Digital Humanities 2016
Kraków, 11–16 July

(http://dh2016.adho.org)

DH Home (http://www.dh2016.adho.org) / Abstracts (/abstracts/) / 230 (/abstracts/230)

Show info How to cite XML Version (/static/data/41.xml)

Terras, M., Baker, J., Hetherington, J., Beavan, D., Welsh, A., O'Neill, H., Finley, W., Duke-Williams, O., Farquhar, A. (2016). Enabling Complex Analysis of Large-Scale Digital Collections: Humanities Research, High Performance Computing, and transforming access to British Library Digital Collections. In *Digital Humanities 2016: Conference Abstracts*. Jagiellonian University & Pedagogical University, Kraków, pp. 376-379.

Enabling Complex Analysis of Large-Scale Digital Collections: Humanities Research, High Performance Computing, and transforming access to British Library Digital Collections

How best can humanities researchers access and analyse large-scale digital datasets 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 to provide insights into historical datasets? This paper 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 using 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, and information professionals from the British Library Digital Scholarship Department [1], UCL Centre for Digital Humanities (UCLDH) [2], UCL Centre for Advanced Spatial Analysis (UCL CASA) [3], and UCL Research IT Services (UCL RITS) [4] to analyse an open-licensed, large-scale dataset 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.

1. 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 datasets 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 (for example, see Schmidt, 2014; Smith et al., 2015; Cordell et al., 2013; 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;
- 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]. Whilst 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).

2. 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 60,000 book dataset covers publication from the 17th, 18th, and 19th centuries, or – seen as data – 224GB of compressed ALTO XML that includes both content (captured using an OCR process) and the location of that content on a page. Using UCL's centrally funded computing facilities [9] we worked from March-July 2015 with 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 visualised, and captured their thoughts on the process through semi-structured interviews.

3. Results

We successfully ran queries across the dataset tracking linguistic change, identifying core phrases, plotting and understanding the placing of illustrations, and mapping locations mentioned within core texts. We found that building queries that generate derived datasets 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. [10]

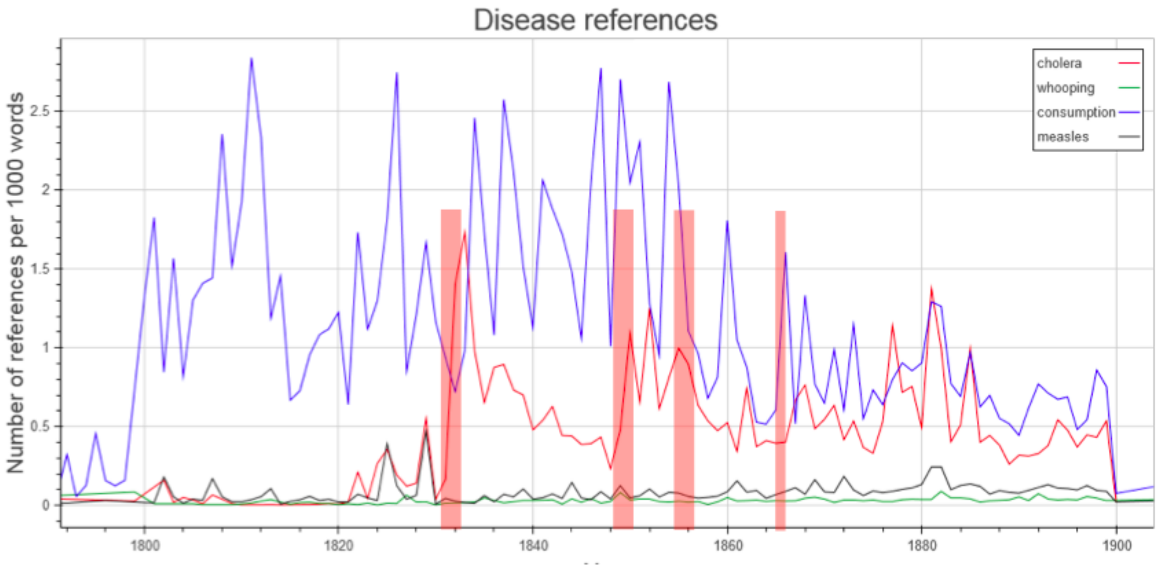

Figure 1: A search for mentions of various infectious diseases (Cholera, Whooping Cough, Consumption, and Measles) across the 60,000 book dataset. We compared the profound spikes for Cholera in the dataset with known data regarding epidemics in the UK (Chadwick, 1842; Wall, 1893) which appear as the 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. There are further pronounced spikes 1870s and 1880s: these are not associated with UK epidemics, but there were outbreaks in the US 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.

