International Journal of Digital Humanities (2023) 5:283–298
https://doi.org/10.1007/s42803-023-00068-9

ORIGINAL PAPER

# Reproducibility, verifiability, and computational historical research

**Toby Burrows**[1,2]

Received: 28 January 2023 / Accepted: 28 July 2023 / Published online: 31 August 2023


## Abstract

Digital humanities methods have been at the heart of a recent series of high-profile historical research projects. But these approaches raise new questions about reproducibility and verifiability in a field of research where grounding one's conclusions in a body of historical evidence is crucial. While there have been extensive debates about the nature and methods of historical research since the nineteenth century, the underlying assumption has generally been that documenting one's sources in a series of footnotes is essential to enable other researchers to test the validity of the research. Even if this approach never amounted to "reproducibility" in the sense of scientific experimentation, it might still be seen as broadly analogous, since the evidence can be reassembled to see the basis for the explanations that were offered and to test their validity. This essay examines how new digital methods like topic modelling, network analysis, knowledge graphs, species models, and various kinds of visualizations are affecting the process of reproducing and verifying historical research. Using examples drawn from recent research projects, it identifies a need for thorough documentation and publication of the different layers of digital research: digital and digitized collections, descriptive metadata, the software used for analysis and visualizations, and the various settings and configurations.

**Keywords** Digital history · Digital humanities · Reproducibility · Verifiability

## 1 Introduction

Digital humanities methods have been at the heart of a recent series of high-profile historical research projects. Machine learning, text analysis, handwritten text recognition, network analysis, knowledge graphs, and even ecological species modelling

✉ Toby Burrows
toby.burrows@uwa.edu.au

[1] Faculty of History, University of Oxford, Oxford, UK

[2] School of Humanities, University of Western Australia, Nedlands, Australia

have been deployed in sophisticated and innovative ways. But these approaches raise new questions about reproducibility and verifiability in a field of research where grounding one's conclusions in a body of historical evidence is crucial.

While there have been extensive debates about the nature and methods of historical research since the nineteenth century, the underlying assumption has generally been that documenting one's sources in a series of footnotes is essential to enable other researchers to test the validity of the research. Explanatory processes are usually seen as separate from this documentation of sources. Even if this approach never amounted to "reproducibility" in the sense of scientific experimentation, it might still be seen as broadly analogous, since the evidence can be reassembled to see the basis for the explanations that were offered and to test their validity.

The essay looks at a range of digital methods – topic modelling, network analysis, knowledge graphs, species modelling, and various kinds of visualizations – and examines how they are affecting the process of reproducing and verifying historical research. Using examples of recent research which rely on each of these methods, the essay identifies a need for thorough documentation and publication of the different layers of digital research: digital and digitized collections, descriptive metadata, the software used for analysis and visualizations, and the various settings and configurations. Without this, it will become increasingly difficult – if not impossible – to reproduce or assess the validity of research findings derived from these digital environments.

## 2 Pre-digital historical research: Footnotes and index cards

In the pre-digital era – up to, say, the 1970s and 1980s – historical research traced its main lineage back to the great nineteenth century German historian Leopold von Ranke (1795–1886), whose methodology, "with its emphasis on primary sources and its insistence on a rigorous critique of both form and content", has had an enormous effect on "all subsequent historical scholarship" (Powell, 1990:xvii-xviii). In particular, Ranke laid the foundation for the centrality of the footnote as a technique for presenting historical narratives in a way which emphasized the accuracy and objectivity in the historian's use of sources. As Anthony Grafton shows in his entertaining account of the history of the footnote, footnotes existed well before Ranke, and it took Ranke some time to develop his own approach. But he was the one who combined them with "systematic, original, critical research" in a way which was much-imitated by subsequent generations of historians (Grafton, 1997:45).

Grafton points out that the footnote actually plays several different roles in publications which report the results of historical research: citing earlier work by historians and others, adding explanatory details which would interrupt the flow of the argument, and referring readers to the specific sources of evidence for the statements made in the text. It is the third of these which is most relevant to questions of reproducibility and verifiability. Conventionally, a distinction was made between "primary" sources" (documents and other forms of evidence contemporaneous with the historical events being described) and "secondary" sources (subsequent work by historians and others describing and analysing historical phenomena). This distinction

is often blurred in footnotes, however, since both types of evidence can be deployed to support explanations, arguments, and narratives. But, at the very least, the footnote should enable the reader to track down the sources used and consult them directly, in order to assess whether the historian's conclusions are justified by the evidence cited.

The difference between the sources used and the arguments and explanations advanced is not always so clear-cut in practice. The writing of history in these circumstances can often appear to be a somewhat mysterious process where the narrative is anchored to the sources embedded in the footnotes, but the exact relationship between them is blurred, sometimes deliberately. The various types of logical errors which can arise in this kind of approach are set out in David Hackett Fischer's scathing typology of historians' fallacies, in which he identifies types of arguments which are not justified by the sources or which apply faulty reasoning to the sources (Fischer, 1970).

