Journal of Siberian Federal University. Humanities & Social Sciences 4 (2016 9) 994-1006
~ ~ ~

УДК 303.436.2:004.352

From Digital Resources to Historical Scholarship with the British Library 19th Century Newspaper Collection

Ian Gregory, Paul Atkinson, Andrew Hardie, Amelia Joulain-Jay, Daniel Kershaw, Catherine Porter, Paul Rayson and CJ Rupp*
Lancaster University
Lancaster LA1 4YT United Kingdom

Received 04.12.2015, received in revised form 20.01.2016, accepted 18.03.2016

It is increasingly acknowledged that the Digital Humanities have placed too much emphasis on data creation and that the major priority should be turning digital sources into contributions to knowledge. While this sounds relatively simple, doing it involves intermediate stages of research that enhance digital sources, develop new methodologies and explore their potential to generate new knowledge from the source. While these stages are familiar in the social sciences they are less so in the humanities. In this paper we explore these stages based on research on the British Library's Nineteenth Century Newspaper Collection, a corpus of many billion words that has much to offer to our understanding of the nineteenth century but whose size and complexity makes it difficult to work with.

Keywords: Corpora, GIS, Resource enhancement, Research Methods, OCR quality.

DOI: 10.17516/1997-1370-2016-9-4-994-1006.

Research area: culture studies.

Introduction

Elsewhere we have argued that the biggest challenge for digital historians is to take the wealth of digital resources that are in existence and use these to create new scholarship that makes applied contributions to our knowledge that are of interest to historians beyond digital history (Gregory 2014). Large amounts of historical sources have been digitised at significant expense to both the public and private sectors and it has been claimed that historians have failed to make a good return on this investment. Critics focus on the relatively poor quality of much digitised material and often argue that the bulk of the work done using these sources simply makes uncritical – and often unacknowledged – use of key-word searches and other basic functionality provided by the web interfaces through which the sources are made available (Hitchcock 2013). Moving beyond this to make better use of the opportunities that digital sources offer is essential if history is to thrive in the digital age. However, the journey

* Corresponding author E-mail address: inna.kizhner@gmail.com

from digitised source to new contributions to knowledge is not a simple one and frequently involves significant amounts of investigation of the opportunities and limitations of the digital source and the methods to be used on it. Intermediate research of this type is common in the social sciences and elsewhere but is unfamiliar within the humanities. Intermediate work in itself provides new and valuable contributions to knowledge as researchers can use the lessons learnt on different sources and different topics. It is, however, important to remember that the researcher's ultimate aim must be to contribute to research that makes an applied contribution to knowledge.

In this paper we explore an example of some of these intermediate stages that lie between a complex electronic source and delivering an applied contribution to knowledge. The source in question is the British Library's Nineteenth Century Newspapers collection, a corpus of around 48 newspaper titles that make up a usually complete run of daily or weekly papers covering the entire century (British Library n.d.). This is a massive and complex resource. It consists of over two million newspaper pages with over 30 billion words that take up around a terabyte of storage space in digital form. Our ultimate aim is to make use of this resource to better understand the historical geographies of nineteenth century Britain and its relationships with the wider world. Getting to this stage is, however, far from easy given the volume and complexity of the source and the need to develop methodologies and conduct pilot analyses to evaluate them – particularly given the sheer volume of data, the digitisation errors that it contains, and our emphasis on understanding geography. We identify at least six intermediate stages to the analysis in two major sections. The first four stages are concerned with understanding the digital version of the source itself and include: cleaning up the original mark-up in the text, attempting to correct the digitisation errors caused by the use of Optical Character Recognition (OCR) technology, exploring the implications of OCR errors, and geoparsing the source to allow geography to be explored. The second section involves conducting exploratory analyses to investigate the techniques and get an overview of the types of results that they return. In this case these include conducting corpus analyses and exploring the text geographically. A possible seventh stage is the need to re-archive resources that have been enhanced in the research process so that other researchers can use them.

