Transactions of the Royal Historical Society (2026), 1–18
doi:10.1017/S0080440126100632

COMMENT

Show Me the Data: New Practices for Historical Sources

Ruth Ahnert[1], Katherine McDonough[2] and Daniel C. S. Wilson[3]

[1]Queen Mary University of London, London, UK; [2]Lancaster University, Lancaster, UK and [3]University College London, London, UK

Corresponding author: Ruth Ahnert; Email: r.r.ahnert@qmul.ac.uk

(Received 19 September 2025; revised 13 January 2026; accepted 27 January 2026)

Abstract

This comment examines the rapidly evolving ecosystem of historical research data in the United Kingdom, where cultural heritage collections are increasingly digitised, commercialised and fragmented. Historians face growing challenges in discovering, accessing and reusing data as resources move behind paywalls, and repositories remain scattered, without a national infrastructure to ensure long-term preservation or discoverability. Drawing on examples from major digital initiatives, we analyse the life cycle of historical research data and highlight the complex interplay of commercial, institutional and scholarly interests that shape access. We distinguish three types of data that emerge from historians' typical engagements with digitised collections: derived, enhanced and aggregated data. We argue that historians must actively participate in the practices relating to the creation, maintenance and reuse of such data. This will involve new forms of citation, favouring open datasets, improving digital skills and building communities around shared resources. The comment concludes with proposals to improve discoverability, sustainability and reuse, urging the discipline to establish common standards and infrastructures to secure an equitable data commons for future research.

Keywords: digitisation; historical research data; open data; digital humanities; collections as data; research infrastructure

Historians are accustomed to studying the workings of power, and in today's information ecosystem, power lies with data. Historical sources are increasingly becoming 'datafied' and accessed via digital platforms, which creates disparities in individuals' ability to access them.[1] Not only do commercial providers continue to concentrate resources behind paywalls (that only well-funded institutions and researchers can

[1]For a lucid overview see Katherine Bode and Lauren Goodlad, 'Data Worlds: An Introduction', *Critical AI*, 1 (2023). https://doi.org/10.1215/2834703X-10734026.

© The Author(s), 2026. Published by Cambridge University Press on behalf of The Royal Historical Society. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/S0080440126100632 Published online by Cambridge University Press

access), but the broader ecosystem of digitised sources and historical data has become so fragmented, complicated and opaque that it hinders both discovery and reuse.

The purpose of this comment is to provide an account of that complex ecosystem as it exists in the UK, and the challenges – as well as opportunities – it presents for historians, their practices and the future of the discipline. Although many examples exist, we draw on our own recent experience as users and creators of British historical datasets. We share the experience of having collaborated on the project 'Data/Culture: Building Sustainable Communities around Arts and Humanities Datasets and Software', which arose in part from the 'Living with Machines' (LwM) project, on which we also collaborated. LwM was a large-scale collaboration between data scientists, humanities scholars and cultural heritage professionals to examine the impact of industrialisation on the lives of ordinary people in nineteenth-century Britain by leveraging *at scale* cultural heritage collections, including textual sources (newspaper and books), visual sources (maps) and tabular data (census returns).[2] But as authors we bring wider experience in digital approaches to early modern archives, modern documents and manuscripts. While circumscribed by our specific trajectories, we engage with key issues that most historians are likely to grapple with regarding discoverability (can we find the data?) and accessibility (can we get hold of it?).

These twin issues are urgent because they concern cultural heritage as a public good. Heritage collections in libraries, archives and other repositories risk being removed from the commons as a consequence of commercially funded digitisation, if this means they become paywalled and no longer physically accessible.[3] We contend that the blanket notion of 'digital collections' needs to be unpicked, as it currently elides a wide variety of material and stages in the life cycle of digitisation and data creation, with different overlapping forms of ownership. The term 'data' is used in the remainder of this article in its most capacious sense to include things historians may not habitually consider as 'data' but which require preservation and are, therefore, integral to historical work. This includes information about provenance, databases, metadata and catalogue descriptions as well as (image) facsimiles of primary sources and digitised documents themselves. Our examples focus mainly on the digitisation of print materials, but the challenges and recommendations apply to historians working with, for example, pre-modern manuscripts or born-digital materials, as well as those considering as-yet-undigitised content, which still accounts for the vast majority of cultural heritage holdings at this time.[4]

[2]For an overview of the project see our website (https://livingwithmachines.ac.uk/) and open-access book (https://read.uolpress.co.uk/projects/living-with-machines) (both accessed 9 Jan. 2026).

[3]Notwithstanding the many barriers to accessing documents under existing regimes, certain materials have been understood to be in the public domain to the extent that they were accessible in research library collections or national archives. The trend we describe here goes against the principles of the Vancouver Statement on Collections as Data, that materials should be 'widely accessible, within the bounds of ethical, legal, and community expectations'. See Thomas Padilla, Hannah Scates Kettler, Stewart Varner and Yasmeen Shorish, 'Vancouver Statement on Collections-as-Data' (2023), 3. https://doi.org/10.5281/zenodo.8341519.

[4]Melissa Terras cites a Europeana study: 'While 82 percent of cultural heritage institutions across Europe have a digital collection or are engaged in digitization [...] after thirty years of large-scale investment in digital, on average 22 percent of heritage collections across Europe have been digitized, and only 58 percent of collections have even been cataloged in a collections database.' See Melissa

https://doi.org/10.1017/S0080440126100632 Published online by Cambridge University Press

Our second purpose is to show where opportunities and responsibilities exist for historians to change the way data is created and used. The article concludes with some pragmatic proposals: we suggest strategies for the community of historians in order to improve the discoverability, sustainability and reuse of data, and to secure equitable common resources for future research. We believe that historians should take ownership of this issue and should not be solely reliant on solutions offered by funders and cultural heritage institutions, notwithstanding recent investments in this area.[5] This is in part because the interests of these institutions are necessarily different from – and, sometimes at odds with – those of historians. A key issue is money. For example, the 2017 Mendoza review by the UK's Department for Culture Media and Sport (DCMS), called on museums to see themselves as 'cultural enterprises' in a climate in which real-term public funding had and would continue to decrease. The report recognised that cultural heritage institutions faced a burden to fund the digitisation of their holdings but steered them to take a pragmatic and 'mixed approach' to paying for it.[6] In reality, this meant partnering with commercial providers. Although such pragmatism helps institutions get their material online in some form, it comes at a cost: libraries whose mission statements claim 'to make our culture and heritage accessible to all to learn, research and enjoy' are – in cases where they choose to collaborate with commercial partners – in fact erecting new barriers around the digital surrogates of ostensibly public holdings.