Reproducibility, in the sense applied to scientific experiments, was not the aim in this footnote-based approach. Rather, the goal was to ensure that the sources on which an argument was based could be identified and consulted, as the primary means of distinguishing fact from interpretation, and of assessing whether the conclusions drawn from the evidence were justifiable. By checking the sources, the reader was able to assess the reliability of the logical and explanatory framework, in both argument and narration, which was built on top of them.

Even in this manual environment, there were processes which intervened between reading the sources and constructing the arguments and narratives. The distinguished Early Modern British historian Keith Thomas once wrote up a detailed account of how he read his sources (Thomas, 2010):

> When I go to libraries or archives, I make notes in a continuous form on sheets of paper, entering the page number and abbreviated title of the source opposite each excerpted passage. When I get home, I copy the bibliographical details of the works I have consulted into an alphabeticised index book, so that I can cite them in my footnotes. I then cut up each sheet with a pair of scissors. The resulting fragments are of varying size, depending on the length of the passage transcribed. These sliced-up pieces of paper pile up on the floor. Periodically, I file them away in old envelopes, devoting a separate envelope to each topic. Along with them go newspaper cuttings, lists of relevant books and articles yet to be read, and notes on anything else which might be helpful when it comes to thinking about the topic more analytically. If the notes on a particular topic are especially voluminous, I put them in a box file or a cardboard container or a drawer in a desk. I also keep an index of the topics on which I have an envelope or a file. The envelopes run into thousands.

This manual form of topic modelling was once the typical way of proceeding for a historian, although index cards were often substituted for Thomas's pieces of paper. The historian and anthropologist Alan Macfarlane describes how he followed Thomas's approach (describing it as "one fact on one card") and notes the similar

approaches of other well-known researchers and writers (Macfarlane, 1999).[1] Having access to these cards or slips and the accompanying categorization schemes might provide valuable insights into the methods and approaches of a particular historian, but it was not necessary for other researchers to see these behind-the-scenes workings in order to assess the arguments and explanations presented in the published work. Going back and reading the sources themselves was the only satisfactory way of doing this.

## 3 The rise and fall of quantitative history

The late 1950s and 1960s saw the advent of what became known as "quantitative history": historical research and writing which relied on statistical techniques imported from applied mathematics. A typical introductory account, published in 1990, focuses on population statistics and covers such topics as populations and sampling, tests for statistical significance, relationships between variables, and "regression as historical explanation." The authors start with examples of journal articles from the 1980s which employed these methods, and argue that their readers "need to be able to critique historical writing that uses quantitative analysis." They link this with Bruno Latour's observations about three possible ways of reading specialized technical work: incomprehension, passive acceptance, and going beyond these to "actively grapple with the author's argument" perhaps by "regroup[ing] the data on a sheet of scratch paper, check[ing] some of the author's statistics, or try[ing] different statistical tests on the data" (Haskins & Jeffrey, 1990:7–10; Latour, 1987:60–61). This third approach involves something akin to reproducibility, though the authors do not use this term. They do, however, point out that statistical methods are usually applied within the framework of a confirmatory or hypothesis-testing style of research, similar to the sciences, rather than the exploratory style more characteristic of narrative-driven history (Haskins & Jeffrey, 1990:171–173).

A more theoretically ambitious and wide-ranging account is that of François Furet, which was originally published in French in 1971 but did not appear in English until 1985 (Furet, 1985). He draws heavily on the work of the *Annales* school, which was dominated by quantitative approaches during the so-called "Labrousse moment" of the 1950s and 1960s (Burguière, 2009:103–132). As well as Ernest Labrousse, its main practitioners included Pierre Chaunu who coined the term "serial history" to refer to any temporal series of homogenous units, both numerical and non-numerical, which can be compared and quantified.

While the *Annales* approach to quantitative or serial history was not initially computational, its practitioners became increasingly aware of the potential of computers for this kind of research. Furet makes the insightful observation that the computer forces historians to "give up methodological naivety" and to be explicit about the choices and hypotheses deployed in structuring the data (Furet, 1985:20). But he does not examine the implications of this observation for the publishing

[1] Macfarlane later converted his card index into a computer database (Macfarlane, 1992).

and footnoting of historical research. A more memorable, but simplistic, formulation was Emmanuel Le Roy Ladurie's often-quoted comment, in a 1968 newspaper article on "The historian and the computer", that "l'historien de demain sera programmeur ou il ne sera plus" (Ladurie, 1968).[2] Ironically, Ladurie himself – and the *Annales* school more generally – turned away from this kind of quantitative and serial history in the 1970s.

## 4 Publishing digital data

At the same time as the *Annales* school was turning away from quantitative history, a group of British historians were experimenting with approaches which can be seen as a precursor of data linkage, big data, and network analysis. Beginning in the 1970s, Alan Macfarlane and his colleagues spent more than twenty years "reconstructing historical communities" – particularly two parishes in Essex and Cumbria between the 16th and the 19th centuries. Initially, this involved manually transcribing and indexing every surviving document, and linking all the references to the same persons, places, dates, and topics.

