Reflections on Infrastructures for Mining Nineteenth-Century Newspaper Data

Julianne Nyhan
Associate Professor of Digital Information Studies
University College London (UK)
j.nyhan@ucl.ac.uk

Tessa Hauswedell
Research Associate
University College London (UK)
t.hauswedell@ucl.ac.uk

Ulrich Tiedau
Associate Professor in Dutch
University College London (UK)
u.tiedau@ucl.ac.uk

Abstract: *In this study we compare and contrast our experiences (as historians and as digital humanities and information studies researchers) of seeking to mine large-scale historical datasets via university-based, high-performance computing infrastructures versus our experiences of using external, cloud-hosted platforms and tools to mine the same data. In particular, we reflect on our recent experiences in two large transnational digital humanities projects:* Asymmetrical Encounters: E-Humanity Approaches to Reference Cultures in Europe, 1815–1992, *which was funded by a Humanities in the European Research Area grant (2013–2016) and* Oceanic Exchanges: Tracing Global Information Networks in Historical Newspaper Repositories 1840–1914, *which was funded through the Transatlantic Partnership for Social Sciences and Humanities 2016 Digging into Data Challenge (2017–2019). As part of the research for both these projects we sought to mine the OCR text of nineteenth-century historical newspapers that had been mounted on UCL's High-Performance Computing Infrastructures from Gale's TDM drives. We compare and contrast our experiences of this with our subsequent experiences of performing comparable tasks via Gale Digital Scholar Lab. We contextualise our experiences and observations within wider discourses and recommendations about infrastructural support for humanities-led analyses of large datasets and discuss the advantages and drawbacks of both approaches. We situate our discussions in the aforementioned infrastructural scenarios with reflections on the human experiences of undertaking this research, which represents a step change for many of those who work in the (digital) humanities. Finally, we conclude by discussing the public and private sector research investments that are needed to support further developments and to facilitate access to and critical interrogation of large-scale digital archives.*

Keywords: digital infrastructures ■ text mining ■ historical newspaper collections ■ high-performance computing ■ critical cultural heritage ■ digital humanities ■ *The Times Digital Archive* ■ the *Times* of London

INTRODUCTION

In this talk we will compare and contrast our experiences as historians, digital humanists, and information studies researchers of seeking to mine large-scale, digitised historical newspaper collections via university-based high-performance computing infrastructures versus our experiences of using external, cloud-hosted tools to mine that same data.[1]

We situate these reflections in the context of two large transnational projects in which we have participated. *Asymmetrical Encounters: E-Humanity Approaches to Reference Cultures in Europe, 1815–1992*, was funded by a Humanities in the European Research Area grant (2013–2016) and included partners from the United Kingdom, Netherlands, and Germany. It was coordinated by the University of Utrecht. With digital technologies it carried out a longitudinal analysis of large digital newspaper and magazine archives and asked how the cultural aspects of European identity changed between 1815 and 1992 (AsymEnc, n.d.). We also draw on the experiences we have gained so far in the project *Oceanic Exchanges: Tracing Global Information Networks in Historical Newspaper Repositories, 1840–1914* (OcEx), funded by the Transatlantic Platform of the Digging into Data programme (2017–2019). The project brings together researchers in Finland, Germany, Mexico, Netherlands, the United Kingdom, and the United States, and is coordinated by Northeastern University. It "examine[s] patterns of information flow across national and linguistic boundaries and to link research across large-scale digital newspaper collections" (Oceanic Exchanges, n.d.).

In this talk, and in line with literature such as that of Susan Leigh Star and Karen Ruhleder (1996) and American Council of Learned Societies (2006), we define infrastructure as comprising, and being shaped by, a complex set of interacting dynamics that enfold not only physical structures but also social, cultural, and institutional processes and contexts. Indeed, Wolfgang Kaltenbrunner has argued that:

> Infrastructure occurs when the tensions between globally valid standards and local contexts, as well as between automated technological processes and tasks performed by human actors can be successfully resolved. An important consequence of this definition of infrastructure is that it develops incrementally—it is not created, it evolves. (Kaltenbrunner 2015, 211)

Other definitions of digital research infrastructure use the analogy of a "digital ecosystem," which provides "services that are built around communities" (Blanke, Kristel, Romary 2015). In our case, the community needs we are concerned with are those of the humanities and specifically those of the field of history and heritage. Accordingly, in this paper we do not limit our observations to the computational infrastructures that we used to undertake our research but we also reflect on the institutional, technical, legal, sociocultural, and labour organisation factors that sometimes aided, and sometimes impeded our work but which are relevant to consider in the context of humanities research. After setting these out, we close with some reflections on evolving infrastructures for undertaking large-scale data mining of cultural heritage sources and artefacts and the future directions they may take.

We begin by describing our experiences of seeking to computationally analyse *The Times Digital Archive*, an online, full-text facsimile of more than 200 years of the [London] *Times*, "one of the most highly regarded resources for eighteenth-, nineteenth-, and twentieth-century news coverage, with every page of every issue from 1785 to 2010" (Gale, n.d.).

