Sussex Research Online

Enabling complex analysis of large-scale digital collections: humanities research, high performance computing, and transforming access to British Library digital collections

Article (Published Version)

Terras, Melissa, Baker, James, Hetherington, James, Beavan, David, Welsh, Anne, O'Neill, Helen, Finley, Will, Duke-Williams, Oliver and Farquhar, Adam (2018) Enabling complex analysis of large-scale digital collections: humanities research, high performance computing, and transforming access to British Library digital collections. Digital Scholarship in the Humanities, 33 (2). pp. 456-466. ISSN 2055-7671

This version is available from Sussex Research Online: http://sro.sussex.ac.uk/id/eprint/67144/

This document is made available in accordance with publisher policies and may differ from the published version or from the version of record. If you wish to cite this item you are advised to consult the publisher's version. Please see the URL above for details on accessing the published version.

Copyright and reuse:
Sussex Research Online is a digital repository of the research output of the University.

Copyright and all moral rights to the version of the paper presented here belong to the individual author(s) and/or other copyright owners. To the extent reasonable and practicable, the material made available in SRO has been checked for eligibility before being made available.

Copies of full text items generally can be reproduced, displayed or performed and given to third parties in any format or medium for personal research or study, educational, or not-for-profit purposes without prior permission or charge, provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.

http://sro.sussex.ac.uk

Enabling complex analysis of large-scale digital collections: humanities research, high-performance computing, and transforming access to British Library digital collections

Melissa Terras
Department of Information Studies, University College London, UK and UCL Centre for Digital Humanities, University College London, UK

James Baker
School of History, Art History and Philosophy, University of Sussex, UK

James Hetherington
Research Software Development Group, Research IT Services, University College London, UK

David Beavan
UCL Centre for Digital Humanities, University College London, UK

Martin Zaltz Austwick
Centre for Advanced Spatial Analysis, University College London, UK

Anne Welsh
Department of Information Studies, University College London, UK

Helen O'Neill
Department of Information Studies, University College London, UK, The London Library, UK

Will Finley
Department of History, University of Sheffield, UK

Oliver Duke-Williams
Department of Information Studies, University College London, UK

Adam Farquhar
Digital Scholarship, British Library, UK

Digital Scholarship in the Humanities © The Author 2017. Published by Oxford University Press on behalf of EADH.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.
doi:10.1093/llc/fqx020

Abstract

Although there has been a drive in the cultural heritage sector to provide large-scale, open data sets for researchers, we have not seen a commensurate rise in humanities researchers undertaking complex analysis of these data sets for their own research purposes. This article reports on a pilot project at University College London, working in collaboration with the British Library, to scope out how best high-performance computing facilities can be used to facilitate the needs of researchers in the humanities. Using institutional data-processing frameworks routinely used to support scientific research, we assisted four humanities researchers in analysing 60,000 digitized books, and we present two resulting case studies here. This research allowed us to identify infrastructural and procedural barriers and make recommendations on resource allocation to best support non-computational researchers in undertaking 'big data' research. We recommend that research software engineer capacity can be most efficiently deployed in maintaining and supporting data sets, while librarians can provide an essential service in running initial, routine queries for humanities scholars. At present there are too many technical hurdles for most individuals in the humanities to consider analysing at scale these increasingly available open data sets, and by building on existing frameworks of support from research computing and library services, we can best support humanities scholars in developing methods and approaches to take advantage of these research opportunities.

Correspondence:
Melissa Terras, UCL Department of Information Studies, Foster Court, University College London, Gower Street, London, WC1E 6BT, UK.
E-mail:
m.terras@ucl.ac.uk

1 Introduction

How best can humanities researchers access and analyse large-scale digital data sets available from institutions in the cultural and heritage sector? What barriers remain in place for those from the humanities wishing to use high-performance computing (HPC) to provide insights into historical data sets, using 'big-data' analytical techniques? This article describes a pilot project that worked in collaboration with non-computationally trained humanities researchers to identify and overcome barriers to complex analysis of large-scale digital collections. It used institutional university frameworks that routinely support the processing of large-scale data sets for research purposes in the sciences. The project brought together humanities researchers, research software engineers (Hettrick, 2016), and information professionals from the British Library Digital Scholarship Department,[1] University College London (UCL) Centre for Digital Humanities,[2] UCL Centre for Advanced Spatial Analysis,[3] and UCL Research IT Services (UCL RITS)[4] to analyse an open-licenced, large-scale data set from the British Library. While useful research results were generated, undertaking this project clarified the technical and procedural barriers that exist when humanities researchers attempt to utilize computational research infrastructures in the pursuit of their own research questions.