Their 1977 book gave a detailed account of the procedures employed, and most of the transcribed and indexed data were published on microfiche in 1980 (Macfarlane et al., 1977; Macfarlane, 1980). These resources were used to answer a series of research questions relating to community structures. By 1983, both the texts and the indexes had been converted to a relational database (Macfarlane, 1983), which was published on the Web after 1994 and eventually transformed into XML in 2002 (Records of an English Village, n.d.).

Reproducibility, in this context, took on a new meaning. The archival sources, the indexes, and the methodologies used were all documented and made available for other researchers to re-visit. The only component missing was the software originally used to query the database, the "Cambridge Database System", though the manual from 1991 is still available (Macfarlane et al., 1991).

Macfarlane's approach was something of a precursor to the developments of the last 25 years, with vast amounts of historical sources, primary and secondary, being digitized or published directly in digital form on the Web. These have been published by collecting institutions, by commercial firms, and by historical research projects, and are usually accompanied by some kind of user interface which enables at least a modest level of searching and browsing. The data themselves are not necessarily accessible separately from the interface, however.

There has been a growing consensus that the data should be made available for download and reuse, separate from any specific software environment. The "Collections as Data" movement has promoted this approach for collecting institutions (Padilla et al., 2019), and libraries, archives, and museums have been increasingly taking up this approach (Candela et al., 2020). This may mean making the

[2] "The historian of tomorrow will be a programmer or will no longer exist." Later re-published in Ladurie, 1973.

digital materials available as files within folders, with no intervening software, as the OPenn digital repository of the University of Pennsylvania Libraries (2023) does. For researchers, journals like the *Journal of Open Humanities Data* now provide a scholarly venue for publishing data with an accompanying explanation of their content and value (McGillivray et al., 2022). Despite this, a great deal of historical data are only available through commercial sources like Cengage, Adam Matthew, and ProQuest or in other digital environments where licensing arrangements limit reuse.

These kinds of digital collections provide the materials from which historical narratives can be constructed and to which analytical and explanatory processes can be applied. It could be argued, as Jesse Torgerson has recently done, that these digital collections demand a new theory of historical practice, in which the distinction between primary and secondary sources is dismissed as false, and replaced by a model which insists on the ubiquity of the progression "sources – data – facts" in both digital and analogue settings (Torgerson, 2022). From the point of view of reproducibility and verifiability, these digital collections certainly add new layers of complexity to the process by which historical research is assessed and related to its sources. These include:

- The principles used in selecting, organizing, describing, and curating the collection;
- The methods by which it was created, including the specific software deployed, e.g., OCR, scanning, image creation, automated transcription (Nockels et al., 2022), and so on;
- The formats, encoding, and other structures applied at the level of individual files;
- How to cite elements from the collection in a stable and consistent way;
- Version control and other methods of recording the level of stability of the digital materials;
- Commercial and other limits on access; and,
- The relationship to analogue originals, where relevant.

Several of these have their equivalents in the non-digital world. Archival collections, for example, have their selection criteria and organizing principles, as well as their descriptive metadata schemas and vocabularies. But the digital world adds new factors which need to be documented and explained. The changeability and instability of digital collections are usually considerably greater, giving a greater prominence to the recording of versions, dates, formats, and the software used to create the data. To verify and reproduce historical research based on these collections, all these factors ought to be documented – both by the creators of the collections and by the researchers using them at specific periods of time.

Alongside these developments, there has been a considerable amount of work on "digital tool criticism", including reports on what Fickers and van der Heijden (2020) call "thinkering" – experiments with digital tools and technologies designed for humanities researchers. Critical infrastructure studies are also of increasing interest. Using the concept of "infrastructure as an analytical tool", for example, Urszula Pawlicka-Deger (2022) looks at issues affecting the development and current state of

global infrastructure for the digital humanities. While digital tools and infrastructure undoubtedly affect the reproducibility of research, my concern here is with the ways in which humanities researchers present and explain the result of specific projects, rather than the broader contexts in which their research is situated.

## 5 Digital analysis

The growing consensus around the need to make humanities research data, including collections data, openly available is an important achievement of the last decade. But this, on its own, is not sufficient to address questions of reproducibility in digital historical research. With the widespread availability of digital sources, digital methods of analysing data have become much more important and more widely deployed. As historical research becomes more reliant on approaches of this kind, they too will need to be documented in a way that makes it possible for other researchers to understand, assess, and reproduce the methods of data analysis which have been used.

This section examines examples of best practice from important research which uses digital methods like topic modelling, network analysis, knowledge graphs,and species modelling, as well as visualizations of various kinds. The software settings and algorithms used will affect the reproducibility of this kind of research just as much as the data, and are a critical element in verifying the appropriateness of the arguments advanced and the conclusions drawn.