ACQUIRING MINING ACCESS TO *THE TIMES DIGITAL ARCHIVE*

Following on from the newspaper digitisation programmes by national libraries and commercial companies that began in the late 1990s (see, e.g., Gooding 2014 for a history), we have seen an exponential increase in the creation and global availability of digital newspaper archives. Large quantities of historical newspapers have been digitised through the public hand (e.g., the Library of Congress), commercial companies (e.g., Gale/Cengage), and public-private partnerships (e.g., the British Newspaper Archive involving the British Library). Despite the many gaps in coverage that exist (Hobbs 2013; Milligan 2013), and issues like variable OCR quality of transcriptions (Smith and Cordell 2018; Tanner et al 2009), the existing collections of national libraries, commercial companies, and, in some cases, public-private partnerships now offer an abundance of material for researchers and other communities to interrogate (see Milligan 2019). Through the interfaces to the materials that providers offer, researchers can run simple and advanced keyword searches of the material. Those who are interested in carrying out more complex queries or investigating the metadata that makes structural and other elements of those digital surrogates (and information about them) machine readable, also have various possibilities open to them. The APIs[2] and data dumps[3] that providers like Europeana make available allow researchers to take copies of open-licenced data and query or transform it according to their needs. For those who wish to work with licenced data and metadata, both Gale and ProQuest provide text and data mining hard drives to paying users (on the history of this in the context of Gale see Fyfe 2016).

Mining *The Times Digital Archive* has been an important aspect of the projects that we have engaged in, and for this it was necessary to get access to the underlying full-text data and metadata of the collection. While our university library had a subscription to *The Times Digital Archive* for years, the collection was accessible to us via the standard search form only. Through a chance encounter with a colleague from Gale/Cengage in 2014, we became aware of the possibility of acquiring a copy of the *Times* data on an external hard drive, a so-called Text and Data Mining Hard Drive or TDM Hard Drive. On it was approximately 2 TB of *The Times Digital Archive*, then covering the years 1785 through 2010 (meanwhile extended to 2013), represented as one XML file per issue plus one TIFF image per page. We paid a moderate three-digit sum for the hard drive and this was mostly for the work that Gale/Cengage did to reformat the data so we could perform text and data mining on it, and also for the delivery of the files to the University College London (UCL) library who made it available to us. As we understand it, the licence essentially positioned the hard drives as extensions to the subscription that UCL libraries already had with Gale/Cengage, and for the duration of those subscriptions only, which explains the reasonably moderate fee. Gale granted UCL a "royalty-free non-exclusive, non-transferrable [sic], non-sublicensable, worldwide right, subject to terms and conditions of the Agreement … to enable Authorised Users to access materials solely for purposes of performing Text and Data Mining activities for non-commercial research purposes" (Gale licence).

Before moving on to recall the steps we next had to take to get the data mounted, and to seek to mine it, we will pause to reflect on the advantages and disadvantages of obtaining the data in this way. Firstly, the advantages: concerns about the risks of monetising cultural heritage through digitisation projects that created gated content have been raised and convincingly made in recent years (e.g., Prescott 2014; Darnton 2010). The *New Renaissance Report of the Comité des Sages* has, for example, cautioned of the potential of creating a digital dark age should the balance between public, private, and public-private digitisation give way (Comité des Sages 2010). Yet, as also made clear in that report and as wide experience has shown, there is much to be gained from mutually beneficial public-private partnerships. Away from the institutional and collection-level benefits of this approach, as individual researchers who needed to secure nineteenth-century newspaper text and metadata to work with, we found the experience with Gale/Cengage to be efficient and straightforward. As university-based researchers our instinct had been to attempt to secure access to the data we needed through our university-based infrastructures and other public institutions. That private providers could facilitate the first level of access to the sources alerted us to the distributed nature of the infrastructures we require to do this research. That we were privileged enough to have the funding required to leverage this route must be noted too. That said, there were particular constraints to working in this way. Secondly, then, to the disadvantages: we sought the TDM drives in the context of a multinational project yet the licence that we signed was conceptualised along national lines and linked to the revenue stream of one university—extending the licences along international lines proved more problematic. As such, our working solution could not scale to the transnational space that the project occurred in.

MOUNTING AND QUERYING

In any case, we had managed to acquire the full text and associated metadata of *The Times Digital Archive* and we in UCL were ready to do something with it. Making the link to UCL Research IT Services (RITS) was also serendipitous and came about during the course of a Digital Excursion that was organised by the UCL Centre for Digital Humanities.[4] UCL RITS[5] was at that time quite keen to show value to the institution outside of the usual data rich disciplines, who routinely undertake computationally intensive research, with which it often works. Their stated aim was to "enable researchers beyond those in science and engineering" (UCL 2017) and at the time, they had just worked with a humanities project for the first time ever: "Enabling complex analysis of large-scale digital collections collaboration between the British Library and UCL," which had been funded as part of the JISC Research Data Spring (Terras 2015). The UCL Research Computing Platforms Service, one of the services of UCL Research IT Services, offers "[s]upport … [for] computationally intensive research at UCL through provision of specialist platforms for high performance and high throughput computing.[6] They make three computing clusters, (Legion, Myriad, and Grace) available to the wider UCL community, who do not pay for the time it takes to run scripts or for storage. Beyond those tasks the service charges about £350 per day to research projects to, for example, write and run bespoke text mining scripts. It was agreed that, with the help of our colleagues in Research IT Services, we could mount *The Times Digital Archive* data and run queries on it with Legion, "a mixed-use cluster hosted in UCL's Bloomsbury data centres."[7]

The first step was to upload the data to a protected account in the integrated Rule-Oriented Data System (iRODS) of UCL Research Data Services (see Wilson 2017), where we were given a 5 TB account. Almost from the outset, though, it became apparent that the system was not originally set up to support humanities data and research. Instead of the massive and highly structured data that the computationally driven disciplines often query (e.g., Stevens 2013), as already mentioned, our data comprised numerous small XML files and TIFF files, grouped into multiple directories. As a result, it took days for it to be uploaded. Once uploaded it was necessary for Research IT Services to rechunk the data so we could merge the XML files for individual newspaper issues into monthly, quarterly, or yearly bundles. We understand that this was necessary not only to support the longitudinal analysis that we hoped to perform on the data but also from a computational perspective. It was also necessary for a parser class for the XML schema to be added but a boon was that the code that had resulted from the JISC Research Data Spring "Enabling complex analysis" project (mentioned above) could be reused.