It may be that money to support more open data will be forthcoming for assets that align with the UK Government's 'AI Opportunities Action Plan', published in January 2025. It states the urgency of 'unlocking' private and proprietary datasets, financing the creation of 'new high-value datasets that meet public sector, academia and startup needs', and explicitly names the potential of large-scale cultural and media data as training data for AI models, which could be developed in partnership with

Terras, 'Digital Humanities and Digitised Cultural Heritage', in *The Bloomsbury Handbook to the Digital Humanities,* ed. James O'Sullivan (2022), 255–66 (p. 256). https://doi.org/10.5040/9781350232143.ch-24. For the original study see Gerhard Jan Nauta, Wietske van den Heuvel, and Stephanie Teunisse, 'D4.4 Report on ENUMERATE Core Survey 4. Europeana DSI 2- Access to Digital Resources of European Heritage', *Europeana* (2017), https://pro.europeana.eu/files/Europeana_Professional/Projects/Project_list/ENUMERATE/deliverables/DSI-2_Deliverable%20D4.4_Europeana_Report%20on%20ENUMERATE%20Core%20Survey%204.pdf.

[5]The UK's Arts and Humanities Research Council (AHRC) recently funded the programme Towards a National Collection, which explored the possibilities of a unified, or networked system of repositories comprising cultural heritage collections. This programme set out to develop an 'inclusive, unified, accessible, interoperable and sustainable UK digital collection'. Rebecca Bailey, Javier Pereda, Chris Michaels and Tom Callahan, 'Unlocking the Potential of Digital Collections: A Call to Action: Towards a National Collection' (2024), 7. https://doi.org/10.5281/zenodo.13838916. A more recent development in parallel is the Museums Data Service, established in 2024, which aims to collate and share museum catalogue data online, see: https://museumdata.uk/ (accessed 9 Jan. 2026).

[6]'The Mendoza Review: An Independent Review of Museums in England', Department for Culture, Media and Sport and Department for Digital, Culture, Media & Sport (14 November 2017), https://www.gov.uk/government/publications/the-mendoza-review-an-independent-review-of-museums-in-england (accessed 9 Jan. 2026).

https://doi.org/10.1017/S0080440126100632 Published online by Cambridge University Press

'the National Archives, Natural History Museum, British Library and the BBC'.[7] While much historical research data may originate in such cultural heritage institutions, its digital life cycle necessarily involves a broader set of actors and stakeholders, including historians who, through their everyday practices, have long shaped the culture around research data rather than being its passive recipients. Indeed, for those historical researchers working in more data-driven areas, there is the new prospect of becoming data custodians, in a way that complements the existing role of cultural heritage institutions.

The life cycle of historical data

Our attempt to make sense of the kinds of data used and shaped by historians suggests an expanded conception of digitisation which goes beyond the custodianship of cultural heritage organisations or even (increasingly) the private photo archives of individual historians. Although the primary stewards of these documents – and their digital afterlives – remain cultural heritage institutions, their wider availability as digital assets will increasingly depend on actors beyond those institutions, including historians. In the following section we lay out a capacious account of historical data. Although it is by no means exhaustive, or representative of all data types our examples are selected to show the variety of opportunities for historians to share in creation, care and reuse of collections as data with our colleagues in cultural heritage organisations.

Digitisation and access

'Digitisation' is ambiguous: it could refer to the scans made of an original document or artefact; metadata about the object from a catalogue or finding aid; or searchable text or visual content from within it. However, each of these features represents different steps of processing and curation. In the case of textual sources, before there is text to search, an image of the document must be created, which might be a scan of a microfilm, a digital photo of a piece of paper, or a scan of a map. Creating a digital image and its metadata (and the conservation work that is often required before or after) may therefore represent only the beginning of a digitisation process, whereas for some materials it might also be its end. To create machine-readable information from an image ('document processing') might require further steps to produce searchable text. Each of these steps is computationally – as well as monetarily – expensive, not least because of the amount of labour required. Thereafter, cultural heritage organisations will also need to consider the costs of rights management (where necessary due to copyright or other restrictions), security, preservation, metadata curation, storage, discovery and data exploration interfaces. These will have implications for how the data is made available, and under what terms.

Broadly speaking, cultural heritage organisations have tended to provide access to digitised material in three ways (sometimes in parallel): direct access via institutional catalogue interfaces or linked digital library systems; online platforms offered as part

[7]Department for Science, Innovation & Technology, 'AI Opportunities Action Plan' (13 January 2025), https://www.gov.uk/government/publications/ai-opportunities-action-plan/ai-opportunities-action-plan (accessed 9 Jan. 2026).

https://doi.org/10.1017/S0080440126100632 Published online by Cambridge University Press

of a third-party, subscription-based product (usually with free in-person access on-site); or data repositories, unlinked to catalogue records.

The example of the British Library's (BL) newspaper collection combines a mix of all three approaches.[8] Although not a comprehensive or representative case study, it is a key national collection of enormous scale which, being familiar to many readers, makes it a useful object lesson. Over more than two decades, the BL has worked with public and private partners to support the gargantuan work of scanning and processing its collection of newspapers, which numbers *c.* 450 million pages. The result has been a patchwork of services and providers offering different access packages to different users. This patchwork allows us to see how the BL's strategies have changed over time, in response to different partnerships, and how these in turn throw light on the range of possible strategies pursued by UK cultural heritage institutions at large.

The UK's Joint Information Systems Committee (JISC) – a quasi-governmental non-profit, but which itself runs as a membership organisation – funded two rounds of newspaper digitisation in the mid-2000s, in partnership with the BL, as well as the academic publisher Gale.[9] These two projects set out to preserve and make available high-quality versions of historical British newspapers, amounting to *c.* 3 million pages selected according to criteria agreed as part of that digitisation project. However, the access arrangements were complicated from the outset, with libraries, universities and the general public offered different routes to the material, portending the splintering of the collection that has followed. Shortly after the conclusion of the second JISC project the BL entered into a partnership with a private genealogy company, FindMyPast, who have since gone on to digitise over thirty times as many newspaper pages as JISC (and counting), and which, outside of the Reading Rooms, are available to paying subscribers to the 'British Newspaper Archive' (BNA).[10]

The BNA search portal is now the default way to access historical British newspapers; however, confusingly, this portal also provides access to 'free to view' material digitised in parallel by publicly funded projects. Alongside the paywalled newspapers, one can also find (if one knows that relevant filters exist) newspapers digitised *ad hoc* by projects within the library, as well as external partners, such the Living with Machines project, on which we worked 2018–2023.[11] University researchers do not typically have access to the BNA because there is no institutional subscription option. Instead, they might have access to the eighteenth- and nineteenth-century collections digitised by Gale-Cengage (depending on the largesse of their library). These two

[8]Although access to public domain material via the catalogue remains awkward at the time of writing.