From a technical perspective, this pilot highlighted various sticking points when using infrastructure developed predominantly for scientific research. 224GB is only moderately large by comparison to the scientific datasets UCL RITS usually encounters, but 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 [11] infrastructure has been constructed specifically to run scientific and engineering simulations, not for search and analysis of heterogeneous datasets. Our task here had a large textual input, a simple calculation, and a small output summary. By comparison, the typical engineering simulation addresses moderately sized numerical input data, runs a long, complicated calculation, and produces a large output. Poor uptake in the arts and humanities (Atkins et al., 2010; Voss, 2010) has meant that these resources have not been optimised for these 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 dataset 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) needed to be installed and optimised 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).

Best practice recommendations for comparable projects emerged from this work: the need to build multiple derived datasets (counts of books and words per year, words and pages per book, etc) to normalise 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 datasets). Pointers to how to process the derived datasets were welcomed, but it was at this stage that the researchers were confident to "go it alone" without our support. 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: for example, keywords in context traced over time; NOT searches for a word or phrase that ignored another word or phrase, etc. As Higher Education Institution (HEI)-based subject librarians regularly handle routine research queries, we contend that training librarians to aid humanities researchers in carrying out defined computational queries via adjustable recipes would improve access to infrastructure, and cut down on the human-resource intensive nature of this approach. In turn, research computing programmers could be invoked as collaborators for their expertise, such as for developing more complex searches beyond the basic recipes.

4. Conclusion

We successfully mounted large-scale humanities data on high performance computing 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. Nevertheless, we found many research questions can be expressed with similar computational queries, albeit with parameters adjusted to suit. We recommend, therefore, that HEIs or HEI clusters looking to build capacity for enabling complex analysis of large-scale digital collections by their non-computationally trained humanities research should consider the following activities:

- Invest in research software engineer capacity to deploy and maintain openly licensed large-scale digital collections from across the GLAM sector in order to facilitate research in the arts, humanities and social and historical sciences,
- 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.

Bibliography

1. **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/
2. **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/
3. **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.
4. **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/.
5. **Lui, A.** (2015). Data Collections and Datasets, Curated by Alan Liu. http://dhresourcesforprojectbuilding.pbworks.com/w/page/69244469/Data%20Collections%20and%20Datasets.
6. **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.
7. **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.
8. **Smith, D., Cordell, R. and Mullen, A.** (2015). Computational Methods for Uncovering Reprinted Texts in Antebellum Newspapers. *American Literary History*, **27**(3).

9. **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

10. **Voss, A., Asgari-Targhi, M., Procter, R. and Fergusson, D.** (2010). Adoption of e-Infrastructure services: configurations of practice. *Philosophical Transactions of the Royal Society A*. DOI: 10.1098/rsta.2010.0162.

11. **Wynne, M.** (2013). The Role of Clarin in Digital Transformations in the Humanities, *International Journal of Humanities and Arts Computing* **7**(1-2): 89-2014.

12. **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] http://britishlibrary.typepad.co.uk/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 datasets.

[6] For example, Mining the History of Medicine (http://nactem.ac.uk/hom/) (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/), Common Language Resources and Technology Infrastructure, and DARIAH (https://www.dariah.eu/) Digital Research Infrastructure for the Arts and Humanities.

[8] The British Library has various digital datasets, including (but not limited to) 7m pages of historic newspapers, 1m out of copyright book illustrations, 100,000s of scientific articles, text from over 60,000 books, 1000s of UK theses, and various digitized medieval manuscripts. We chose here just one of its large scale datasets 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] https://wiki.rc.ucl.ac.uk/wiki/Legion, just one of the High Performance Computing facilities available at UCL for researchers, see http://www.ucl.ac.uk/isd/services/research-it/research-computing.

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

[11] 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 Centers in Europe (http://plan-europe.eu/). In the United States, the equivalent of eScience is known as Cyberinfrastructure, see the National Science Foundation's guides: http://www.nsf.gov/div/index.jsp?div=ACI.