2 Overview

The drive in the gallery, library, archive, and museum (GLAM) sector towards opening up collections data,[5] as well as the growth in data published by publicly funded research projects, means humanities researchers have a wealth of large-scale digital collections available to them (Lui, 2015; Terras, 2015). Many of these data sets are released under open licences that permit uninhibited use by anyone with an Internet collection and modest storage capacity. A few humanities researchers have exploited these resources, and their interpretations make claims that change our understanding of cultural phenomena (Smith *et al.*, 2013; Schmidt, 2014; Smith *et al.*, 2015; Huber, 2007; Leetaru, 2015). Nevertheless, there remain major barriers to the widespread uptake of these data sets, and related

computational approaches, by humanities researchers, which risks diminishing the relevance of the humanities in 'big data' analysis (Wynne, 2015). These barriers include:

- fragmentation of communities, resources, and tools;
- lack of interoperability;
- complexity and incompleteness of heterogeneous cultural heritage data sets (Terras, 2009); and
- lack of technical skills: 'mainstream researchers in the humanities and social sciences often don't know what the new possibilities are' (ibid.) and seldom have the technical experience to experiment (Hughes, 2009; Mahony and Pierazzo, 2012).

A common response to this lack of awareness and computational skills is to build Web-based interfaces to data[6] or federated services and infrastructures.[7] While these interfaces play a positive role in introducing humanities researchers to large-scale digital collections, they rarely fulfil the complex needs of humanities research which constantly questions received approaches and results, or allow researchers to tailor analysis without being limited by shared assumptions and methods (Wynne, 2013).

3 Method

We explored the challenges associated with deploying and working with large-scale digital collections suitable for humanities research, using a public domain digital collection provided by the British Library.[8] This circa 60,000-book data set covers fiction and non-fiction publications from the 17th, 18th, and 19th centuries, or—seen as data—224 GB of compressed ALTO XML that includes both content (captured using an Optical Character Recognition (OCR) process) and the location of that content on a page.[9] Using UCL's centrally funded computing facilities,[10] we worked from March–July 2015 with UCL RITS and a cohort of four humanities researchers (from doctoral candidates to mid-career scholars) to ask queries that could not be satisfied by search- and discovery-orientated graphical user interfaces. Working in collaboration, we turned their research questions into computational queries, explored ways in which the returned data could be visualized, and captured their thoughts on the process through semi-structured interviews.

4 Results

We successfully ran queries across the data set that tracked linguistic change, identified core phrases, plotted the placement of illustrations, and mapped locations mentioned within core texts. The semi-structured interviews conducted with non-computationally trained humanities researchers at various stages during the collaborative work supported four key findings. First, that breaking down a research question into a series of more defined computational queries was time-consuming and challenging. Secondly, that the iterative nature of this research methodology puts pressure on the time taken to execute queries, and that long processing times were frustrating. Thirdly, that full comprehension of the programming code was not necessary to process data and use their outputs in research, though understanding the inputs, outputs, and effects of parameters was required. Fourthly, that creating derived data sets of a size manageable by desktop PCs[11] opened up further investigation using established methods. Indeed, we found that building queries that generate derived data sets from large-scale digital collections (small enough to be worked on locally with familiar tools) is an effective means of empowering non-computationally trained humanities researchers to develop the skill sets required to undertake complex analysis of humanities data.[12]

Our case studies deepen and add nuance to these findings. Two of our case studies were interested in looking at instances of particular words or phrases in the corpus (for example, 'professor'), or particular combinations of phrases within the corpus ('higher education'), to identify a particular institution and group of persons across time. The requirements from the researchers were to return the complete page of text that surrounded each example. This was found to be technically quite