[9]Jane Shaw, '10 BILLION WORDS: THE BRITISH LIBRARY BRITISH NEWSPAPERS 1800–1900 PROJECT: Some Guidelines for Large-Scale Newspaper Digitization', in *International Newspaper Librarianship for the 21st Century*, ed. Hartmut Walravens (Berlin and New York, 2006), 27–44. https://doi.org/doi:10.1515/9783598440205.1.27.

[10]Over 95 million pages at the time of writing; subscriptions are currently £14.99 per month . See http://www.britishnewspaperarchive.co.uk (9 Jan. 2026). 'Free to view' pages are those funded publicly, and do not require a subscription.

[11]See Luke McKernan, https://web.archive.org/web/20191223141449/https://blogs.bl.uk/thenewsroom/2019/01/heritage-made-digital-the-newspapers.html (accessed 9 Jan. 2026); Giorga Tolfo et al., 'Hunting for Treasure: Living with Machines and the British Library Newspaper Collection', in *Digitised Newspapers – A New Eldorado for Historians?* ed. Estelle Bunout, Maud Ehrmann and Frédéric Clavert (Berlin, 2023), 25–47. https://doi.org/10.1515/9783110729214.

https://doi.org/10.1017/S0080440126100632 Published online by Cambridge University Press

platforms are emblematic of the types of products offered by commercial providers of cultural heritage collections across various document types: one caters to academic users (e.g. Gale or Proquest), where subscriptions are taken on by institutions, whereas the other serves a more general audience (e.g. FindMyPast and Ancestry), using a subscription model and search interface designed for the burgeoning genealogy and family history market. In both cases, users are restricted to browsing the collection or keyword search via a simplified web interface, rather than having direct access to the text of the newspapers in bulk – something that is necessary if historians wish to ask more nuanced corpus-wide questions than can be approached via simple search terms.[12]

In parallel with the development of subscription portals, the BL has more tentatively made those subsets of its newspaper collection not under commercial licence available in other ways. As an important actor within the 'collections as data' movement, the BL has an established history of sharing open data, which since 2018 they have been making available via the public British Library Research Repository. For example, portions of the open-access digitisation efforts undertaken internally and with external partners can be downloaded as XML files (e.g. the outputs from the scanning and OCR process) directly from this repository, albeit without the wrapper of a user-friendly interface offered by the BNA. Likewise, the full text of the JISC newspaper collections, mentioned above, can be requested from BL Labs.[13] The example of newspapers shows how cultural heritage organisations have tried to square the circle: to get their collections digitised for the benefit of future generations with the help of commercial partners, while also carving out free-to-access sections where possible. The latter enables projects such as Impresso to access the 'open' BL newspaper titles (i.e. those that are 'free to view' via the BNA), so that they might be included in the construction of an international newspaper corpus.[14]

That impetus can be seen in efforts across the sector, with repositories like the National Library of Scotland's 'Data Foundry' leading the way. Through this platform the NLS openly shares not only its organisational data (e.g. catalogue metadata) and digitised collection files (e.g. scans of periodicals, gazetteers, and encyclopaedias), but also research datasets derived from these resources by third parties, including historians.[15] However, with resources in the sector scarce – especially for smaller cultural heritage institutions – progress in this direction remains piecemeal despite

[12]More detailed accounts of the long history of newspaper digitisation can be found in Paul Fyfe, 'An Archaeology of Victorian Newspapers', *Victorian Periodicals Review*, 49 (29 Dec. 2016), 546–77. https://doi.org/10.1353/vpr.2016.0039; and Yann Ryan, 'Accessing and Using Historical Newspaper Data' (2023), https://yann-ryan.github.io/newspapers (accessed 9 Jan. 2026).

[13]https://labs.biblios.tech/ (accessed 26 Feb. 2025).

[14]Marten Düring, Matteo Romanello, Maud Ehrmann, Kaspar Beelen, Daniele Guido, Brecht Deseure, Estelle Bunout, Jana Keck and Petros Apostolopoulos., 'Impresso Text Reuse at Scale: An Interface for the Exploration of Text Reuse Data in Semantically Enriched Historical Newspapers', *Frontiers in Big Data,* 6 (Nov. 2023), https://doi.org/10.3389/fdata.2023.1249469. See also https://impresso-project.ch/news/2025/09/18/major-release.html (accessed 9 Jan. 2026).

[15]See https://data.nls.uk (accessed 9 Jan. 2026). The value of these datasets is underscored by the way they are integral to the Library Carpentry Text and Data Mining lesson. See http://librarycarpentry.org/lc-tdm/ (accessed 9 Jan. 2026). Sarah Ames and Stuart Lewis, 'Disrupting the Library: Digital Scholarship and Big Data at the National Library of Scotland', *Big Data & Society*, 7.2 (2020). https://doi.org/10.1177/2053951720970576. Tim Sherrat's GLAM Workbench has also made it easier for researchers to

https://doi.org/10.1017/S0080440126100632 Published online by Cambridge University Press

there being excellent precedents and models. The purpose of the UK's Towards a National Collection Programme (TaNC) was to scope the best route to a UK-wide Digital Collection; but the study undertaken by Daniel Belteki, Arran J. Rees, and Anna-Maria Sichani within TaNC highlights the manifold barriers to scaling up current efforts due to inconsistencies between institutions' operational practices and data standards as well as a lack of necessary technical or specialist skills (not to mention that AI has made navigating copyright 'an increasingly time-consuming, yet vital responsibility for cultural heritage organizations').[16] The creation of accessible digital assets requires 'complex interactions across institutional structures as well as between software suppliers, cultural heritage organisations, and third-party partners': which in turn require investments in technical and human infrastructure, as well as a willingness for cultural heritage institutions to collaborate and reach consensus.[17]

Historians creating data

The 'collections as data' movement has created a fruitful meeting point between institutions holding collections and researchers working with them. Historians might share in the creation of datasets as part of the digitisation process, either within projects (by securing funding for targeted digitisation) or even by taking what we might term a *self-service* approach (although this lies beyond the scope of this article).[18] In what follows, we draw attention to the kinds of data historians are involved in creating *following* the digitisation (or in the case of born-digital materials), which we categorise broadly into three types: *derived* data, *enhanced* data and *aggregated* data. Many historians derive, enhance and aggregate data without explicitly saying so. By formally recognising these activities we hope to set the stage for historians' engagement with data to be better acknowledged.

Derived data

This denotes new information or datasets created from existing ones, for example where data is subject to a degree of abstraction or transformation such that it constitutes a new dataset, potentially thereby becoming free of any copyright in existence on the original. For example, if a library makes available images of book pages, and *then* OCR software is used to create full-text data, the latter would be derived data. If the library already provides that text, a new dataset could be derived by processing the text in one of several ways to enable different forms of research. This could

reuse Australian data deposits using Jupyter Notebooks he developed, see https://glam-workbench.net/ (accessed 9 Jan. 2026).