Once the data was finally mounted it was possible to reuse this code and other specifically written R and Python scripts to mine the data. We continued, however, to note that the computing infrastructure was not really set up to support humanities research. For example, the idea behind the reference data sets on Legion is that they shouldn't be changed once stored. As David Smith and Ryan Cordell note, "the proportion of erroneous words in nineteenth-century newspapers can exceed 40% error rates can be even higher for other languages and earlier periods" (2018, 5). In a situation where our licence would have allowed us to undertake, for example, some form of post-correction of the OCR transcripts, or even some combination of automated and manual correction, the time barrier involved in remounting the data would have been significant enough to deter us.

Moving on to the experience of seeking to query the material, although we were immensely grateful to UCL Research IT Services, we found our efforts to use university-based high-performance computing (HPC) to perform text mining to be a fraught experience from a number of perspectives. The example that we will now present relates again to the peculiarities of the cultural heritage datasets that we sought to work with and the questions that we sought to bring

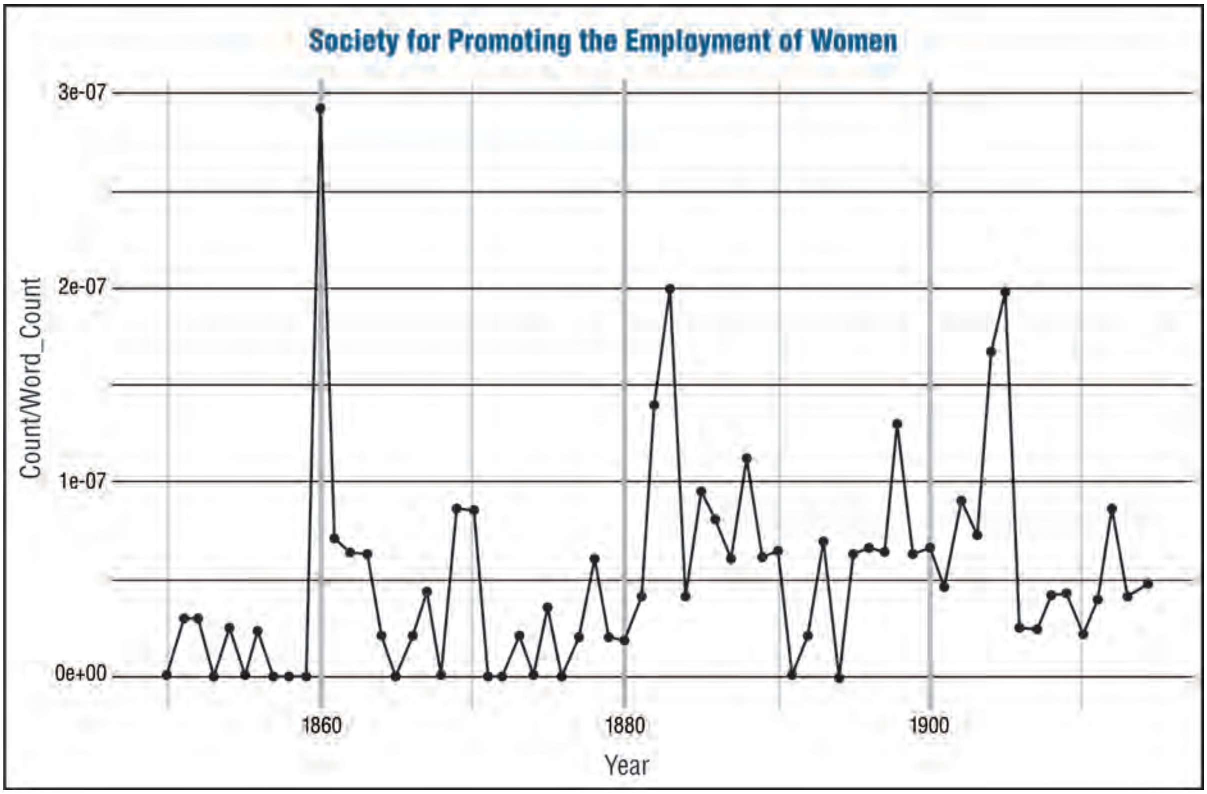

Figure 1. Plot showing the normalised occurrences of the phrase "Society for Promoting the Employment of Women" from 1850 through 1914 in the TDA dataset.

to bear on them. As stated above, digitised newspaper text is relatively unstructured and tends to be riddled with OCR errors. This can be especially challenging when dealing with multilingual sources according to language-specific rules, which are challenging for both the humanities researchers and the research software engineers in terms of the linguistic and computationally linguistic skill sets required. Even simple queries on monolingual texts can involve an element of unpredictability, since even this might retrieve an unmanageable amount of results or not return any meaningful results, confounding expectations and necessitating either further refining or widening the search parameters.