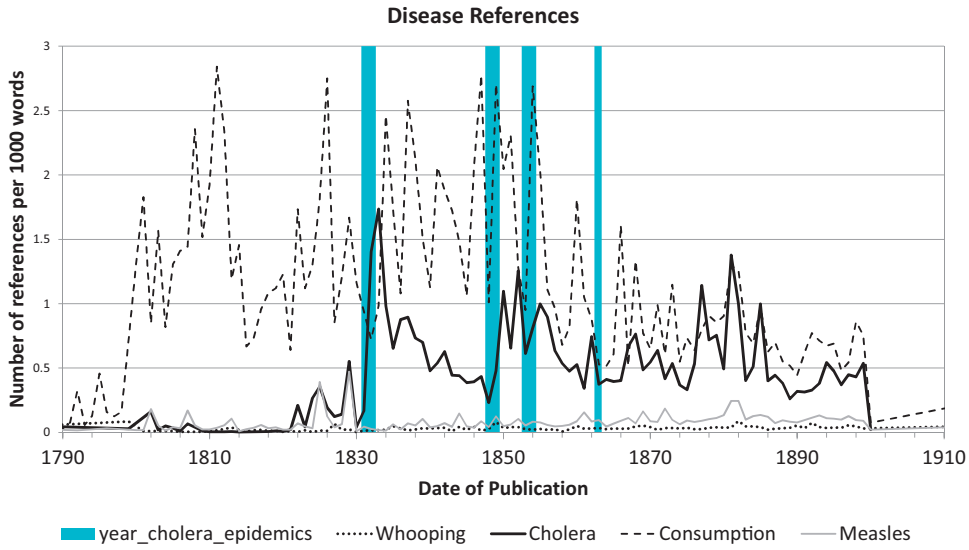

Fig. 1 A search for mentions of various infectious diseases (cholera, whooping cough, consumption, and measles) across the 60,000-book data set. We compared the profound spikes for cholera in the data set with known data regarding epidemics in the UK (Chadwick, 1842; Wall, 1893) which appear as the bars on the graph, showing a relationship between the first major UK outbreak of cholera and its appearance within the written record of the time (in 1831–32), and again with the second UK epidemic (1848–49). Later outbreaks (1853–54 and 1863) do not see this same correlation. There are further pronounced spikes for mentions of cholera in the1870s and 1880s: these are not associated with UK epidemics, but there were outbreaks in the USA and elsewhere. Identifying the texts that refer to these outbreaks allows us to look more closely at these clusters and to understand the relationship between public health, epidemiology, and the published historical record

straightforward, and resulted in a text file being delivered to the Humanities Scholars which they could then 'close read' to analyse each instance of the search term within a given page of the book in the corpus. Analysis in this case entails finding instances of the search term in question; however, there are further possibilities that can interrogate the data set further, in procedurally and methodologically novel ways. We present here two more ambitious case studies that allowed for further visualization and analysis.

4.1 Case Study 1: history of medicine

Duke-Williams is a senior lecturer in Digital Information Studies in the Department of Information Studies at UCL,[13] and his research interests include the presentation of spatial data and dissemination of demographic data, and the past, present, and future of demographic data capture in the UK. Visualization of these kinds of data can be used to explore issues around the spread of diseases, and the research questions were how does the occurrence of diseases in published literature compare to known epidemics in the 19th century? Can we see any correlation between the occurrence of infectious diseases in society and reference to these diseases in both fiction and nonfiction?

Variations in the number of mentions of cholera (Fig. 1, continuous black line) were compared to recorded epidemics (shaded bars on Fig. 1). A sharp rise in mentions coincides with the first cholera epidemic in the UK, of 1831–32; a similar but less pronounced rise is coincident with the 1848–49 epidemic. A more volatile pattern of mentions is observed after this point; subsequent spikes may be associated with epidemics within and beyond the UK, or may be less directly related to

disease incidence. Identifying the range and type of texts (whether epidemiological reports or works aimed at a wider audience) may help to inform and understand the cultural response to disease.

This work opens up possibilities for our understanding of trends in both fiction and non-fiction, and could be linked into further data sets (for example, of digitized historical newspaper data). In the case of our pilot project, it demonstrated that we could graph and visualize searches based on the corpus to present overviews that were useful to our researcher, but only in conjunction with both our research software engineer and our information visualization expert: this service then—as a result of the person hours required—does not scale in practice, demonstrating both the potential in the data set and the current limited opportunities historians, epidemiologists, and historians of science have to generate such visualizations from open-licenced data sets.

4.2 Case Study 2: the history of images

Finley is a doctoral candidate on the British Library and University of Sheffield Collaborative PhD Studentship 'The Printed Image 1750–1850: towards a Digital History of Printed Book Illustration'.[14] Between 1750 and 1850, changes in printing technology enabled several kinds of image to proliferate and for image and text to be brought together in novel and unexpected ways. Existing printing technologies—such as woodcuts—continued alongside new printing technologies, shaping the dissemination, reuse, and meaning of the designs they conveyed (Stijnman, 2012; Maidment, 2013). To understand these changes, scholars have so far sampled small, hand-crafted collections of images, an approach repeated in the fields of art and cultural history (Donald, 1996; Thomas, 2004). Yet digital sources allow us to study these changes with a much larger sample to use visual content as well as metadata to grapple with past phenomena at scale.

Finley's research focuses on the digital images from the same 60,000-book data set our project uses. The research addresses questions such as: How did changes in image techniques and the size of images map onto the different genres over time? What do quantitative findings reveal about the changing meanings of images from one genre to the next? How do the findings made possible using digital humanities techniques and digital sources compare to those using traditional methods and small, hand-crafted collections?