[16]Daniel Belteki, Arran J. Rees, and Anna-Maria Sichani, 'Datafication and Cultural Heritage Collections Data Infrastructures: Critical Perspectives on Documentation, Cataloguing and Data-Sharing in Cultural Heritage Institutions', *Journal of Open Humanities Data*, 11.1 (2025), 14.

[17]*Ibid*.

[18]Melissa Terras notes that 'both the copyright exceptions that allow digital copies of material to be created for nonprofit data-mining research purposes, and the advancing technological infrastructures mean that, even with a modern smartphone and a few choice apps, high-quality data regarding historical collections can now be efficiently created for later processing and analysis'. Terras, 'Digital Humanities and Digitised Cultural Heritage', 256.

https://doi.org/10.1017/S0080440126100632 Published online by Cambridge University Press

include producing lists of named entities or word (or n-gram) frequencies. An n-gram is a sequence of *n* words which allows the analysis of texts based on word frequencies, but which does not preserve their position in the text and therefore prevents the original text being reproduced. An example of this is the n-gram data released for FindMyPast's BNA titles (published by the LwM team, and a previous collaborative project), comprising time series for the most frequently occurring words from the corpus, but without the full-text from which they were derived, because it remains licenced.[19]

Enhanced data

This denotes datasets that have been enhanced through processes such as cleaning, categorisation or enrichment with other features such as geo-coordinates, or linked open data, which link to stable identifiers on the web and ground references in sources to authoritative statements. A key feature of enhanced data is that it creates machine-readable data that can be processed computationally. An example of this is an enhanced version of the newspaper n-grams published by LwM colleagues mentioned above, which linked newspaper text to descriptive metadata for each newspaper title from which they were derived (such as places of publication, price and so forth).[20] Another example of enhanced data that will be familiar to historians of the nineteenth-century is the Integrated Census Microdata (I-CeM): individual-level census records for Great Britain from 1851 to 1911 (England and Wales, 1851–61 and 1881–1921; Scotland, 1851–1901). This dataset itself has a history comprising multiple transformations, undertaken by different organisations over time with profound implications for access and reuse. The original census enumerators' returns are held by The National Archives (TNA) and were scanned and processed by FMP in a similar arrangement to that struck with the BL regarding its newspaper collections. However, and in contrast to the case of the newspapers, these digitised census returns have been cleaned and coded by the I-CeM project (funded by the Economic and Social Research Council), which added consistent geographies over time and standardised coding schemes for many census variables, such as professions.[21]

A further example is the 'Tudor Networks of Power (TNoP) – Correspondence Network Dataset', co-created by one of the present authors.[22] The history of this dataset spans several institutions: the original letters are among the Tudor State Papers, mostly held by TNA, but also within BL collections and Hatfield House, amongst others. These were digitised by Gale as 'State Papers Online', which links scans of most

[19]As new pages are added to the underlying BNA collection, new versions of the derived data can be created and documented. Most recently, see Kaspar Beelen, 'NewsWords Data (Word Counts)', Zenodo (1 Mar. 2025). https://doi.org/10.5281/zenodo.14826348 released by LwM; and previously: Saatviga Sudhahar, Nello Cristianini, Thomas Lansdall-Welfare, and The FindMyPast Newspaper Team 'FindMyPast Yearly N-grams and Entities Dataset' (2016). https://doi.org/10.5523/bris.dobuvuu00mh51q773bo8ybkdz.

[20]Kaspar Beelen, 'NewsWords Data (Contextualized Word Counts)', Zenodo (9 Mar. 2025). https://doi.org/10.5281/zenodo.14996278.

[21]https://icem.ukdataservice.ac.uk/ (accessed 9 Jan. 2026).

[22]See Ruth Ahnert, Sebastian E. Ahnert, Jose Cree, and Lotte Fikkers, 'Tudor Networks of Power – Correspondence Network Dataset', https://doi.org/10.17863/CAM.99562, as part of the AHRC-funded Tudor Networks of Power (TNoP) project.

https://doi.org/10.1017/S0080440126100632 Published online by Cambridge University Press

of the manuscripts with the Calendars of the State Papers, a long-running project to catalogue the manuscripts chronologically, begun in the nineteenth century (and still incomplete). The TNoP team gained access to the underlying SPO XML data and scans, from which they extracted documents likely to be letters (i.e. documents with a full sender and recipient metadata field), and cleaned and enriched the metadata for these records. They de-duplicated and disambiguated the names (sender and recipient) fields, creating unique resource identifiers and providing LOD identifiers where available. They did the same for place names, which were also geo-referenced; dates were rationalised to the Gregorian calendar, where possible; and ambiguous time-windows narrowed. Finally, the network of communications was reconstructed, with people (the nodes) connected by letters (edges).[23]

Aggregated data

This denotes something akin to the production of traditional reference works, involving work to aggregate or categorise data (or both). The names of some typical examples reveal their affinity with more traditional reference works: the World Historical Gazetteer, the Universal Short Title Catalogue and Early Modern Letters Online (which is described as a union catalogue) extend the long histories of cataloguing and the compiling of gazetteers or directories.[24] An example of the latter from our own work is the Structured Timeline of Passenger Stations in Great Britain (SToPSGB): a dataset created by LwM.[25] The underlying dataset was aggregated from a vast range of sources by Michael Quick and published in his *Railway Passenger Stations in Great Britain: a Chronology*. The book is a directory listing passenger station locations, as well as dates of opening and closure, making it a uniquely rich and detailed account of Britain's changing railway infrastructure over time. With permission, the LwM project produced and released this dataset in order to provide the rich detail of this exceptional resource in a structured format, making it an open resource for historical research.[26]

Digital editions can be viewed as another example of data aggregation, bringing together documents that otherwise would not be accessible in one place. Editions traditionally published transcriptions of specific works or papers, collected according to some principle of selection such as theme or author. In both its analogue and digital manifestations editions thereby aggregate information by uniting scattered documents under one cover or web address. We might also note that digital edition projects are also examples of enhanced data as often these projects rely on the Text Encoding Initiative, which has, by virtue of setting common guidelines for publication,

[23]For a fuller account of the data provenance and cleaning process, see Ruth Ahnert and Sebastian E. Ahnert, *Tudor Networks of Power* (Oxford, 2023), 3–26.

[24]https://whgazetteer.org/; https://www.ustc.ac.uk/; http://emlo.bodleian.ox.ac.uk/ (all accessed 9 Jan. 2026).

[25]Mariona Coll Ardanuy, Kaspar Beelen, Jon Lawrence, Katherine McDonough and Federico Nanni, 'StopsGB: Structured Timeline of Passenger Stations in Great Britain', https://doi.org/10.23636/wvva-3d67.

