan, J., Xia, W., Wang, Y., Wang, Z., Ding, H., Cheng, Z., Cao, W., Feng, Z., He, S., Yan, S., Chen, J., He, X., Jiang, C., Ye, W., Yu, K., & **Li, X.** (2025). *HSSBench: Benchmarking humanities and social sciences ability for multimodal large language models* (No. arXiv:2506.03922). arXiv. https://doi.org/10.48550/arXiv.2506.03922

Karjus, A. (2024). *Machine-assisted quantitizing designs: Augmenting humanities and social sciences with artificial intelligence* (No. arXiv:2309.14379). arXiv. https://doi.org/10.1057/s41599-025-04503-w

Kim, W., Choi, B.-J., Hong, E.-K., Kim, S.-K., & **Lee, D.** (2003). A taxonomy of dirty data. *Data mining and knowledge discovery*, *7*(1), 81–99. https://doi.org/10.1023/A:1021564703268

Kneidel, G., Nelson, B., & **Dase, K.** (2021, 30 May–3 June). *Striking a NERV: The case for networking English verse* [Conference presentation]. Canadian Society of Digital Humanities/Société Canadienne des humanités numériques, online.

Ladd, J. R. (2021). Imaginative networks: Tracing connections among early modern book dedications. *Journal of cultural analytics*, *3*, 64–101. https://doi.org/10.22148/001c.21993

Leahey, E. (2008). Overseeing research practice: The case of data editing. *Science, technology, & human Values*, *33*(5), 605–630. https://doi.org/10.1177/0162243907306702

Leonelli, S., Rappert, B., & **Davies, G.** (2017). Special issue introduction: Data shadows: Knowledge, openness, and absence. *Science, technology, & human Values*, *42*(2), 191–202. https://doi.org/10.1177/0162243916687039

Levy, F. J. (1982). How information spread among the gentry, 1550–1640. *Journal of British studies*, *21*, 11–34. https://doi.org/10.1086/385788

Levy, M., Elyoseph, Z., & **Goldberg, Y.** (2025). *Humans perceive wrong narratives from AI reasoning texts* (No. arXiv:2508.16599). arXiv. https://doi.org/10.48550/arXiv.2508.16599

Lewis, M., Bosse, A., Hotson, H., Wallnig, T., & **van Miert, D.** (2019). Time. In H. Hotson & T. Wallnig (Eds.), *Reassembling the Republic of Letters* (pp. 97–117). Göttingen University Press.

Lohr, S. (2014, August 18). For big-data scientists, "janitor work" is key hurdle to insights. *The New York times*. https://www.nytimes.com/2014/08/18/technology/for-big-data-scientists-hurdle-to-insights-is-janitor-work.html

Manjavacas Arevalo, E., & **Fonteyn, L.** (2021). MacBERTh: Development and evaluation of a historically pre-trained language model for English (1450–1950). In M. Hämäläinen, K. Alnajjar, N. Partanen, & J. Rueter (Eds.), *Proceedings of the workshop on natural language processing for digital humanities* (pp. 23–36). NLP Association of India (NLPAI). https://aclanthology.org/2021.nlp4dh-1.4/

Manjavacas, E., & **Fonteyn, L.** (2022). Adapting vs. pre-training language models for historical languages. *Journal of data mining & digital humanities*, *NLP4DH*(Digital humanities in languages). https://doi.org/10.46298/jdmdh.9152

Marçais, G., DeBlasio, D., Pandey, P., & **Kingsford, C.** (2019). Locality-sensitive hashing for the edit distance. *Bioinformatics*, *35*(14), i127–i135. https://doi.org/10.1093/bioinformatics/btz354

Marotti, A. F. (1995). *Manuscript, print, and the English Renaissance lyric*. Cornell University Press. https://doi.org/10.7591/9781501728501

May, S. W. (2004). The future of manuscript studies in early modern poetry. *Shakespeare studies*, *32*, 56–62.

May, S. W., & **Marotti, A. F.** (2014). *Ink, stink bait, revenge, and Queen Elizabeth: A Yorkshire yeoman's household book*. Cornell University Press.

May, S. W., & **Wolfe, H.** (2010). Manuscripts in Tudor England'. In K. Cartwright (Ed.), *A companion to Tudor literature* (pp. 125–39). Wiley-Blackwell. https://doi.org/10.1002/9781444317213.ch8

Messeri, L., & **Crockett, M. J.** (2024). Artificial intelligence and illusions of understanding in scientific research. *Nature*, *627*(8002), 49–58. https://doi.org/10.1038/s41586-024-07146-0

Meta. (n.d.). *Faiss*. Retrieved February 27, 2026, from https://ai.meta.com/tools/faiss/

Microsoft Corporation. (2025). *LightGBM*. Retrieved February 27, 2026, from https://lightgbm.readthedocs.io/en/stable/

Miller, M., & **Vielfaure, N.** (2022). OpenRefine: An approachable open tool to clean research data. *Bulletin – Association of Canadian Map Libraries and Archives (ACMLA)*, 170. https://doi.org/10.15353/acmla.n170.4873

Millstone, N. (2016). *Manuscript circulation and the invention of politics in early Stuart England*. Cambridge University Press. https://doi.org/10.1017/CBO9781316343111

Nurmi, A. (2010). The English language of the early modern period. In M. Hattaway (Ed.), *A new companion to English Renaissance literature and culture* (pp. 15–26). Blackwell Publishing. https://doi.org/10.1002/9781444319019.ch2

Open Knowledge Foundation. (n.d.). *Open Data Editor*. Retrieved February 27, 2026, from https://okfn.org/en/projects/open-data-editor/