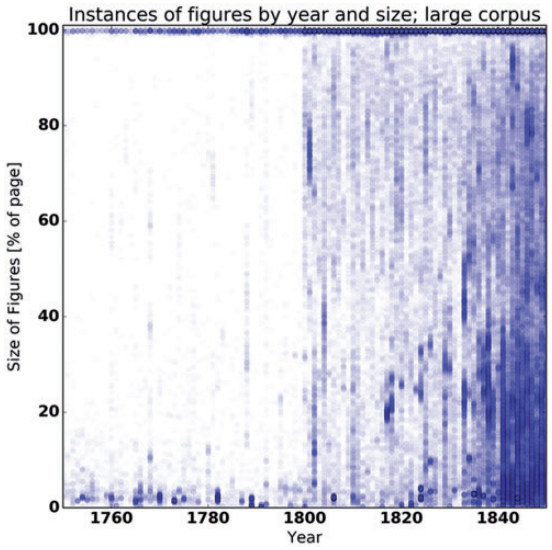

Fig. 2 A search for figures between 1750 and 1850, plotted according to the size of each figure in relation to the size of the page

To support these research questions, we queried the book XML to extract the coordinates of the boundary boxes put around each area the OCR process defined as an image. The resulting derived data lists the title, author, place of publication, and reference number for each book. For each of the 1 million images in these books, the derived data lists the page number it appears on, x-position of the top left corner, y-position of the top left corner, its width, its height, and its overall size as a percentage of the page. We then took two approaches to turn Finley's research questions into computational queries.

First, we used the data derived from the HPC to generate a graph of the instances of images by their size as a percentage of the page over time (Fig. 2). This enabled Finley to observe the dominance of full page and very small images (<15% of the page) between the 1750s and 1810s, after which time—driven by novel deployment of woodcuts and lithographs in books—the range of figure sizes

Fig. 3 A search for figures in *A new and complete System of Modern Geography*, two volumes (Mackenzie, 1817) plotted by page (*x*-axis), percentage of page the each figure occupies (*y*-axis), and separated by volume

diversified. Although the graph is not normalized by the number of images in the data set for each year, and is therefore dominated by the greater volume of books in the data set after 1800, it has proved a useful reference and a new way into the macro patterns of book illustration.

Secondly, we wrote a script that could be run locally in R to create graphs based on image data for single books. Plotting the page number on the *x*-axis and figures as a percentage of the page on the *y*-axis, the script generated a visual representation of the size of illustrations in a book. Finley selected books for analysis to observe patterns in the use of illustrations in books on history, geology, and topography (the subjects of his doctoral research). Here (Fig. 3) we see this for the 1817 *A new and complete System of Modern Geography*, a two-volume work published in Newcastle upon Tyne by Mackenzie (1817). The discrepancies in use of illustrations between the two volumes took Finley back to the physical books to assess how the placement and size of images changed the reading experience between volumes and to compare the findings with similar multi-volume works.

Subsequent to the project, Finley has continued to use the project data and scripts in his research. For example, he has used the image location data to plot and compare the average position images in books. This has underscored the value of generating derived data that can be used locally by a researcher outside the context of a HPC facility and a funded project.

5 Infrastructural recommendations

From a technical perspective, this pilot highlighted various sticking points when using infrastructure developed predominantly for scientific research. The combined data input and output volume undertaken during our work (less than 300 GB) is only moderately large by comparison to the scientific data sets UCL RITS usually encounters, for although there are shared assumptions between research infrastructures (adoption of technical standards, and the sharing of tools, approaches and research outputs (Wynne, 2015)), most of the UK's university eScience[15] infrastructure has been constructed specifically to run scientific and engineering simulations, not for search and analysis of the types of heterogeneous data sets we see emanating from cultural heritage institutions. We had a large textual input (224 GB), a simple calculation, and a small output summary of only a few KB. By comparison, the typical engineering simulation addresses moderately sized numerical input data, runs a long, complicated calculation, and produces a large output (multiple TBs). The average data size of project using the UCL data storage service is 4.4 TB (Hetherington, 2017). For example, the work of the UCL Centre for Computational Science[16] on brain blood flow simulations takes an input file of around 1 GB and, for a full production simulation recording a snapshot once every 200 time steps, produces 20 TB of output (Groen *et al.*, 2013). Poor uptake in the Arts and Humanities (Atkins *et al.*, 2010; Voss *et al.*, 2010) has meant that these computational systems have not been optimized for Arts and Humanities workloads. The file system and network configuration of

Legion—UCL RITS's centrally funded resource for running complex and large computational scientific queries across a large number of cores—did not match the way that the data set in question was structured (a large number of small zipped XML files).

The complexities associated with redeploying architectures designed to work with scientific data (massive yet very structured) to the processing of humanities data (not massive but more unstructured) should not be understated, and are a major finding of this project. Relevant libraries (such as an efficient XML processor) were needed to be installed and optimized for the hardware. Also, the data needed to be transformed to a structure that the parallel file system (Lustre) could address efficiently (that is fewer, larger files). We found that the architecture at UCL, which was configured for effective compute of scientific data, was input/output limited for our processing requirements, rather than computationally limited. Understanding the needs of our user community has already fed into the procurement and development of HPC facilities at UCL to ensure that the systems—which are available to all researchers—can deal with the variety and type of data that digital humanists wish to analyse, in future.