[26]Mariona Coll Ardanuy, Kaspar Beelen, Jon Lawrence, Katherine McDonough, Federico Nanni, Joshua Rhodes, Giorgia Tolfo and Daniel C. S. Wilson, 'Station to Station: Linking and Enriching Historical British Railway Data', *Proceedings* http://ceur-ws. org ISSN 1613 (2021): 0073. https://ceur-ws.org/Vol-2989/long_paper29.pdf.

https://doi.org/10.1017/S0080440126100632 Published online by Cambridge University Press

enabled editorial labour to enter the digital age, underpinning important initiatives such as *The Women Writers Project* (a collection of early-modern women's writings) and the *Old Bailey Online* (which publishes London's central criminal court *Proceedings* from 1674 to 1913).[27]

These brief examples of derived, enhanced and aggregated datasets issue from academic research projects led by historians (or other humanities scholars), and are designed for use primarily by historians. Sometimes these projects involved active collaboration with colleagues within cultural heritage organisations (e.g. the outputs of the LwM project), but others can be seen more as an asynchronous form of collaboration, building on the outputs of library and archive digitisation (e.g. the census microdata of I-CeM, or the Tudor letters dataset of TNoP), or cataloguing (e.g. the union catalogue of EMLO). These asynchronous engagements, however, often create new stages of life for the data, often away from the institutional homes in which the physical documents reside. In developing these datasets, historians must begin to address questions of how they are made accessible and maintained – practices that historians have not customarily been trained to perform. As the next section suggests, the variety of ways in which historical data is currently being made available to potential users makes for a complex set of practices, access arrangements and possible futures for the discipline, which requires us to re-imagine what a data 'commons' might look like.

Show me the data

As things stand presently, the great effort of creating the derived, enhanced or aggregated datasets just discussed does not always translate into benefits to the wider community of historians because of a lack of agreed-upon practices, policies and venues for their deposit. Notwithstanding the creation of principles such as FAIR (the recommendation that data be Findable, Accessible, Interoperable, and Reuseable), these require a reliably-funded, common infrastructure in order to function in practice.[28] Historians may wonder why they should care about such dry, technical issues, but in fact many of these resources will benefit both digital history and more traditional forms of scholarship by making sources available and interrogable in new ways.

Some of the best-known digital projects have published their data on self-hosted platforms, including aforementioned examples such as *World Historical Gazetteer, The Women Writers Project* and *Old Bailey Online*. Perhaps due to the prominence and success of these examples, amongst various others, many new initiatives entering the digital space will aspire to this form of output – despite the many horror stories of legacy projects delivering '404 not found' messages when grant funding dries up and institutional support is not forthcoming.[29] Such approaches will rarely be feasible in the future because they require funding on a much larger scale than most

[27] https://tei-c.org/; https://www.wwp.northeastern.edu/; https://www.oldbaileyonline.org/ (all accessed 9 Jan. 2026).

[28] See https://ukdataservice.ac.uk/learning-hub/research-data-management/plan-to-share/fair-data-principles/ (accessed 9 Jan. 2026).

[29] In response to by now well-documented losses of digital projects, see the proposals of *The Endings Project*: https://endings.uvic.ca/index.html (accessed 9 Jan. 2026).

https://doi.org/10.1017/S0080440126100632 Published online by Cambridge University Press

researchers or projects can hope to secure and sustain, especially in the current climate in research and Higher Education in the UK (and around the world), which faces indefinite austerity.

Researchers may wish then, at a minimum, to deposit their data in a repository: but which repository is the right one? To take the example of I-CeM census microdata: it was a requirement of ESRC funding that the project data be deposited with the UK Data Service (UKDS), where they can be accessed by registered users, for the purposes of research sustainability, reproducibility and reuse. This policy aims both to ensure public return on investments, and to ward against proprietary attitudes to data.[30] Moreover, it presents a centralised home for economic and social datasets because non-ESRC-funded datasets can also be deposited subject to panel approval. However, not all data in this repository is open. For example, in the case of the I-CeM data, compromises have been forced by the commercial interests of FindMyPast, who invoked a clause to 'safeguard' the data. This means that users must either opt to use an anonymised version, or apply for a special licence to access the full dataset – a mechanism, as Richard Rodger has pointed out, that effectively excludes the public.[31]

At present there does not exist an equivalent repository for the wider discipline of History. The Arts and Humanities Data Service (AHDS) was established in 1996 by JISC and the AHRC, but was decommissioned in 2008 – before its potential was realised – on the basis that researchers falling under their remit could deposit their data either in their university's repositories, or in one of several other centralised repositories.[32] The centralised repositories purported to meet this need included the Archaeology Data Service (originally one of the five disciplinary-defined distributed services that made up AHDS), the Literary and Linguistic Data Service (which now houses the Oxford Text Archive collections),[33] the Shared Research Repository for cultural and heritage organisations (which stores the research outputs of the BL and current partner organisations: the British Museum, Museum of London Archaeology, National Museums Scotland, National Trust, Science Museum Group and Royal Botanic Gardens, Kew)[34] and the more recently established DiSSCo UK (Distributed System of Scientific Collections, a ten-year £155.6 million infrastructure programme to digitise natural science collections, involving up to ninety organisations across the UK).[35]

This patchwork of provision may serve cultural heritage partners to some degree, but is not easily comprehensible to historians and remains vulnerable, as some of the

[30]For discussion of proprietary behaviour in the sciences, see Jasmine Jamshidi-Naeini et al, 'Guest Editorial: Data availability statements', COPE: Committee on Publication Ethics (24 July 2023), https://publicationethics.org/news/guest-editorial-data-availability-statements.

[31]Richard Rodger, 'Making the Census Count: Revealing Edinburgh 1760–1900', *Journal of Scottish Historical Studies*, 40 (2020), 134–48.

[32]AHDS was established in 1996 and ceased operation in 2008. For details of the archiving of the website and data, see https://ahds.ac.uk/ (accessed 9 Jan. 2026). For comparison, the US's National Endowment for the Humanities recently announced Knowledge Commons as its public access repository of the humanities. See https://about.hcommons.org/2024/12/10/kcworks-named-designated-public-access-repository-of-the-national-endowment-for-the-humanities/ (accessed 9 Jan. 2026).

[33]https://llds.ling-phil.ox.ac.uk/llds/xmlui/ (accessed 9 Jan. 2026).

[34]https://iro.bl.uk/ (accessed 9 Jan. 2026).