Open Refine. (n.d.). *Open Refine*. Retrieved February 27, 2026, from https://openrefine.org

Oryx Digital Ltd. (2025). *Transform data into information*. Retrieved February 27, 2026, from https://www.easydatatransform.com

Palmer, A., Smith, N. A., & **Spirling, A.** (2024). Using proprietary language models in academic research requires explicit justification. *Nature computational science*, *4*(1), 2–3. https://doi.org/10.1038/s43588-023-00585-1

Posner, M. (2022). Agile and the long crisis of software. *Logic(s) magazine*, *16*. Retrieved 9 September 2025, from https://logicmag.io/clouds/agile-and-the-long-crisis-of-software/

Rawson, K., & **Muñoz, T.** (2019). Against cleaning. In M. K. Gold & L. F. Klein (Eds.), *Debates in the digital humanities 2019* (pp. 270–292). University of Minnesota Press. https://doi.org/10.5749/j.ctvg251hk.26

Raymond, J., & **Moxham, N.** (Eds.) (2016). *News networks in early modern Europe*. Library of the written word, vol. 47. Brill. https://doi.org/10.1163/9789004277199

Rehbein, M. (2014). From the scholarly edition to visualization: Re-using encoded data for historical research. *International journal of humanities and arts computing*, *8*, 81–105. https://doi.org/10.3366/ijhac.2014.0121

Ryan, Y., Ahnert, S. E., & **Ahnert, R.** (2020). Networking archives: quantitative history and the contingent archive. *Proceedings of the Workshop on Computational Humanities Research*, *2723*, 385–396. http://ceur-ws.org/Vol-2723/

SAND. (n.d.). Retrieved February 27, 2026, from https://github.com/usc-isi-i2/sand

Scott-Warren, J. (2000). Reconstructing manuscript networks: The textual transactions of Sir Stephen Powle. In A. Shepard & P. Withington (Eds.), *Communities in early modern England: Networks, place, rhetoric* (pp. 18–37). Manchester University Press.

Smith, D. S. (2014). *John Donne and the Conway papers*. Oxford University Press.

Strocchia, S. T. (2014). Introduction: Women and healthcare in early modern Europe. *Renaissance studies*, *28*, 496–514. https://doi.org/10.1111/rest.12076

Tanselle, G. T. (1989). *A rationale of textual criticism*. University of Pennsylvania Press.

Verweij, S. (2016). *The literary culture of early modern Scotland: Manuscript production and transmission, 1560–1625*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198757290.001.0001

Vine, A. (2019). *Miscellaneous order: Manuscript culture and the early modern organization of knowledge*. Oxford University Press. https://doi.org/10.1093/oso/9780198809708.001.0001

Volz, J., Bizer, C., Gaedke, M., & **Kobilarov, G.** (2009). Silk – A link discovery framework for the web of data. In *Proceedings of the Linked Data on the Web Workshop* (LDOW2009), Madrid, Spain, April 20, 2009, CEUR Workshop Proceedings, ISSN 1613–0073, online. https://ceur-ws.org/Vol-538/ldow2009_paper13.pdf

Walford, A. (2020). Data aesthetics. In T. Carroll, A. Walford, & S. Walton (Eds.), *Lineages and advancements in material culture studies* (pp. 205–217). Routledge. https://doi.org/10.4324/9781003085867-15

Warren, C. N., Shore, D., Otis, J., Wang, L., Finegold, M., & **Shalizi, C.** (2016). Six degrees of Francis Bacon: A statistical method for reconstructing large historical social networks. *Digital humanities quarterly*, *10*(3). https://doi.org/10.17613/mdwdc-tne88

Weber, L. M., Saelens, W., Cannoodt, R., Soneson, C., Hapfelmeier, A., Gardner, P. P., Boulesteix, A.-L., Saeys, Y., & **Robinson, M. D.** (2019). Essential guidelines for computational method benchmarking. *Genome biology*, *20*(1), 125. https://doi.org/10.1186/s13059-019-1738-8

Wickham, H. (2014). Tidy data. *Journal of statistical software, 59*, 1–23. https://doi.org/10.18637/jss.v059.i10

Wilcox, K. R. (2012). American women's writing in the colonial period. In D. M. Bauer (Ed.), *The Cambridge history of American women's literature* (pp. 55–73). Cambridge University Press. https://doi.org/10.1017/CHOL9781107001374.005

Woudhuysen, H. R. (1996). *Sir Philip Sidney and the circulation of manuscripts 1558–1640*. Clarendon Press. https://doi.org/10.1093/acprof:oso/9780198129660.001.0001

Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., & **Yang, D.** (2024). Can large language models transform computational social science? *Computational linguistics*, *50*(1), 237–291. https://doi.org/10.1162/coli_a_00502

Zou, G. (2012). Sample size formulas for estimating intraclass correlation coefficients with precision and assurance. *Statistics in medicine*, *31*(29), 3972–3981. https://doi.org/10.1002/sim.5466

TO CITE THIS ARTICLE:
McCarthy, E. A. (2026). Are We There Yet? Notes Towards Benchmarking an Experimental AI-Assisted Workflow for Humanities Data Cleaning and Reconciliation. *Journal of Open Humanities Data*, 12: 46, pp. 1–14. DOI: https://doi.org/10.5334/johd.490

Submitted: 27 November 2025
Accepted: 20 February 2026
Published: 18 March 2026

COPYRIGHT:
© 2026 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See http://creativecommons.org/licenses/by/4.0/.

Journal of Open Humanities Data is a peer-reviewed open access journal published by Ubiquity Press.