Best practice recommendations for similar projects emerged from this work: the need to build multiple derived data sets (counts of books and words per year, words and pages per book, etc.) to normalize results and maintain statistical validity; the necessity of documenting decisions taken when processing data and metadata; and the value of having fixed, definable data for researchers to explain results in relation to (and in turn, the risks associated with iterating data sets). We also discovered that a core set of four or five queries gave most of the humanities researchers the type of information they required to take a subset of data away to process effectively themselves: searches for all variants of a word, searches that return keywords in context traced over time, NOT searches for a word or phrase that ignored another word or phrase, searches for a word when in close proximity to a second word, and searches based on image metadata. It is the subset of the data set that most humanities scholars required, and were happy to be presented with for further analysis (with most researchers wishing to see their search term in context, presented with the complete page of the text it was found within to allow informed understanding).

A main finding of this pilot was, given most humanities researchers have a research problem that can be facilitated by a standard set of queries across large-scale textual data, that it would be more efficient to train a focussed group of service providers to be able to generate the results needed by researchers, than providing widespread training of humanities academics in this area. Higher Education already employs librarians to assist in searching and training for searching (information literacy), and providing this professional group with adjustable 'recipes' for defined computational queries and background training on their use would situate access to infrastructure in the resource to which humanists already turn for assistance—their subject librarian—and thereby normalize such computational work within the general humanities workflow.[17] In turn, research software engineers could be invoked as collaborators for their expertise, such as for developing more complex searches beyond the basic recipes, rather than having to repeat the defined searches across data for different researchers which would allow limited resources to be used efficiently, and to build on existing frameworks of support from both the library and computing services.

Given issues in resourcing such facilities at every University, it may be more efficient for multiple Higher Education Institutions to support a specialist service, perhaps under the umbrella of the likes of Jisc Historical Texts (http://historicaltexts.jisc.ac.uk/) or national or legal deposit libraries. Expertise and approaches, if not the service itself, could also be facilitated through the likes of Digital Research Infrastructure for the Arts and Humanities (DARIAH) (http://www.dariah.eu/) and Common Language Resources and Technology Infrastructure (CLARIN) (https://www.clarin.eu/). However, local support for researchers wishing to utilize existing eScience technologies within the Higher Education sector should be possible. Such support enables Arts and Humanities researchers to develop ongoing, mutually beneficial relationships with research

computing within their own institution. This, in turn, can encourage other researchers to use these resources (rather than them being only available as a specialist service which users have to seek out). Research computing infrastructure across the university sector will not meet the needs of Arts, Humanities, and Social Sciences researchers unless academics in these fields becomes active users of the systems, and their requirements can be taken into account, going forward.

6 Conclusion

We successfully mounted large-scale humanities data on HPC University infrastructure in an interdisciplinary project that required input from many professionals to aid the humanities scholars in their research tasks. The collaborative approach we undertook in this project is labour-intensive and does not scale. This should not, however, discourage the sector from taking this work forward. We found that many research questions can be expressed with similar computational queries, albeit with parameters adjusted to suit. We recommend, therefore, that Higher Education Institutions or HEI clusters looking to build capacity for enabling complex analysis of large-scale digital collections by their non-computationally trained humanities researchers should consider the following activities:

(1) Invest in research software engineer capacity to deploy and maintain openly licensed large-scale digital collections from across the GLAM sector to facilitate research in the arts, humanities and social and historical sciences.
(2) Invest in training library staff to run these initial queries in collaboration with humanities faculty, to support work with subsets of data that are produced, and to document and manage resulting code and derived data.

Our pilot project demonstrates that there are at present too many technical hurdles for most individuals in the arts and humanities to consider analysing large-scale open data sets. Those hurdles can be removed with initial help in ingest and deployment of the data, and the provision of specific, structured, training and support which will allow humanities researchers to get to a subset of useful data they can comfortably and more simply process themselves, without the need for extensive support. While we, together with our partners, have plans to continue expanding the range and depth of research carried out on our chosen data set, this project has signposted many of the barriers to encourage greater uptake of 'big data' research across the Arts and Humanities. These findings should be of use to researchers wishing to use comparable approaches, and to service providers in research computing aiming to encourage the use of shared computational facilities by the Arts and Humanities community.

Acknowledgements

This project was funded by a pilot grant from the Jisc Research Data Spring (JISC 3548) between April and July 2015. See https://www.jisc.ac.uk/rd/projects/research-data-spring for further details. The authors thank our colleagues in the British Library and UCL RITS (particularly Clare Gryce) for their support, Geoffrey Rockwell and Stéfan Sinclair for discussions on the current limitations of text analysis for individual humanities scholars, and Scott B. Weingart for his assistance.