[35]https://dissco-uk.org/ (accessed 9 Jan. 2026).

https://doi.org/10.1017/S0080440126100632 Published online by Cambridge University Press

components rely on continuing institutional support.[36] At present there is no requirement for AHRC-funded projects to deposit their data publicly; it would be hard to mandate such a requirement in the absence of a national system for doing so. Such infrastructure would not only provide security and longevity, but also set the standard for unfunded research, helping establish best practice for data deposit by historians and others. The UK Government's *AI Opportunities Action Plan* gestured at the desideratum for a centralised cultural repository of data, but in specific relation to training 'A.I.' models.[37] It remains unclear what is being envisaged in this regard, but the framing of this suggestion has already received criticism as a 'sell-off' of UK data and passes over the existing and more immediate needs of researchers to share data with their communities.[38]

What then are the options for historians creating data today? General-purpose repositories such as Zenodo and Figshare are options sometimes used for research outputs that do not fit into discipline-specific repositories, on an ad hoc basis.[39] Likewise, many universities will have their own repositories, but how discoverable these are is open to question.[40] A set of thematically focused data collectives have also emerged in recent years, which publish period-specific datasets, including the Post45 and (fledgling) Nineteenth-Century Data Collective.[41] These sit alongside journals such as *The Journal of Open Humanities Data (JOHD), The Journal of Cultural Analytics*, and the even newer DARIAH journal *Transformations*, which have encouraged a category of publication known as a 'data paper', which provides a citable object that not only leads to deposited datasets but also documents how data was prepared and might be reused.[42]

This brief overview aims to show that, although many scholars are thinking carefully about ways to ensure that data is both available and well documented, the plethora of options for doing this means that resources remain scattered, often unindexed, and therefore hard to find, especially for the uninitiated. This impacts discoverability and reuse, and means a poor return on investment for publicly-funded projects. Historians – both those with and without digital skills – often do not know about the existence of datasets that could or would be of use to them, let alone where to find them.[43] Our purpose here is not to persuade readers that they must embrace

[36]Dorothea Strecker, Heinz Pampel, Rouven Schabinger, Nina Leonie Weisweiler, 'Disappearing Repositories: Taking an Infrastructure Perspective on the Long-Term Availability of Research Data', *Quantitative Science Studies*, 4 (2023), 839–56. We might look to the example of the Oxford Text Archive, which needed to find a new home when the Bodleian Library could no longer host it.

[37]See https://www.gov.uk/government/publications/ai-opportunities-action-plan/ai-opportunities-action-plan (accessed 9 Jan. 2026).

[38]See for example, Jon Agar's comment: https://web.archive.org/web/20251228140217/https://bsky.app/profile/jonagar.bsky.social/post/3lfmnsyhodk2l (accessed 9 Jan. 2026).

[39]https://help.zenodo.org/docs/get-started/ (accessed 9 Jan. 2026).

[40]For instance, the 'Tudor Networks of Power – Correspondence Network Dataset' was voluntarily deposited in the University of Cambridge's repository, 'Apollo', by its creators https://www.repository.cam.ac.uk/items/0be239a0-909c-44d7-a4bc-39f9e8668f40 (accessed 9 Jan. 2026).

[41]https://data.post45.org/ and https://c19datacollective.com/ (both accessed 9 Jan. 2026).

[42]See https://openhumanitiesdata.metajnl.com/, https://culturalanalytics.org/about, and https://transformations.episciences.org/ (accessed 9 Jan. 2026). The *CA* dataverse can be found here: https://dataverse.harvard.edu/dataverse/culturalanalytics (accessed 9 Jan. 2026).

[43]There is a certain irony that the research environment in the UK now expects most outputs to be published open access, but there are still few structures in place to ensure the data that underlies

https://doi.org/10.1017/S0080440126100632 Published online by Cambridge University Press

digital methods in their research (although we do believe that tentative engagement is part of the solution, as we outline below);[44] our primary point is that the efforts of scholars creating digital resources and open data are to the benefit of the entire discipline because of their potential to democratise access to historical knowledge. But for those benefits to be fully realised, the wider discipline needs to engage with how they are maintained and made accessible.

Proposals

So what is to be done? In this final section we set out a series of principles that we believe will improve the access to and reuse of historical research data. These are pragmatic because they work within the current structural challenges that we outline above. In an ideal world our institutions and infrastructure would receive the necessary investment to support the digitisation, storage and discovery of research data in a sustainable and equitable way. Unfortunately, such investment appears unlikely, notwithstanding the relatively modest proposals set out in research council-funded reports, such as those from the TaNC initiative.[45] High level changes may perhaps be more likely, due to governmental pushes to facilitate the AI-industry in the UK, with its requirements for ever more training data. Such initiatives are likely to be problematic, but may end up providing unintended benefits. Trapped between the Scylla of austerity and the Charybdis of AI, historians can nevertheless find ways to take responsibility for their practices in relation to data such that they are not dictated by the needs of other institutions and corporations.

The following proposals are intended as a means of encouraging change through our actions as a community, which we divide according to two imagined historian-personae wanting either to *use* or to *create* historical research data. In terms of creation we are still focusing on the derived, enhanced or aggregated data that builds on prior digitisation or cataloguing efforts by cultural heritage organisations, rather than the self-service digitisation approach.

For users of data

1. *Change citation practices to account for the nature of data.* Historians are trained in the art of finding and citing research materials. However, best practices for which version of a non-unique document is cited remain inchoate, despite important research and advocacy around citation practices for physical versus digitised copies.[46] For many, it is still the norm to cite the material document that sits behind a digital resource rather than the digital document or dataset

those publications meets the same standards. On why closed data is bad for good research, see Nathalie Cooke and Ronny Litvack-Katzman, 'Open Times: The Future of Critique in the Age of (Un)replicability', *International Journal of Digital Humanities,* 5 (2024), 2–3. https://doi.org/10.1007/s42803-023-00081-y.

[44]On the failure of 'digital historians', since the early postwar period, to persuade significant numbers of historians to take the digital turn, see Max Kemman, *Trading Zones of Digital History* (Berlin, 2021), 1–38.

[45]See Bailey et al., 'Unlocking the Potential of Digital Collections'.

[46]Jonathan Blaney and Judith Siefring, 'A Culture of Non-Citation: Assessing the Digital Impact of British History Online and the Early English Books Online Text Creation Partnership', *Digital Humanities Quarterly* 11 (2017), https://www.digitalhumanities.org/dhq/vol/11/1/000282/000282.html; Meaghan Brown, Paige Morgan and Jessica Otis, 'Identifying Early Modern Books: Challenges for Citation Practices

https://doi.org/10.1017/S0080440126100632 Published online by Cambridge University Press

(e.g. the edition of the early printed book, rather than *Early English Books Online*). This not only obscures what the researcher has actually observed in order to make their claim; it also hides how hugely reliant we are now as a research community on the products of digitisation and online portals. This occlusion means that it remains extremely hard to measure how much historical research data (in terms of both commercial platforms and data deposited on repositories) is actually being used and contributing to research. Historians must play a role in demonstrating the value of digitised collections by choosing to cite the digital sources. Doing so allows data creators such as cultural heritage organisations and research projects to measure the use of their data and demonstrate their impact.