For example, we wished to use the TDA hard drives to conduct research on how female emigration was reported in nineteenth-century newspapers. We were especially interested in how these papers discussed and reported the emigration of women to British colonies of the time, like South Africa, New Zealand, and Australia. Via our initial queries, we searched for articles containing the names of the female emigration societies that were operating during the second half of the nineteenth century. During this time, emigration was increasingly being facilitated, assisted, and managed by charitable societies who saw it as their mission to send women overseas to populate the countries they had colonised (Constantine 1991, 95; Bush 1994). Our initial queries concentrated on plotting the occurrence of mentions of these emigration societies in newspaper articles during the period 1850 through 1914. To support this, we researched and drew up a list of names of emigration societies and sent this to the software engineers with whom we were working. Given the length of the names of these societies (for example, "British Women's Emigration Association) and the vagaries of OCR quality, each of the names was manually assigned a defined "error tolerance" based on word density per title. Our aim here was to avoid, on the one hand, underfitting of the results whilst, on the other, ensuring we did not introduce too many false

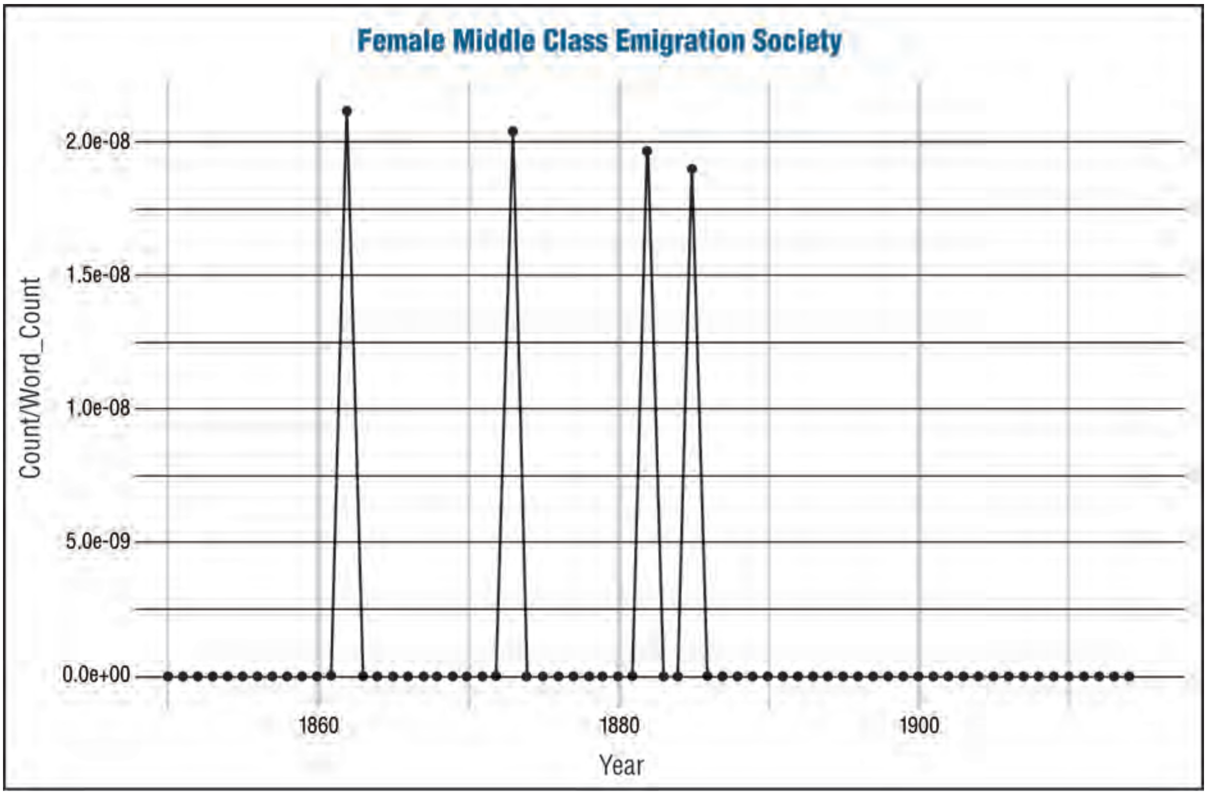

Figure 2. Plot showing the normalised occurrences of the phrase "Female Middle Class Emigration Society" from 1850 through 1910.

positives in the process. When we received back the first plots, however, there was huge variance in the results for each of the emigration societies.

The results for the "Society for Promoting the Employment of Women" indicated that the society was mentioned and discussed in newspaper articles across the decades. The plot for the "Female Middle Class Emigration Society" included hardly any mentions from the mid-1880s. Yet, the existing historical research about this society indicates that it was especially active from the 1880s onwards (Chilton 2007). Why was this not borne out by the plots? It became clear to us that it would be necessary to refine and recalibrate the assigned error tolerances in order to produce more reliably representative results. However, without clear knowledge about the OCR quality for each year and issue of the TDA, it was clear that this process would involve several iterations through our list of about 25 emigration societies.