### 5.1 Topic modelling

Topic modelling is a widely used approach for analysing the content of large corpora of texts, although Karsdorp et al. (Karsdorp et al., 2021:294) caution that the term is "not particularly informative" and that "mixed-membership model" is more accurate. Graham et al. (Graham et al., 2016:119) remind us that there are "many possible algorithms and approaches", although LDA (Latent Dirichlet Allocation) is the most popular among historians. There are also numerous types of software which will carry out "topic modelling" analyses, including software libraries to be used in programming environments like R and Python as well as standalone text analysis toolkits, both open source and commercial, like GTMT, MALLET, and STMT. The attraction of this kind of analysis is that it can "read" and explore large amounts of text, and capture the key themes and trends over time. These results can lead to more focused analyses based on hypotheses, as well as to narrower investigations of specific concepts and phenomena. In a sense, this offers a more rigorous and consistent – if less imaginative – version of the kind of reading and annotation carried out by historians like Keith Thomas.

Jo Guldi makes this analogy in an article describing her use of topic modelling to explore discussions of infrastructure in British Parliamentary proceedings of the nineteenth century. She suggests that the topic model could "act like a particularly well-gardened card catalogue, directing the researcher to names, places, keywords,

and dates that would illuminate the parallel events we have found, the changing discourses, and the complementary landscapes" (Guldi, 2019:33). An annotated version of this article, published in 2021, presents the results in detail but does not describe the software or the settings and configurations used to carry out this kind of analysis. Another paper by the same author, in which a nested topic model approach is applied to the same Parliamentary debates, does make reproducibility possible, however. It incorporates appendices containing the LDA topic models created with MALLET as well as the R code for reading these and for creating sunburst diagrams (Guldi & Williams, 2018). These are housed on the publisher's Web site; other researchers have used open repositories like GitHub or CERL's Zenodo for the same purpose.

### 5.2 Network analysis

While the mapping and exploration of networks of relationships are now relatively commonplace in historical research, network analysis in the formal sense is a more specific approach, based on the mathematical model of a network or graph with nodes and edges. Various mathematical calculations can be applied to measure such factors as centrality, density, paths, hubs and bridges, and weak ties. Whether it is appropriate or not to use this kind of approach for analysing relationships and patterns – and one group of experts urges caution in the use of network analysis for historical research (Graham et al., 2016:294) – there is still the pressing problem for readers to assess how valid the conclusions are.

Ruth and Sebastian Ahnert have been applying network analysis to the study of Early Modern European correspondence since 2015 (Ahnert & Ahnert, 2015), and have co-written an introductory book on the subject (Ahnert et al., 2021). In a recently annotated version of their 2015 article, they comment on the relationship between readability and reproducibility, acknowledging that the original article contained "short and accessible prose descriptions of the technical processes" but did not "get into the details of the code, which we thought might act as a barrier to our less technical readers." Anyone who wanted to reproduce their results would have needed to approach them for the data and code (Ahnert & Ahnert, 2021). For their current collaborations, they comment, they have realized that this is not enough, and are planning to make all the data and code "available alongside the research outcomes so that the results can be recreated, or the data/code adopted and used for other purposes." For the data, they note that this will be "where permitted", since their work is based in part on the commercial *State Papers Online* digital product (Ahnert et al., 2020).

### 5.3 Knowledge graphs

Similar issues arise for knowledge graphs, which also organize data into network-type structures based on formal ontologies supporting computational reasoning and querying across the graph. When this kind of Linked Open Data (LOD) approach is used as the basis for historical research, there are a number of different layers

deployed in managing the data which need to be explained and exposed before other researchers can reproduce the research. These layers are likely to include: the original data (if not in LOD-compatible formats), the transformation process and the transformed data (usually in RDF – Resource Description Framework), the data models and the ontologies on which they are based, and the methods and software used for publishing, accessing, querying, analysing, and visualizing the data.

The Mapping Manuscript Migrations project assembled a knowledge graph with historical data for nearly a quarter of a million medieval manuscripts, using a combination of the $FRBR_{oo}$ and CIDOC-CRM ontologies (Koho et al., 2021). Reasoning across the data to compare and visualize, for example, the sales patterns of particular British book dealers, or the changing page layouts over time of different types of liturgical manuscripts, involved the construction of a series of relatively complex SPARQL queries. Documenting this approach so that other researchers could reproduce the results required not only the exposure of the "raw" data as RDF triples, the publication of the data model used, and the provision of a public SPARQL endpoint to the data, but also the publication of the SPARQL queries themselves, together with documentation of the choices made in developing these queries in relation to the assumptions applied when constructing the dataset (Burrows et al., 2022).

### 5.4 Species models

In some cases, digital historical analysis is going beyond widely used approaches like topic modelling and network analysis. In 2022, Mike Kestemont and colleagues used "unseen species models" from ecology to estimate the proportion of medieval literary manuscripts which have survived to the present day (Kestemont et al., 2022a, b, c). The specific technique used – the Chao1 estimator – is a set of biostatistical calculations which can be implemented in various software environments. The authors of this paper argue that they have demonstrated the successful use of this approach, and comment that "these analyses call for a wider application of these methods across the heritage sciences."