2. *Choose open data over paywalled versions where possible.* In addition to choosing to cite the digital version of documents, we suggest that where possible, researchers should seek out open rather than paywalled versions of data. We must, in other words, vote with our feet. This may mean some level of compromise, by asking whether a similar data source that is open access could be chosen instead of one that is perhaps more readily accessible, but behind a paywall. For example, if researchers want to access historical British newspapers, as we already outlined above, there are alternatives to the paid-for services of Gale-Cengage and the British Newspaper Archive (BNA) in the form of freely available full text versions, available on the British Library's Research Repository.[47] There will be equivalents for other sources and data types, although we recognise that this recommendation will not be straightforward for many practising historians, who may know neither how to manipulate and search such files, nor that this data even exists outside of the commercial platforms and portals to which they are ordinarily directed. However, with the right support and training where required, this may in fact involve a lower overall institutional cost; for example, where there is no or limited access to institutional subscriptions for paid-for platforms. The behavioural shift in seeking out open-access data will help provide cultural heritage colleagues and policy makers with evidence and encouragement in arguing for open-access digitisation and discovery platforms in future. As things stand, paywalled solutions – with their many limitations – will continue to be offered as a primary point of access.
3. *Educate yourselves and your students in open-access data.* Demonstrating the value of open-access data is essential to change behaviour at a disciplinary level. Educators should be familiarising themselves with the data available through open-access repositories and passing this information on to their students as part of training in research skills. We suggest that the availability of new data should, to some extent, be steering our research and teaching, and that the availability of new open datasets could spark research and ideas for student

in Book History and Early Modern Studies', *Archives Journal* (Nov. 2017), https://www.archivejournal.net/essays/identifying-early-modern-books/.

[47]See, for example, https://www.britishnewspaperarchive.co.uk/titles/stockton-herald-south-durham-and-cleveland-advertiser (accessed 9 Jan. 2026).

https://doi.org/10.1017/S0080440126100632 Published online by Cambridge University Press

projects.[48] If a new generation of scholars change their behaviours by rejecting subscription models, and collaborating with colleagues in university libraries on initiatives to find and access open data, it has the potential to cause change.

4. *Improve digital skills and collaborate.* How are historians to work with this open-access data once they have it? For most historians, accustomed to user interfaces with keyword search boxes, the prospect of working directly with data may feel daunting. However, open-access datasets can be rendered navigable with off-the-shelf tools, which allow researchers to move beyond basic search and browse functions, to perform more nuanced and powerful queries, tailored to their research questions. This is, perhaps, the bigger change in behaviour, and as such will require digital skills to be baked into the undergraduate curriculum. Established researchers who are amenable to such approaches will need either to undertake a degree of self-education, or to seek out and establish collaborations with new colleagues. There are a multitude of online and in-person opportunities available to the curious (such as *The Programming Historian* or the *Oxford Digital Humanities Summer School*), while those in universities may have access to advice from or opportunities for collaboration with colleagues in IT, Research Software Engineering, the Library, Computer Science, Linguistics, and Digital Humanities.[49] Joining forces with colleagues with computational skills and experience may not require funding in the first instance if historians can seek out areas of complementary scholarly interest. The effort involved in such collaborations lies in establishing relationships, and working out the 'trading zones' between your disciplinary interests in ways whose value is hard to predict in advance, but which may be highly rewarding.[50]

For creators of data

5. *Create canonical datasets.* FAIR research data principles recommend that data should be Findable, Accessible, Interoperable and Reusable.[51] To ensure that data is *accessible*, we encourage colleagues to work with their communities to develop canonical, open-access datasets. In many fields certain online databases have become, by default, the scholarly standard; however, where these are behind paywalls, they are not accessible to all. Not only is this not equitable, it also means that any historical arguments based on such documents are not

[48] Ian Milligan has highlighted the extent to which decisions about which newspapers get scanned have shaped which ones are used as sources in Canadian History Ph.D.s and journal articles in the *Canadian Historical Review*: Ian Milligan, 'Illusionary Order: Online Databases, Optical Character Recognition, and Canadian History, 1997–2010', *Canadian Historical Review*, 94 (2013), 540–69.

[49] See https://programminghistorian.org/ (accessed 9 Jan. 2026) for an ever-growing archive of hands-on, peer-reviewed tutorials.

[50] For further discussion about establishing interdisciplinary collaborations, and exploration of Peter Galison's concept of 'trading zones', see Kemman, *Trading Zones*; and Ruth Ahnert, Emma Griffin, Mia Ridge and Giorgia Tolfo, *Collaborative Historical Research in the Age of Big Data: Lessons for Interdisciplinary Collaboration* (Cambridge, 2023), ch. 1. https://doi.org/10.1017/9781009175548.

[51] https://www.go-fair.org/fair-principles/ (accessed 9 Jan. 2026). These principles have been extended by the Global Indigenous Data Alliance to consider the dimensions of Collective Benefit, Authority to Control, Responsibility, and Ethics, see https://www.gida-global.org/care (accessed 9 Jan. 2026).

https://doi.org/10.1017/S0080440126100632 Published online by Cambridge University Press

necessarily reproducible (which is increasingly an expectation for rigorous *digital* historical research). We advocate community efforts to establish alternative datasets that are openly available to all. This may involve compromising on their size; or, having only derived, abstracted features, such as n-grams for research data that may otherwise be restricted, due to commercial interests or copyright. One of the additional benefits of focusing attention on building canonical datasets is that they can be selected for quality, or to meet certain standards of representativeness that a larger dataset might not be able to provide – issues that frequently dog historical collections, digitised or otherwise. Initiatives in this direction also enable collective effort to ensure iterative approaches to data cleaning or curation, and documentation. The benefits for scholarship should be clear: if research in digital or computationally inflected history is to develop a rich critical literature that parallels existing historiographies, then it must become possible for communities of scholars to gather around particular datasets (much as historians have returned to the same archives) in order to extend and challenge each other's interpretations. This will only be possible if such datasets are *freely* open and available, without gatekeeping (such as making data available only on request). For a model of how this can work, we can look to computational literary studies, where Andrew Piper, Ted Underwood and their collaborators have published derived datasets from the Hathi Trust, which have been used at the centre of numerous studies.[52]

6. *Consider the discoverability of your data.* We know that the proliferation of repositories is a barrier to data being *findable* (the first of the FAIR data principles). As the TaNC programme explores, having a national collections portal is a powerful way to facilitate discoverability (as well as interoperability); but the recommendations of TaNC will take time to be enacted, and there is unlikely to be the scale of funding to do for humanities data what DISSCO UK will do to unify scientific collections in a single platform. More realistically, we would recommend that sub-communities within history work to establish local norms by choosing to submit their research data to the same locations, which, crucially, should provide a durable Digital Object Identifier (DOI). The coverage of UKDS beyond ESRC-funded projects has been user-driven and could easily be expanded with coordination; and it has the benefit of resources to help users think about how qualitative data can be stored, contextualised and accessed.[53] But equally a coordinated effort might opt for an institutional or disciplinary repository, the repo of the cultural heritage institution where the data originated, or even an open science repository such as Zenodo. As noted already above, efforts to create discipline or period-specific communities have been trialled by the Post45 and Nineteenth-century Data collectives. The fact that the latter has only three