}Many digital humanities projects that work with noisy data stress that the process should involve iterative stages of cleaning or otherwise refining data, generating and adapting queries, and further refining the working hypothesis of the research (see Schöch 2013). Yet, remember that this work is costed at £350 a day. Even in impressive-looking major research grants (in the high six-digit area), when institutional overheads are taken off and the project budget is split between multiple project partners, resources are tight. Indeed, it quickly becomes apparent that the luxury of undertaking the desired iterative approach may exceed project funds. In this way, our experience of working on projects like this is that they proceed by trial and error. When the TDM drives arrive one can feel like a world of opportunity is opening up, given the TBs of content they offer. Yet, as we have summarised, grappling with the institutional, logistical, financial, bureaucratic, and other hurdles that follow take a significant amount of time that cannot adequately be reflected in the published outputs of the project.

GALE DIGITAL SCHOLAR LAB

From a number of perspectives, then, cloud-hosted tools like Gale Digital Scholar Lab offer an immense step forward. The struggles outlined above in terms of securing and negotiating access to the underlying data; mounting the data; thinking through the peculiarities or limitations of the data and how this might be accounted for computationally; running queries on the data and seeking to work iteratively so that emerging insights can be integrated are, to varying extents (and providing that your institution has the appropriate subscriptions), well provided for.

Gale Digital Scholar Lab (DSL) provides an interface that is highly usable: a number of core text-mining routines can be applied to the data: clustering, n-grams, sentiment analysis, and topic modeling. Named entity recognition and part of speech tagging are also supported and the annotated data can be downloaded for further transformation, like network analysis. The lab is straightforward to use and thus invites explorative research and playful engagement with text. Put in a form that popularises these approaches, it offers a low-barrier entry into digital scholarship–especially for students unfamiliar with text mining. The explanations about the tools that Gale DSL provides are concise and helpful; they do not overburden the more casual and inexperienced user. For these reasons, the platform lends itself for teaching purposes. Much potential can also be noticed in the way that Gale DSL can help to further embed digital forms of enquiry in the scholarship of people who might not otherwise have undertaken elements of this research. In this way Gale DSL contributes to the task of raising the visibility of digital humanities research across the disciplines.

In addition, Gale Digital Scholar Lab also goes some way to meeting the needs of the more advanced user and researcher. Amongst its offerings, we note, is the ability to create custom content and to personalise folders with subsets of relevant articles in a way that is easy and intuitive to manage and organise. Searches can be retained to work with these personalised corpora. Upon getting trial access to the lab, we were concerned that it offered access to licenced material only, and about the new silos of open versus licenced content analysis that the lab might cement. Yet, as we understand it, the possibility of adding external, open licence collections will follow (of course, some would rather not see open licenced materials used within a proprietary platform such as this, but that is another issue).

Moreover, because the original primary source document and OCR text are presented next to each other, it is immediately possible to compare the two and assess the OCR against the original source. Additionally, the user receives detailed metadata information together with an OCR confidence rating, which provides them with some guidance as to how to interpret and treat the results they have generated.

Information about OCR confidence is usually missing from standard digital newspaper search interfaces, but incorporating it in the Gale Digital Scholar Lab will be welcome news for researchers as it gives a better understanding of the collection in question. This openness about the quality of transcription and the provenance of the data is, we believe, crucial to fostering a more critical engagement with digital sources and we hope that more providers will consider implementing such functionalities going forward.

Further, the ability to export tabular data allows more advanced users to take control of their results and to use these for additional visualisations or simply to document their workflow. This is an important aspect for those researchers who are seeking to publish their results and require their findings to be reproducible and repeatable. Again, as it currently stands, when researchers rely on standard search interfaces, it is often difficult to document the workflow in a satisfactory way that meets these requirements. The Gale Digital Scholar Lab, therefore, goes some way towards helping scholars achieve these standards because it can facilitate transparency about how the results of their text mining analysis has been arrived at.

Nevertheless, and we do not mean to imply that this is part of Gale/Cengage's proposition, the availability of cloud-based tools like Gale Digital Scholar Lab do not obviate the need to use high-performance computing or other infrastructures to undertake deep dive text mining–led research.

To discuss our experiences with the Gale DSL in a bit more detail, we will return to our previous example about female emigration societies. We ran a search of the TDA across entire documents (one can choose from, e.g., document title or place of publication) restricted to the years 1850 through 1914. We further restricted our search to content types labelled as "newspapers and periodicals." Exact matches of the phrase "Society for Promoting the Employment of Women" numbered 54 (or 58 if not an exact match). This was starkly different from the number of results that we had obtained from the TDA drives (162 matches). The difference is apparently caused by the error rates that the lab sets, which are different from the error rates that we had assigned when working with the TDA drives. It was not possible for us to change the Gale Digital Scholar Lab's error rates, though the use of regular expressions is permitted in searches. This may offer one way of forcing an adjustment of the lab's error tolerances, but it is not an approach that would scale well to the running of a large number of queries.

In terms of the results that were returned, we found the lab's interface and presentation of metadata helped us to swiftly gain an overview of the kind of material that comprised those 54 occurrences of the search term. We could immediately see the document types in which the results occurred: article (34), advertisement (10), and letter to the editor (10). A breakdown of subjects is also given: 11 in total with the top 3 being working women (9), associations (5), chambers of commerce (1). Articles can also be clustered under those subject headings. One illustrated work, specifically a table, was also returned. The timeframe of the articles that the keywords occurred in was from 1860 through 1913. The OCR confidence levels varied widely, from in excess of 90% down to the low 40%. After adding the documents to a content set, a list of 10 of the authors of those documents also became available. When it came to interrogating those results, we found that we could use the lab to generate some views of our data that may offer some interesting ways of reading it. For example, we could generate a chart of sentiment across the years 1860 to 1914 that might further avenues of investigation (see figure 3).