The publication of this research in *Science* included a set of supplementary materials which gave an account of the methods used, explained the statistical calculations in more detail, provided additional figures and tables, and compared the "survival ratio" results with those produced with three other kinds of estimators, among them one which had been previously applied to the survival of early printed books (Green et al., 2011). Beyond this, however, lie the data and the code used to carry out this kind of analysis. The data have been published in a GitHub repository in the form of Excel spreadsheets, together with a set of Jupyter notebooks containing the Python code used for the analysis and additional "experiments" with R and SI code (Kestemont & Karsdorp, 2022a, b, Feb. 2). Releases of the GitHub materials have been mirrored on the public Zenodo repository (Kestemont et al., 2022a, b, c, Feb. 2). The software package *copia* which underlies this work has also been made available on GitHub with associated documentation (Kestemont & Karsdorp, 2022a, b, Aug. 22).

This is an exemplary approach to reproducibility, weakened only slightly by the lack of links to the data and code from the published paper and its supplementary materials. This information can be found from the "Open Science" section of an accompanying Web site for the Forgotten Books project (Kestemont et al., 2022a, b, c).

### 5.5 Visualizations

Visualizations are an important part of many of these kinds of analyses – particularly in relation to networks, maps, and statistics. While they may be used by researchers in a diagnostic sense for exploring the data, their more visible and influential use is for communicating views of the data and of the analyses (Drucker, 2020). In some cases, the visualizations can be produced by the same software which is used to organize, explore, and analyse the data. The *nodegoat* database software, for example, produces visualizations of geographical distributions and network graphs, with a chronological time slider (Nodegoat, 2022). In other cases, visualizations are produced by transporting (and sometimes transforming) the data into a different software environment. This is likely to be software specifically designed to produce visualizations, like *Gephi* or *Palladio*. But it may also be bespoke software produced for a particular project, such as the "Tudor Networks" visualization by Kim Albrecht (Ahnert et al., 2020).

In all these scenarios, being able to reproduce and evaluate the visualizations is likely to require an understanding of which software was used, and with what settings and configurations. If project-specific programming was involved, then the code will also be needed, perhaps in the form of a Jupyter notebook if Python or R have been used (Pryke, 2020).

### 5.6 Arguments

The white paper produced by the Arguing with Digital History working group (2017) makes the case that the selection and structuring of digital collections, and the descriptive metadata applied to them, are forms of historical argument in themselves. Visualizations are also incorporated into this broad definition, although the authors concede that an accompanying narrative may be necessary to make the argument explicit. They also observe that each step in the process of computational text analysis involves an interpretative act and "needs to be articulated in relation to the research question and sources justified as part of elaborating an argument" (Arguing with Digital History, 2017:11).

Digital processes like constructing collections and carrying out data analysis undoubtedly involve interpretative and contestable decisions and perspectives at each step of the way. Their products, in the form of graphs, tables, maps, and other visualizations, can be reproduced if the same set of data and the same software are used in the same order. Whether "arguments" of a more narrative and explanatory kind can also be reproduced and replicated is much more doubtful. Is this just

another way of saying that the author's arguments should be re-traceable to the sources and analyses on which they are based?

The authors of the CIDOC Conceptual Reference Model (CRM) have drafted "a formal ontology intended to be used as a global schema for integrating metadata about argumentation and inference making in descriptive and empirical sciences." The main purpose of CRMinf, as this ontology is titled, is to facilitate "the management, integration, mediation, interchange and access to data about reasoning by a description of the semantic relationships between the premises, conclusions and activities of reasoning" (Paveprime Ltd., 2019).

This might also be seen as a possible model for reproducing arguments and explanations in historical research, analogous perhaps to Fischer's typology of historians' fallacies. The only attempt to implement CRMinf in a working system to date seems to be in the British Museum's ResearchSpace software, where it is described as moving "from knowledge map to semantic narrative" (ResearchSpace, 2021). If this approach is more widely adopted, it may provide a basis for publishing data about historians' reasoning and arguments as well as about digital collections and methods.

### 5.7 New forms of publication

The discussion above draws on research which builds on the traditional model of publication, inasmuch as the data and code are made available alongside journal articles reporting the results of the research. The narrative article is linked with the supporting evidence through footnotes and references. A different approach aimed at integrating these elements more closely has been taken by the *Journal of Digital History*, launched in October 2021. It publishes articles which combine "data-driven scholarship and … transmedia storytelling", using a three-layered approach described in the following terms:

- a narrative layer exploring the possibilities of multimedia storytelling;
- a hermeneutic layer highlighting the methodological implications of using digital tools, data and code;
- a data layer providing access to data and making it reusable (when possible).

Of the nine research articles published in 2021–2022, only one includes an explicit "data download" section (Hoehne, 2022). Some refer to publicly accessible datasets and one uses data from Twitter, but for some it is unclear how to replicate or obtain the specific datasets used for the research.

The articles are assembled in the form of layered Jupyter notebooks, supporting programming in R and Python, although all those published so far make use of Python. The methodological aspects of the research projects are presented in a way that can be re-run in an open Jupyter environment like MyBinder. This is an imaginative and important contribution to the publication of digital history research environments, but the Juypter environment does come with significant learning curve, especially for writing up the research but also for reading the results.