References

Atkins, D. E., Borgman, C. L., Bindhoff, N., Ellisman, M., Felman, S., Foster, I., and Heck, A. (2010). "*RCUK Review of e-Science 2009*." Research Councils UK. https://www.epsrc.ac.uk/newsevents/pubs/rcuk-review-of-e-science-2009-building-a-uk-foundation-for-the-transformative-enhancement-of-research-and-innovation/.

Bates, M. J. (1996). The Getty end-user online searching project in the humanities: report no. 6: overview and conclusions. *College and Research Libraries*, **57**(6): 514–23.

Bedi, S. and Walde, C. (2016). Transforming roles: Canadian academic librarians embedded in faculty research projects. *College and Research Libraries*. http://

crl.acrl.org/content/early/2016/03/22/crl16-871.abstract.

Bradigan, P. S. and Mularski, C. A. (1989). End-user searching in a medical school curriculum: an evaluated modular approach. *Bulletin of the Medical Library Association*, **77**(4): 348–56.

Brettle, A., Maden-Jenkins, M., Anderson, L., McNally, R., Pratchett, T., Tancock, J., Thornton, D., and Webb, A. (2011). Evaluating clinical librarian services: A systematic review. *Health Information and Libraries Journal*, **28**(1): 3–22.

Burke, J. and Tumbleson, B. (2016). LMS embedded librarianship and the educational role of librarians. *Library Technology Reports*, **52**(2): 5–9.

Chadwick, E. (1842). Report on the sanitary condition of the labouring population of Great Britain: supplementary report on the results of special inquiry into the practice of interment in towns (Vol. 1). HM Stationery Office.

Donald, D. (1996). *The Age of Caricature: Satirical Prints in the Reign of George III*. New Haven: Published for the Paul Mellon Centre for Studies in British Art by Yale University Press.

Farber, M. and Shoham, S. (2002). Users, end-users, and end-user searchers of online information: a historical overview. *Online Information Review*, **26**(2): 92–100.

Feliu, V. and Frazer, H. (2012). Embedded librarians: teaching legal research as a lawyering skill. *Journal of Legal Education*, **61**(4): 540–59.

Groen, D., Hetherington, J., Carver, H. B., Nash, R. W., Bernabeu, M. O., and Coveney, P. V. (2013). Analysing and modelling the performance of the HemeLB lattice-Boltzmann simulation environment. *Journal of Computational Science*, **4**(5): 412–22.

Hetherington, J. (2017). Question about data capacity and use at UCL. *Response to Melissa Terras via email*, 27 January 2017.

Hettrick, S. (2016). *A not-so-brief history of Research Software Engineers*. Software Sustainability Institute. https://www.software.ac.uk/blog/2016-08-19-not-so-brief-history-research-software-engineers.

Huber, M. (2007). The Old Bailey Proceedings, 1674-1834 Evaluating and annotating a corpus of 18th - and 19th-century spoken English. *Studies in Variation, Contacts and Change in English 1: Annotating Variation and Change*. http://www.helsinki.fi/varieng/series/volumes/01/huber/.

Hughes, A. (2009). *Higher Education in a Web 2.0 World*. JISC. http://www.webarchive.org.uk/wayback/archive/20140614042502/http://www.jisc.ac.uk/publications/generalpublications/2009/heweb2.aspx.

Ikeda, N. R. and Schwartz, D. G. (1992). Impact of end-user training on Pharmacy students: a four-year follow-up study. *Bulletin of the Medical Library Association*, **80**(2): 124–30.

Lancaster, F. W., Elzy, C., Zeter, M. J., Metzler, L., and Low, Y. M. (1994). Searching databases on CD-ROM: comparison of the results of end-user searching with results from two modes of searching by skilled intermediaries. *RQ*, **33**(3): 370–86.

Leetaru, K. (2015). *History as Big Data: 500 Years of Book Images and Mapping Millions of Books*. Forbes, Tech. http://www.forbes.com/sites/kalevleetaru/2015/09/16/history-as-big-data-500-years-of-book-images-and-mapping-millions-of-books/.

Ludwig, L., Mixter, J. K., and Emanuele, M. A. (1988). User attitudes toward end-user literature searching. *Bulletin of the Medical Library Association*, **76**(1): 7–13.

Lui, A. (2015). *Data Collections and Datasets*. Curated by Alan Liu. http://dhresourcesforprojectbuilding.pbworks.com/w/page/69244469/Data%20Collections%20and%20Datasets.

Mackenzie, E. (1817). A new and complete system of modern geography: containing an accurate delineation of the world. 2 Vols, Printed and published by Mackenzie and Dent, Newcastle-upon-Tyne.