[52]See, for example, Sunyam Bagga and Andrew Piper, 'HATHI 1M: Introducing a Million Page Historical Prose Dataset in English from the Hathi Trust', *Journal of Open Humanities Data*, 8 (2022), https://doi.org/10.5334/johd.71; Sil Hamilton, and Andrew Piper, 'Multihathi: A Complete Collection of Multilingual Prose Fiction in the Hathitrust Digital Library', *Journal of Open Humanities Data*, 9 (2023), https://doi.org/10.5334/johd.95; and Ted Underwood, Patrick Kimutis, and Jessica Witte, 'NovelTM Datasets for English-Language Fiction, 1700–2009', *Journal of Cultural Analytics,* 5 (2020), https://doi.org/10.22148/001c.13147.

[53]https://ukdataservice.ac.uk/learning-hub/qualitative-data/ (accessed 9 Jan. 2026).

https://doi.org/10.1017/S0080440126100632 Published online by Cambridge University Press

datasets goes to show that establishing new community practices can be a slow process, even with willing collaborators. We therefore suggest that any future efforts in this direction should seek the support of national and international associations and learned societies, such as the Royal Historical Society, perhaps underscored by updated requirements to publish data from the journals linked to these organisations.

7. *Create greater awareness of new datasets to drive reuse.* Depositing data without signposting it is like burying it. One key way of raising the profile of new historical research data is to share it as a data paper in a venue such as *The Journal of Cultural Analytics* or *JOHD*. Data papers make datasets more accessible and discoverable – both practically and *intellectually* – to other historians. By providing an account of a dataset's provenance, structure, affordances and value for future research, data papers pave the way for research data to be more readily *reusable* (the fourth of the FAIR principles). Perhaps just as importantly, these accounts also make visible the labour of the data creators for considerations such as promotion. Other modes of signposting may include the creation of published links to datasets' DOI through cross depositing, indexing via reporting mechanisms (e.g. Researchfish, which populates the UKRI 'Gateway to Research', although this is to be shut down in 2027)[54] and citations in linked publications. The more links we make, the more pathways there are to the data.
8. *Build communities around your data.* The fourth of the FAIR principles – reusability – can also be supported by building communities around datasets Our experience collaborating on the Data/Culture project has demonstrated that some of the most effective ways to build and maintain a community is by delivering workshops and training, publishing tutorials and undertaking 'community calls' or other mechanisms for providing ongoing support to users of a given dataset.[55] In our case the community call mechanism was trialled on a piece of software we developed on LwM (MapReader). We met online every month to share examples of ways our material was being reused by others, as well as offering technical support and critical discussion of the software itself. This format is common among software developers but could easily be adapted to build community and collaborations around a given historical dataset. The benefit of building such a community is that it supports the canonisation of research datasets by driving reuse and citation, as well as creating a community of practice.[56]
9. *Develop and support mapping exercises.* While there is hope that future investment may provide a one-stop solution for finding historical data, it is likely to be steered to some extent by the UK government's AI agenda, and will probably be delivered in partnership with cultural heritage institutions.[57] Because

[54] https://gtr.ukri.org/ (accessed 9 Jan. 2026).

[55] https://web.archive.org/web/20241111143907/https://www.turing.ac.uk/research/research-projects/dataculture-building-sustainable-communities-around-arts-and-humanities (accessed 9 Jan. 2026).

[56] Daniel C. S. Wilson, 'Working at Scale: What Do Computational Methods Mean for Research Using Cases, Models and Collections?', *Science Museum Group Journal* 18, (2023). https://doi.org/10.15180/221805.

[57] See Bailey et al., 'Unlocking the Potential of Digital Collections'.

https://doi.org/10.1017/S0080440126100632 Published online by Cambridge University Press

these interests may not necessarily align with the interests and needs of historians, we suggest that mapping exercises are essential to record the development of the wider data landscape. This may need to be implemented from below by historians. The Data/Culture project suggested to the AHRC that existing reporting mechanisms could be harnessed for the purposes of indexing UK arts and humanities research data. UKRI-funded projects already have a reporting requirement through Researchfish. Despite known problems with the interface and the UK's 'Gateway to Research' website, the reporting platform could be used to collect new datasets for a central index automatically, which could then link out to distributed repositories. Mapping data generated by *completed* projects is a harder task. One route would be to audit all projects funded over the last twenty years and to check which have resulted in data outputs, if they are still accessible, and if so, where they currently reside. Similar projects could be undertaken in other geographical areas in collaboration with other national funding bodies. More nuance could be provided to the resultant picture via a crowdsourced mapping exercise of data assets known and used within the broader communities whose work has been research-council funded. Comparing these two approaches would help show which projects have been effective at creating and sharing data, and which have allowed data to be forgotten or kept private; and relatedly, what factors distinguish one type of project from the other.

These strategies are needed due to the absence of systematic national policies and infrastructures for storing and accessing data created by and for historians. Of course, as we have gestured above, the slow progress towards existing recommendations might be set to change due to the big promises that are being made in the UK to support AI innovation. It might be that some of these endeavours will indirectly benefit the history community by pulling cultural heritage assets out from behind paywalls, and by financing the creation of new open datasets. Arguably our proposals are even more pertinent if those promises are delivered on, because they are being designed to serve different constituencies. Because the data landscape is likely to change rapidly in the coming years, we need to establish a set of practices that ensure the data we are creating and will rely on as a discipline is available for the entire community to use.

Author contributions. Ahnert acted as first author, with substantial drafting and revision of the manuscript from Wilson who, together with McDonough, helped formulate and draft the argument. All authors edited and reviewed the final manuscript.

Acknowledgements. Our thanks to Pieter Francois and Kalle Westerling, as well as the journal's two anonymous reviewers, for their comments and feedback on earlier versions of this manuscript.

Financial support. This work emerges from the project 'Data/Culture: Building sustainable communities around Arts and Humanities datasets and software' (AH/Y00745X/1, PI Peter Francois), which was funded by the Arts and Humanities Research Council, UK.

Cite this article: Ruth Ahnert, Katherine McDonough and Daniel C. S. Wilson, 'Show Me the Data: New Practices for Historical Sources,' *Transactions of the Royal Historical Society* (2026), 1–18. https://doi.org/10.1017/S0080440126100632

https://doi.org/10.1017/S0080440126100632 Published online by Cambridge University Press