## 6 Conclusions

History itself is not reproducible. Re-running or re-visiting historical events in order to study them directly is the stuff of science fiction narratives. At least one of these, Ernst Jünger's novel *Eumeswil*, envisages a computer-like machine which provides exactly this kind of reproducibility. Known as the "luminar", it enables a user to re-play historical scenes as if participating in them. Described as "conjuration" and "resurrection", this process summons the people of the past to re-enact historical events. "According to my whim," says the narrator, "I can sit with the Montagnards or the Girondists in the Convention, occupy the seat of the chairman or the concierge, who may have the best overview of the sitation. I am at once plaintiff, lawyer, and defendant – whichever I like" (Jünger, 1995:315). "Sometimes I play similar scenes at the luminar – say, from the history of the caesars, or the Russians before and after the revolution. I close the door, draw the curtains; I am in the abyss. Then I take over the role of the monarch – say, Nero, once he is notified that his bodyguards have left" (Jünger, 1995:156).

To reproduce historical research, rather than the past itself, researchers are expected to share the sources they used and the methods they deployed to develop their explanations, make their arguments, and produce their conclusions. In the era of digital history, this covers a wider range of different components than was previously the case. It is important to note that this means more than just sharing the data, however. Most universities provide digital repositories for researchers to deposit their datasets, together with guidance and policies for data sharing. This kind of guidance is also repeated in various publications.

The guide to *Doing Digital History* produced by the Institute for Historical Research at the University of London contains a great deal of helpful and sensible advice on data management and data sharing (Blaney et al., 2021:106–122). The authors suggest making data available for reuse, with suitable documentation, through an institutional or disciplinary repository rather than a bespoke Web site, while also recommending the use of Git for effective version control of files. They also discuss issues around licensing regimes for this kind of research data. But the question of reproducibility is not addressed explicitly; nor is the relationship between the data, the software used for analysis, and the arguments advanced in the published results of the research.

A somewhat earlier guide to software for "exploring big historical data" observes in passing that "historians are not accustomed to thinking about reproducibility, but it will become an issue" (Graham et al., 2016:156–157). The authors offer a cautionary tale of their own experience in trying (and failing) to reproduce some of their own data analysis using an Open Source tool, and advise researchers to archive a copy of the software they use (e.g., by forking it on GitHub) and to document carefully the conditions under which it worked. This, they note, "can be crucial to ensuring the reproducibility of your research" – although the context seems to be that of reproducing your own analysis, rather than publishing these materials to enable other researchers to evaluate your work.

For real reproducibility of digital historical research processes, several layers of evidence ought to be made available. At a minimum, what is needed in order to verify whether the conclusions reached are reasonable and supportable is access to the same set of data, the ability to run the same software processes, and the ability to compare the outputs from those processes. Making the data, the software, and the outputs available is only one aspect of these requirements, however. It is equally important to document the parameters of the processes, including software settings and configurations, and the specific states and versions of the data which were used.

Best practice for doing this can be extrapolated from some of the case studies above. As yet, these are the exceptions rather than the rule, and there are no agreed approaches to ensuring that reproducible components of digital historical research are documented and shared in consistent ways. But developing such approaches is becoming increasingly urgent, given the rapidly increasing amount of historical research which relies on digital methodologies. Evaluating the results of specific investigations and assessing the suitability of digital techniques require the documentation of the parameters and processes behind digital collections, digital analyses, and digital publications. Otherwise, the study of history runs the risk of developing a "reproducibility crisis" of its own (Bausell, 2021).

This approach to reproducibility in historical research sits within the broader context of government and university policies and practices relating to Open Science. Knöchelmann (2019), Arthur and Hearn (2021), and Gilby et al. (2022), amongst others, have argued that humanities researchers need to develop their own approach to the imperatives around "open research", recognizing that the Open Science framework can be difficult to translate into the different environment of the humanities. Gilby and colleagues, for instance, suggest that the FAIR (Free, Accessible, Interoperable, Reproducible) data principles in Open Science are better reconceptualized for the humanities as CORE (Collected, Organised, Recontextualised and Explained). Even when these concerns about difference and uniqueness are taken into account, however, the increasing use of digital methods for analysis and presentation makes reproducibility a relevant concept for the humanities, and for historical research in particular. Explaining and documenting the specific ways in which such methods have been applied is vital for assessing and validating that research. Being able to take the data and re-run the analyses and visualizations is an essential component in the evaluation of the interpretative narratives which rely on these outputs. Even if Open Science policies do not take the distinctive nature of the humanities sufficiently into account, there is now a clear case for the relevance of reproducibility to digital historical research.

**Authors' contributions** Not applicable.

**Funding** Open Access funding enabled and organized by CAUL and its Member Institutions

**Data availability** No data were generated or analyzed in the preparation of this paper.

## Declarations

**Ethical approval** Not applicable.

**Competing interests** The author has no relevant financial or non-financial interests to disclose.



## References

Ahnert, R., & Ahnert, S. (2015). Protestant letter networks in the reign of Mary I: A quantitative approach. *ELH, 82*(1), 1–33.