Mahony, S. and Pierazzo, E. (2012). Teaching skills or teaching methodology? In Hirsch, B. D. (ed.), *Digital Humanities Pedagogy: Practices, Principles and Politics*. Open Book Publishers. http://www.openbookpublishers.com/product/161/digital-humanities-pedagogy–practices–principles-and-politics.

Maidment, B. (2013) *Comedy, Caricature and the Social Order 1820-1850*. Manchester: Manchester University Press.

Markey, K. (2007a). Twenty-five years of end-user searching. Part 1: research findings. *Journal of the American Society for Information Science and Technology*, **58**(8): 1071–81.

Markey, K. (2007b). Twenty-five years of end-user searching. Part 2: future research directions. *Journal of the American Society for Information Science and Technology*, **58**(8): 1123–30.

Murray, T. (2016).The forecast for special libraries. *Journal of Library Administration*, **56**(2): 188–98.

Rethlefsen, M., Farrell, A., Osterhaus Trzasko, L., and Brigham, T. (2015). Librarian co-authors correlated with higher quality reported search strategies in general internal medicine systematic reviews. *Journal of Clinical Epidemiology*, **68**(6): 617–26.

Rockwell, G. (2017). Question about maximum capacity of Voyant tools. *Response to Melissa Terras via email*, 27 January 2017.

Rockwell, G. and Sinclair, S. 2016. *Hermeneutica: Computer-Assisted Interpretation in the Humanities*. Cambridge: MIT Press.

Roper, T. (2015). The impact of the clinical librarian: a review. *Journal of EAHIL*, **11**(4): 19–22.

Schmidt, B. (2014). Shipping maps and how states see. *Sapping Attention Blog*. http://sappingattention.blogspot.co.uk/2014/03/shipping-maps-and-how-states-see.html.

Sinclair, S. (2017). Question about maximum capacity of Voyant tools. *Response to Melissa Terras via email*, 27 January 2017.

Smith, D. A., Cordell, R., and Dillon, E. M. (2013). Infectious texts: modeling text reuse in nineteenth-century newspapers. In *IEEE International Conference on Big Data, October 6-9, 2013*. Santa Clara, California: IEEE, pp. 86–94. 10.1109/BigData.2013.6691675.

Smith, D., Cordell, R., and Mullen, A. (2015). Computational methods for uncovering reprinted texts in Antebellum newspapers. *American Literary History*, **27**(3).

Starr, S. S. and Renford, B. L. (1987). Evaluation of a program to teach health professionals to search MEDLINE. *Bulletin of the Medical Library Association*, **75**(3): 193–201.

Stijnman, A. (2012). *Engraving and Etching, 1400-2000: A History of the Development of Manual Intaglio Printmaking Processes*. London; Houten, Netherlands: Archetype Publications; in association with HES and DE GRAAF Publishers.

Straus, S., Eisinga, A., and Sackett, D. (2016). What drove the evidence cart? Bringing the library to the bedside. *Journal of the Royal Society of Medicine*, **109**(6): 241–7.

Streipe, T. and Talley, M. (2013). Embedded librarianship. In Kroski, E. (ed.), *Law Librarianship in the Digital Age*. Plymouth: Scarecrow, pp. 13–30.

Terras, M. (2009). *Potentials and Problems in Applying High Performance Computing for Research in the Arts and Humanities: Researching e-Science Analysis of Census Holdings*. Digital Humanities Quarterly. http://www.digitalhumanities.org/dhq/vol/3/4/000070/000070.html.

Terras, M. (2015). Opening access to collections: the making and using of open digitised cultural content. *Online Information Review*, **39**(5): 733–52. http://www.emeraldinsight.com/doi/full/10.1108/OIR-06-2015-0193.

Thomas, J. (2004). *Pictorial Victorians: The Inscription of Values in Word and Image*. Athens: Ohio University Press.

Voss, A., Asgari-Targhi, M., Procter, R., and Fergusson, D. (2010). Adoption of e-Infrastructure services: configurations of practice. *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences*, **368:** 4161–76; doi: 10.1098/rsta.2010.0162.

Wall, A. J. (1893). *Asiatic cholera: its history, pathology, and modern treatment*. London: H K Lewis.

Wittek, S., Sinclair, S., and Milner, M. (2015). *DREaM: Distant Reading Early Modernity*. Digital Humanities 2015, Global Digital Humanities, Sydney, Australia, 29 June–3 July 2015. http://dh2015.org/abstracts/xml/WITTEK_Stephen_DREaM__Distant_Reading_Early_Moder/WITTEK_Stephen_DREaM__Distant_Reading_Early_Modernity.html.

Wynne, M. (2013). The role of Clarin in digital transformations in the humanities. *International Journal of Humanities and Arts Computing*, **7**(1-2): 89.

Wynne, M. (2015). Big Data and Digital Transformations in the Humanities: are we there yet? Textual Digital Humanities and Social Sciences Data, Aberdeen, 21–22 September 2015. http://www.slideshare.net/martinwynne/big-data-and-digital-transformations-in-the-humanities.