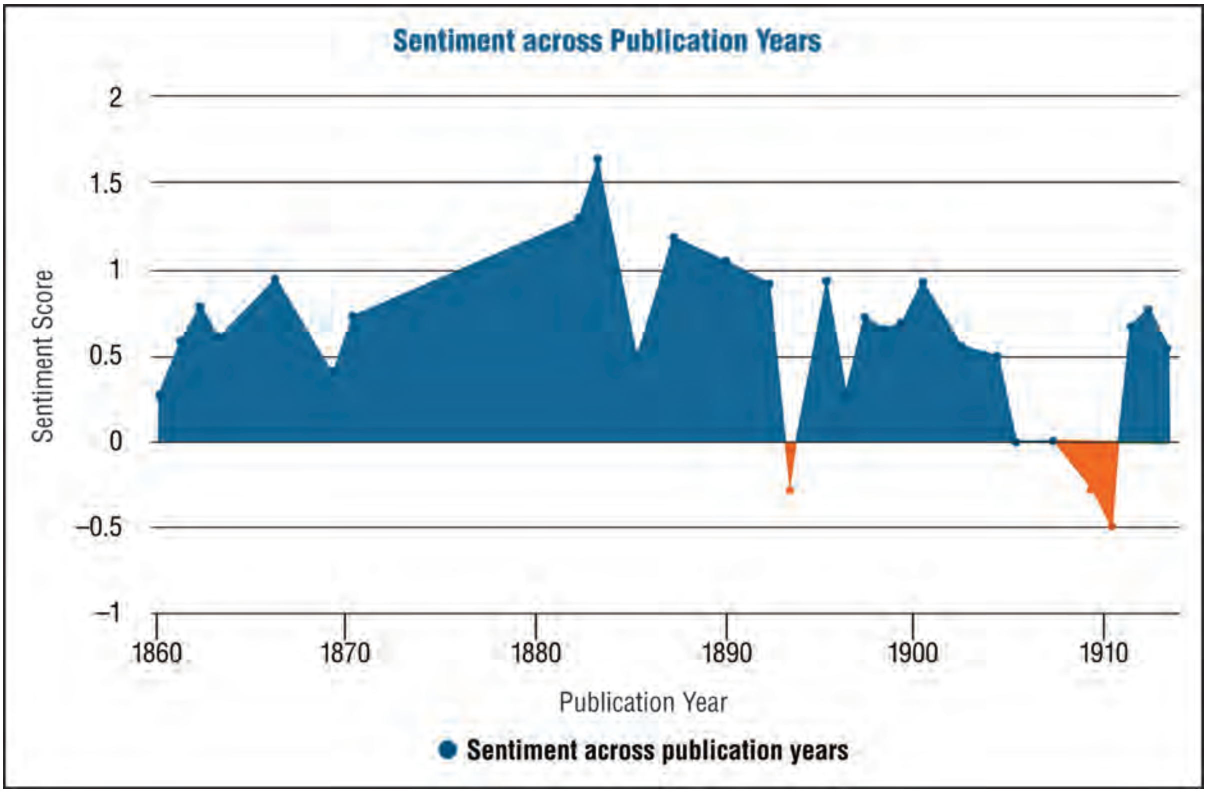

Figure 3. Plot showing the sentiment across the publication years 1860 through 1914.

However, the lab does not currently offer the routines that would allow us to plot the data in the way that we require for our research. Indeed, within the Digital Scholar Lab it is not possible to write one's queries and so the usefulness of the lab for the research we had hoped to undertake is limited. That said, we certainly can imagine a scenario where researchers could use the lab as a site to explore and critically evaluate their datasets, and that this process could then inform and enhance their subsequent use of costly and more bespoke HPC-based interrogation of their data.

With regard to the lab, we would also point out that currently it does not seem possible to use the output of one routine as the input of another so that cascading digital workflows can be built up. Custom scripting on HPCs like those offered by our own Research IT Services (or commercial alternatives like Amazon Web Services) are needed for that. Also, at present, Gale Digital Scholar Lab's support for collaborative work is limited. So, while it will be most useful in some contexts, like teaching on introductory modules and for individual users, it is currently not scaled towards sustained text mining work undertaken by large-scale and geographically dispersed project teams. These projects would need dedicated functionalities for larger group accounts in which results can be stored and accessed by different team members, who might require some bespoke functionalities.

Yet of course, we are aware that no one tool can meet the needs of every type of researcher, let alone every type of research group. It might well be the case, therefore, that cloud-based platforms will function as an intermediary that researchers can trust to perform the "standard" text mining functionalities to exacting and reliable standards. For the great majority of users these will be sufficient, while complex operations that require custom scripting will continue to be performed in other settings.

CONCLUSION

We will close with some general reflections on our experiences of seeking to use university-based high-performance computing infrastructures to mine large-scale, commercially digitised nineteenth-century historical newspaper collections. We are certainly not the first researchers to communicate the difficulties that we have encountered when attempting to undertake this work (and it is in this context writ large that wider pan-European initiatives like CLARIN operate (Wynne 2013) as well as the open-source, web-based Voyant Tools (Rockwell and Sinclair 2016)). Yet, even though, "DH needs have been growing more complex, as humanists—like scholars in many other fields—tackle research questions with larger-scale data than was previously possible" (Dombrowski and Lippincott 2018), the infrastructures required to support this kind of computationally intensive work in the humanities are evolving unsteadily. What to do? The recommendations of the "Enabling large scale data-analysis" project mentioned above included the following:

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 (Terras et al. 2018, 463).