Ahnert, R. & Ahnert, S. (2021). Protestant letter networks in the reign of Mary I: A quantitative approach. Annotated version published in 2021 as part of *Models of Argument-Driven History*. https://doi.org/10.31835/ma.2021.04

Ahnert, R., Ahnert, S., & Albrecht, K. (2020). *Tudor networks.* Retrieved January 28, 2023, from http://tudornetworks.net/

Ahnert, R., Ahnert, S., Coleman, C., & Weingart, S. (2021). *The network turn: Changing perspectives in the humanities*. Cambridge University Press.

Arguing with Digital History working group (2017). Digital history and argument: White paper. Roy Rosenzweig Center for History and New Media. Retrieved January 28, 2023, from https://rrchnm.org/argument-white-paper/

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

Bausell, R. (2021). *The problem with science: The reproducibility crisis and what to do about it*. Oxford Academic. https://doi.org/10.1093/oso/9780197536537.001.0001

Blaney, J., Milligan, S., Steer, M., & Winters, J. (2021). *Doing digital history: A beginner's guide to working with text as data*. Manchester University Press.

Burguière, A. (2009). *The Annales school: An intellectual history*. Cornell University Press.

Burrows, T., Cleaver, L., Emery, D., Hyvönen, E., Koho, M., Ransom, L., Thomson, E., & Wijsman, H. (2022). Medieval manuscripts and their migrations: Using SPARQL to investigate the research potential of an aggregated knowledge graph. *Digital Medievalist*, 15(1). https://journal.digitalmedievalist.org/article/id/8064/

Candela, G., Sáez, M. D., Escobar Esteban, M. P., & Marco-Such, M. (2020). Reusing digital collections from GLAM institutions. *Journal of Information Science*. https://doi.org/10.1177/0165551520950246

Drucker, J. (2020). *Visualization and interpretation: Humanistic approaches to display*. MIT Press.

Fickers, A., & van der Heijden, T. (2020). Inside the trading zone: Thinkering in a digital history lab. *DHQ*, *14*(3). http://www.digitalhumanities.org/dhq/vol/14/3/000472/000472.html

Fischer, D. (1970). *Historians' fallacies: Toward a logic of historical thought*. Harper & Row.

Furet, F. (1985). Quantitative methods in history. In J. Le Goff & P. Nora (Eds.), *Constructing the past: Essays in historical methodology* (pp. 13–27). Cambridge University Press.

Gilby, E., Ammon, M., Leow, R., & Moore, S. (2022). *Open research and the arts and humanities: Opportunities and challenges*. University of Cambridge Working Group on Open Research in the Arts and Humanities. https://doi.org/10.17863/CAM.86734

Grafton, A. (1997). *The footnote: A curious history*. Harvard University Press.

Graham, S., Milligan, I., & Weingart, S. (2016). *Exploring big historical data: The historian's macroscope*. Imperial College Press.

Green, J., McIntyre, F., & Needham, P. (2011). The shape of incunable survival and statistical estimation of lost editions. *Papers of the Bibliographical Society of America, 105*, 141–175.

Guldi, J. (2019). Parliament's debates about infrastructure: An exercise in using dynamic topic models to synthesize historical change. *Technology and Culture, 60*(1), 1–33.

Guldi , J. & Williams, B. (2018). Synthesis and large-scale textual corpora: A nested topic model of britain's debates over landed property in the nineteenth century. *Current Research in Digital History 1*, 1. https://crdh.rrchnm.org/essays/v01-01-synthesis-and-large-scale-textual-corpora/

Haskins, L., & Jeffrey, K. (1990). *Understanding quantitative history*. MIT Press.

Hoehne, P. (2022). "Murderous, unwarrantable, and very cold": Mapping the rise of extralegal collective killing in the United States, 1783-1865. *Journal of Digital History*, *2*(1). https://doi.org/10.1515/JDH-2021-1007?locatt=label:JDHFUL

Jünger, E. (1995). *Eumeswil*. Quartet Books.

Karsdorp, F., Kestemont, M., & Riddell, A. (2021). *Humanities data analysis: Case studies with Python*. Princeton University Press.

Kestemont, M. & Karsdorp, F. (2022a). *Copia: Estimating the survival of cultural heritage artifacts with unseen species models from ecology.* Retrieved January 28, 2023, from https://github.com/mikekestemont/copia

Kestemont, M. & Karsdorp, F. (2022b). *Forgotten books* [code and data]. Retrieved January 28, 2023, from https://github.com/mikekestemont/forgotten-books

Kestemont, M., Karsdorp, F., de Bruijn, E., Driscoll, M., Kapitan, K., Macháin, P.Ó., Sawyer, D., Sleiderink, R., & Chao, A. (2022a). *Forgotten books: Supplementary materials (data and code).* Retrieved January 28, 2023, from https://zenodo.org/record/5947206