Notes

1 https://www.bl.uk/subjects/digital-scholarship.
2 http://www.ucl.ac.uk/dh.
3 http://www.bartlett.ucl.ac.uk/casa.
4 http://www.ucl.ac.uk/isd/services/research-it.
5 http://openglam.org/, an initiative to promote free and open access to digital cultural heritage data sets.
6 For example, Mining the History of Medicine (http://nactem.ac.uk/hom/) or Language of the State of the Union (http://www.theatlantic.com/politics/archive/2015/01/the-language-of-the-state-of-the-union/384575/).
7 For example, CLARIN (http://clarin.eu/) and DARIAH (https://www.dariah.eu/).

8 The British Library has various digital data sets, including (but not limited to) 7 million pages of historic newspapers, 1 million out of copyright book illustrations, 100,000s of scientific articles, text from over 60,000 books, 1,000s of UK theses, and various digitized medieval manuscripts. We chose here just one of its large-scale data sets to work with in this pilot phase. For the terms under which the British Library makes collections available, see http://www.bl.uk/aboutus/terms/copyright/.

9 The books cover a wide range of subject areas including philosophy, history, poetry, and literature; most are in English. For a full list of metadata for this book collection, see doi: 10.21250/DB21.

10 https://wiki.rc.ucl.ac.uk/wiki/Legion, just one of the HPC facilities available at UCL for researchers, see http://www.ucl.ac.uk/isd/services/research-it/research-computing.

11 It is difficult to put a figure on what a manageable text file size would be for a researcher, if we take that as being able to search through it on their own machine without further technical assistance. This is dependent on access to processing power, which varies considerably between desktop and laptop computers, and the complexity of the search the researcher is intending to carry out, as well as the format and granularity of the source text. Most humanities researchers should now be able to comfortably manipulate text files of around 100 MB if they have access to a modern machine: the amount of text those files will contain is dependent on how complex the file formats are. Programs and tools available to assist in text analysis include Voyant Tools (https://voyant-tools.org/) and R (https://www.r-project.org/). The upper limit to using server-based Voyant is 20 MB of text, which should correlate to around 1 million words depending on format (Sinclair, 2017). Voyant Server can be downloaded and used locally, with an upper limit of around 100 MB on a standard PC/Mac (Rockwell, 2017). Indexing engines such as Lucene (http://lucene.apache.org/) can search larger amounts of text; however, visualization of results will require further processing and can be complex. For further exploration of 'computer-assisted interpretation in the humanities' (particularly using Voyant), see Rockwell and Sinclair (2016), and an example of the approach of generating manageable smaller corpora from a large-scale Early Modern Sources can be found in Wittek *et al.* (2015).

12 All code, data, visualizations and other outputs from this pilot project are freely available at https://github.com/UCL-dataspring.

13 https://www.ucl.ac.uk/dis/people/oliverdukewilliams.

14 https://www.shef.ac.uk/history/research/students/william-finley.

15 For more on the UK's eScience infrastructure, see the work of the eScience Institute, http://www.esi.ac.uk/. Plan-Europe is the Platform of National eScience Centres in Europe (http://plan-europe.eu/). In the USA, the equivalent of eScience is known as Cyberinfrastructure; see the National Science Foundation's guides: http://www.nsf.gov/div/index.jsp?div=ACI.

16 http://ccs.chem.ucl.ac.uk/.

17 This approach is similar to that employed in the 1980s when computerized searching of specialist databases was first available (Farber and Shoham, 2002; Markey, 2007a, 2007b). At that time, evaluations carried out on programs to cascade train end users via librarians reported increased user uptake and efficiency as compared to direct end-user training by database providers (see Starr and Renford, 1987; Bradigan and Mularski, 1989; Ikeda and Schwartz, 1992 on the library-based training of medical staff to use MEDLINE). Significantly, studies showed that while end users wanted to be able to search databases themselves, they also desired access to searches mediated by librarians and/or a team approach to search utilizing the end user's specific knowledge of the language and jargon of their discipline and the information professional's understanding of BOOLEAN and other advanced search techniques (Ludwig *et al.*, 1988; Lancaster *et al.*, 1994; Bates, 1996). The move towards embedded librarians in law practices (Feliu and Frazer 2012; Streipe and Talley, 2013; Murray, 2016) and clinical librarians in hospitals (Brettle *et al.*, 2011; Roper, 2015; Straus *et al.*, 2016) can be seen to be a modern iteration of a team approach to search. In Higher Education, subject librarians are well-positioned to form part of a research team using computational methods and/or provide information literacy training that upskills humanists in computational methods (Rethlefsen *et al.*, 2015; Bedi and Walde, 2016; Burke and Tumbleson, 2016).