These recommendations are useful and we hope that they will be taken up. Nevertheless, it seems important to also draw attention to a broader point of research strategy and foresight regarding computationally intensive humanities work that we believe our experiences and these recommendations speak to.

King's College London's digital lab is a rarity in that it employs research software engineers on permanent contracts, in a dedicated centre, with clear professional progression paths, many of whom have a wealth of experience of working as and with humanities and digital humanities researchers to model and implement research software for cultural heritage and arts and humanities sources.[8] Elsewhere, this investment is much more uneven. So, yes, the case needs to be made to engage research software engineers to work with open data sets, but a much wider case needs to be made too about the crucial and more wide-ranging role that research software engineers (and the infrastructures they participate in and enable) will play as research in the humanities becomes ever more digital. At the present time, for the humanities researcher it can be difficult even to find out that research software engineering posts exist in their university and, given the often-unclear career paths for these posts in universities, and the much better conditions that research software engineers can attain outside of the university, these posts can have a high turnover rate. This is unfortunate for the individuals involved and difficult in terms of the management of limited-term grants.

As such, we believe that our experiences in the difficulties of seeking to mine large-scale, digitised cultural heritage collections, whether open or gated, is symptomatic of the struggle of many universities to adequately respond to the wider investment proposition that is required and opened by the digital turn. To return to Kaltenbrunner's discussion of infrastructure and how it involves the interlacing of the global and the local, it seems that the difficulties faced in local contexts can accordingly have implications of global reach. The difficulties faced when seeking to mine large data sets can have global implications in terms of those wider conversations and contexts that humanities scholars may be impeded from contributing to. The difficulties of doing this work limits not only the questions that humanities researchers can ask of sources at scale; it also reduces the visibility of the humanities, and potentially impedes initiatives to develop partnerships between universities and other external actors like commercial companies and the creative economy. These difficulties also limit our opportunities to pursue and shape the economic proposition of digital cultural heritage in the wider digital economy.

Returning, then, to Gale Digital Scholar Lab, perhaps one of the most interesting aspects of it is how it invites us to think about innovation and enterprise in digital humanities, and the potential—and of course pitfalls—of working with industry in order to find perhaps less familiar and extra-university-based ways to facilitate crucial aspects of the infrastructures that are needed to undertake, for example, computationally intensive interrogation of large-scale digital cultural heritage data. Mutually beneficial collaborations undertaken in this space may result in new ways to convince universities of the value proposition of roles like research software engineers, of the need to further invest in the infrastructures that can support digital research in the humanities, and of the multifaceted impacts of digital humanities. Through the private-public partnerships that have been undertaken to digitise cultural heritage materials much has been learned, by both public and private bodies, of the benefits and pitfalls of working together. We must continue to draw on these experiences as we think through the kinds of infrastructures and access we require to facilitate the digitally led interrogation of large-scale, open, and licensed data cultural heritage collections.

NOTES

1. The views that we present in this paper are our personal views as individual researchers; they should not be taken to express the views or opinions of University College London.
2. https://pro.europeana.eu/resources/apis [accessed May 1, 2019].
3. https://pro.europeana.eu/post/experimental-text-dumps-from-europeana-newspapers [accessed May 1, 2019].
4. https://www.ucl.ac.uk/digital-humanities/.
5. https://www.ucl.ac.uk/research-it-services/ [accessed May 1, 2019].
6. https://www.ucl.ac.uk/isd/services/research-it-services.
7. https://wiki.rc.ucl.ac.uk/wiki/Cluster_Computing.
8. See https://www.kdl.kcl.ac.uk/.

REFERENCES

American Council of Learned Societies. 2006. *Our Cultural Commonwealth: The Report of the American Council of Learned Societies Commission on Cyberinfrastructure for the Humanities and Social Sciences*. Accessed June 13, 2009. https://www.acls.org/uploadedFiles/Publications/Programs/Our_Cultural_Commonwealth.pdf.

Anderson, Sheila, and Tobias Blanke. 2012. "Taking the Long View: From e-Science Humanities to Humanities Digital Ecosystems." *Historical Social Research* 37, no.3: 147–64.

AsymEnc (website). Accessed May 1, 2019. http://asyenc.eu.

Blanke, Tobias, Conny Kristel, and Laurent Romary. 2015. "Crowds for Clouds: Recent Trends in Humanities Research Infrastructures." Accessed July 8, 2019. http://arxiv.org/abs/1601.00533.

Bush, Julia. 1994. "'The Right Sort of Woman': Female Emigrators and Emigration to the British Empire, 1890–1910." *Women's History Review* 3, no. 3: 385–409.

Chilton, Lisa. 2007. *Agents of Empire: British Female Migration to Canada and Australia, 1860s–1930*. Toronto: University of Toronto Press.

Comité des Sages. 2010. *The New Renaissance. Report of the 'Comité Des Sages.'* Reflection Group on Bringing Europe's Cultural Heritage Online. https://ec.europa.eu/digital-single-market/sites/digital-agenda/files/final_report_cds_0.pdf.