Kestemont, M., Karsdorp, F., de Bruijn, E., Driscoll, M., Kapitan, K., Macháin, P.Ó., Sawyer, D., Sleiderink, R., & Chao, A. (2022b). Forgotten books: The application of unseen species models to the survival of culture. Science, 375(6582), 765–769. https://www.science.org/doi/10.1126/science.abl7655

Kestemont, M., Karsdorp, F., de Bruijn, E., Driscoll, M., Kapitan, K., Macháin, P.Ó., Sawyer, D., Sleiderink, R., & Chao, A. (2022c). *Forgotten books: The application of unseen species models to the survival of culture*. [Web site]. Retrieved January 28, 2023, from https://forgotten-books.netlify.app/#details

Knöchelmann, M. (2019). Open Science in the humanities or: open humanities? *Publications 7*(4) 65. https://doi.org/10.3390/publications7040065

Koho, M., Burrows, T., Hyvönen, E., Ikkala, E., Page, K., Ransom, L., Tuominen, J., Emery, D., Fraas, M., Heller, B., Lewis, D., Morrison, A., Porte, G., Thomson, E., Velios, A., & Wijsman, H. (2021). Harmonizing and publishing heterogeneous premodern manuscript metadata as linked open data. *Journal of the Association for Information Science and Technology, 73*(2), 240–257. https://doi.org/10.1002/asi.24499

Ladurie, E. (1968). L'historien et l'ordinateur. *Le nouvel observateur,* 8 mai 1968, 2–3.

Ladurie, E. (1973). *Le territoire de l'historien*. Gallimard.

Latour, B. (1987). *Science in action: How to follow scientists and engineers through society*. Harvard University Press.

Macfarlane, A. (1980). *Records of an English village, Earls Colne, 1400–1750*. Chadwyck-Healey Ltd..

Macfarlane, A. (1983). Reconstructing historical communities with a computer: Final report to the Social Science Research Council. Retrieved January 28, 2023, from https://www.alanmacfarlane.com/TEXTS/Report.pdf

Macfarlane, A. (1992). Paper slips to computers: Notes on setting up the 'topics' database. (Unpublished paper c. 1992). Retrieved January 28, 2023, from https://www.alanmacfarlane.com/TEXTS/Connect2.pdf

Macfarlane, A. (1999). Only connect: Some thoughts on discovery and creativity. (Unpublished paper 1992, rev. 1999). Retrieved January 28, 2023, from https://www.alanmacfarlane.com/TEXTS/METHOD3.pdf

Macfarlane, A., Harrison, S., & Jardine, C. (1977). *Reconstructing historical communities*. Cambridge University Press.

Macfarlane, A., Porter, M., & Bryant, M. (1991). *The Cambridge Database System*, version 1.5. Retrieved January 28, 2023, from https://www.alanmacfarlane.com/TEXTS/CDS_manual.pdf

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. https://doi.org/10.3390/publications10040039

Nockels, J., Gooding, P., Ames, S., & Terras, M. (2022). Understanding the application of handwritten text recognition technology in heritage contexts: A systematic review of Transkribus in published research. *Archival Science, 22*, 367–392.

Nodegoat (2022). *Visualise your data*. Retrieved January 28, 2023, from https://nodegoat.net/guide.s/111/visualiseyourdata

Padilla, T., Allen, L., Frost, H., Potvin, S., Russey Roke, E., & Varner, S. (2019). *Final Report --- Always Already Computational: Collections as Data*. Retrieved January 28, 2023, from https://doi.org/10.5281/zenodo.3152935

Paveprime Ltd. (2019). *CRMinf: the argumentation model: An extension of CIDOC-CRM to support argumentation.* Version 0.10.1. Retrieved January 28, 2023, from https://cidoc-crm.org/crminf/sites/default/files/CRMinf%20ver%2010.1.pdf

Pawlicka-Deger, U. (2022). Infrastructuring digital humanities: On relational infrastructure and global reconfiguration of the field. *Digital Scholarship in the Humanities, 37*(2), 534–550. https://doi.org/10.1093/llc/fqab086

Powell, J. (1990). Introduction. In G. Iggers & J. Powell (Eds.), *Leopold von Ranke and the shaping of the historical discipline*. Syracuse University Press.

Pryke, B. (2020). *How to use Jupyter Notebook: A beginner's tutorial*. Dataquest. Retrieved January 28, 2023, from https://www.dataquest.io/blog/jupyter-notebook-tutorial/

Records of an English Village 1375–1854. (n.d.). Retrieved January 28, 2023, from https://wwwe.lib.cam.ac.uk/earls_colne/contents.htm

ResearchSpace (2021). *Argument & uncertainty*. British Museum. Retrieved January 28, 2023, from https://researchspace.org/argument/

Thomas, K. (2010). Diary: Working methods. *London Review of Books, 32*(11). 10 June 2010. https://www.lrb.co.uk/the-paper/v32/n11/keith-thomas/diary

Torgerson, J. (2022). Historical practice in the era of digital history. *History and Theory, 61*(4), 37–63.

University of Pennsylvania Libraries (2023). *OPenn: Read me*. Retrieved January 28, 2023, from https://openn.library.upenn.edu/ReadMe.html