Constantine, S. 1991. "Empire Migration and Social Reform 1880–1950." In *Migrants, Emigrants, and Immigrants: A Social History of Migration*, ed. by Colin G. Pooley and Ian Whyte. New York: Routledge, pp. 62–87.

Darnton, Robert. 2010. *The Case for Books: Past, Present, and Future*. New York: PublicAffairs.

Dombrowski, Quinn, and Joan Lippincott. 2018. "Moving Ahead with Support for Digital Humanities." *EDUCAUSE Review*, March 12, 2018. Accessed July 8, 2019. https://er.educause.edu/articles/2018/3/moving-ahead-with-support-for-digital-humanities.

Fyfe, Paul. 2016. "An Archaeology of Victorian Newspapers." *Victorian Periodicals Review* 49, no. 4: 546–77. https://doi.org/10.1353/vpr.2016.0039.

Gale. "The Times Digital Archives 1785–2013." https://www.gale.com/intl/c/the-times-digital-archive.

Gooding, Paul. 2014. "Search All about It: A Mixed Methods Study into the Impact of Large-Scale Newspaper Digitisation." PhD diss., University College London.

Hobbs, Andrew. 2013. "The Deleterious Dominance of *The Times* in Nineteenth-Century Scholarship." *Journal of Victorian Culture* 18, no. 4: 472–97. https://doi.org/10.1080/13555502.2013.854519.

Kaltenbrunner, Wolfgang. 2015. "Scholarly Labour and Digital Collaboration in Literary Studies." *Social Epistemology* 29, no. 2: 207–33. https://doi.org/10.1080/02691728.2014.907834.

Milligan, Ian. 2019. *History in the Age of Abundance?: How the Web Is Transforming Historical Research*. Montreal: McGill-Queen's University Press.

Milligan, Ian. 2013. "Illusionary Order: Online Databases, Optical Character Recognition, and Canadian History, 1997–2010." *The Canadian Historical Review* 94, no. 4: 540–69.

Oceanic Exchanges (website). Accessed May 1, 2019. https://oceanicexchanges.org.

Prescott, Andrew. 2014. "I'd Rather Be a Librarian: A Response to Tim Hitchcock, 'Confronting the Digital.'" *Cultural and Social History* 11, no. 3: 335–41. https://doi.org/10.2752/147800414X13983595303192.

Rockwell, Geoffrey, and Stéfan Sinclair. 2016. *Hermeneutica: Computer-Assisted Interpretation in the Humanities*. Cambridge, MA: The MIT Press.

Schöch, Christof. 2013. "Big? Smart? Clean? Messy? Data in the Humanities." *Journal of Digital Humanities* 2, no. 3 (Summer 2103). Accessed June 10, 2019. http://journalofdigitalhumanities.org/2-3/big-smart-clean-messy-data-in-the-humanities/.

Smith, David A., and Ryan Cordell. 2018. *A Research Agenda for Historical and Multilingual Optical Character Recognition*. Boston: Northeastern University; NULab; The Andrew W. Mellon Foundation. Accessed June 8, 2019. https://ocr.northeastern.edu/report/.

Star, Susan Leigh, and Karen Ruhleder. 1996. "Steps Toward an Ecology of Infrastructure: Design and Access for Large Information Spaces." *Information Systems Research* 7, no. 1 (March 1996): 111–37. https://pdfs.semanticscholar.org/9cfc/d2dfe7927451f2c39617e6ac0aa499fd2edb.pdf.

Stevens, Hallam. 2013. *Life Out of Sequence: A Data-Driven History of Bioinformatics*. Chicago: University of Chicago Press.

Tanner, Simon, Trevor Muñoz, and Pich Hemy Ros. 2009. "Measuring Mass Text Digitization Quality and Usefulness." *D-Lib Magazine* 15, no. 7/8 (July/August 2009). https://doi.org/10.1045/july2009-munoz.

Terras, Melissa M. 2011. "The Rise of Digitization: An Overview." In *Digitisation Perspectives*, edited by Ruth Rikowski, 3–20. Rotterdam: Sense Publishers.

Terras, Melissa M. 2015. "Bluclobber, or: Enabling Complex Analysis of Large Scale Digital Collections." UCLDH Blog. Accessed May 7, 2015. https://blogs.ucl.ac.uk/dh/2015/05/07/bluclobber-or-enabling-complex-analysis-of-large-scale-digital-collections/.

Terras, Melissa, James Baker, James Hetherington, David Beavan, Martin Zaltz Austwick, et al. 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, no. 2: 456–66. https://doi.org/10.1093/llc/fqx020.

UCL. 2017. “Partners in Time: HPC Opens New Horizons for Humanities Research.” Research IT Services, October 24, 2017. Accessed 8 July 2019. https://www.ucl.ac.uk/research-it-services/news/2017/oct/partners-time-hpc-opens-new-horizons-humanities-research.

Wilson, James A. J. 2017. “The Research Data Storage Service at UCL – A LEARN Case Study.” In *LEARN Toolkit of Best Practice for Research Data Management*, Leaders Activating Research Networks (LEARN), pp. 78-81. https://doi.org/10.14324/000.learn.00.

Wynne, Martin. 2013. “The Role of CLARIN in Digital Transformations in the Humanities.” *International Journal of Humanities and Arts Computing* 7, no. 1–2: 89–104. https://doi.org/10.3366/ijhac.2013.0083.