Understanding and enhancing the digital source

The Nineteenth Century Newspapers corpus arrived with us on disk in the format shown in Fig. 1. This fragment shows the start of one article from the Liverpool Mercury on the 25th July 1873. There are only eleven words of actual text in the fragment "MERSEY DOCKS BOARD. The weekly meeting of the Mersey Docks ar.(d [sic]"; the rest of the information is metadata associated with the article and the individual words within it, marked up in XML (eXtensible Mark-up Language) format. This metadata gives some useful information, such as the article title (<dc:Title>), the newspaper title (<title>) and the date (<printedDate>), but much that is irrelevant for our purposes, especially the coordinates of the individual words on the page (<articleWord coord="…">). The first requirement was to convert the text into a suitable form for analysis in CQPweb, a corpus software package (Hardie 2012). This entailed writing a program that would convert the required mark-up to a suitable format while removing the remaining superfluous mark-up. Due to the size of the corpus this required parallel processing on a Hadoop cluster. Once this had been done the resulting text could be run through part of speech and semantic taggers

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE BL_newspaper SYSTEM "BL_newspaper.dtd">
<BL_newspaper xmlns:dc="http://purl.org/dc/elements/1.1/" xmlns:dcterms=
<BL_article>
<title>Liverpool Mercury etc</title>
<normalisedTitle>Liverpool Mercury</normalisedTitle>
<name>Liverpool Mercury etc</name>
<placeOfPublication>Liverpool</placeOfPublication>
<issue_metadata>
<volumeNumber></volumeNumber>
<issueNumber>7960</issueNumber>
<printedDate>FRIDAY, JULY 25, 1873</printedDate>
<normalisedDate>1873.07.25</normalisedDate>
<dc_metadata>
<dc:Title>MERSEY DOCKS BOARD.</dc:Title>
<dc:Subject></dc:Subject>
<dcterms:issued>1873.07.25</dcterms:issued>
<dc:Type>Image</dc:Type>
<dc:Type>Newspaper article</dc:Type>
<dc:Type>News</dc:Type>
<dc:Identifier>WO1_LVMR_1873_07_25-0003-008.xml</dc:Identifier>
</dc_metadata>
<articleImage>
<articleSequence>0003-008</articleSequence>
<articleImageFile>WO1_LVMR_1873_07_25-0003-008.tif</articleImageFile>
<articleCoordinates>4253,3306,5003,8025</articleCoordinates>
<articleText>
<articleWord coord="106,21,328,76">MERSEY</articleWord>
<articleWord coord="303,20,480,75">DOCKS</articleWord>
<articleWord coord="462,20,642,74">BOARD.</articleWord>
<articleWord coord="60,102,138,139">The</articleWord>
<articleWord coord="132,102,251,139">weekly</articleWord>
<articleWord coord="237,102,374,138">meeting</articleWord>
<articleWord coord="365,102,411,137">of</articleWord>
<articleWord coord="412,101,473,137">the</articleWord>
<articleWord coord="458,101,583,137">Mersey</articleWord>
<articleWord coord="569,100,672,137">Docks</articleWord>
<articleWord coord="667,100,737,136">an(d</articleWord>
```

Fig. 1. A fragment of the raw mark-up from the British Library Nineteenth Century Newspapers Collection.

(Garside et al 1987; Rayson et al 2004; Wattam et al 2014) and finally loaded into CQPweb. This pipeline took approximately a week to run on the *Era*, a weekly paper published from 1838 that consists of 375 million words.

Although in theory the text is now ready for analysis there were two subsequent stages that we wanted to explore to evaluate the impact of digitising errors. The digital text was captured using OCR technology which attempts to automatically recognise the letters and words from image scans of the original source. Given that the original was newsprint that was over a century old and frequently in poor condition this inevitably resulted in significant numbers of OCR errors (Tanner et al 2009). A simple example of this is visible in Fig. 1 where a "(" has been added to the word "and". Our first question was whether we could use automated techniques to correct some of these errors, the second to explore the extent to which the errors undermine subsequent analyses.

Automated correction of OCR errors is receiving an increasing amount of research attention (see, for example, Daðason et al 2014; Reynaert 2008; Volk et al 2011; Wick et al 2007). After evaluating the various possible methods we decided to further explore three techniques: the first was a line-filtering method designed to remove the areas of the text with the worst OCR errors. This was implemented by finding how many words within each line of text could be located in a lexicon and how many could not. When the ratio between these fell below a certain

threshold it was assumed that the OCR was so garbled as to be unusable and it would therefore be dropped from the final text. All subsequent lines would be dropped until the ratio rose above a second threshold. This was found to be effective at removing areas where the OCR could not be expected to succeed, such as adverts with unusual fonts, and the most garbled parts of the OCR. It therefore reduced the size of the corpus and made it more readable but potentially removed material of interest.

Rather than remove text, the other two methods attempted to correct errors. The first of these used the freely-available VARD software developed at Lancaster University to standardise historical texts (Baron and Rayson 2008). The second used a commercial correction method called OverProof (Evershed & Fitch 2014; Project Computing n.d.). The correction methods were evaluated against a 160,000-word gold-standard corpus derived from the Corpus of Nineteenth Century Newspaper English (CNNE), a hand-corrected corpus that overlaps with the British Library Nineteenth Century Newspapers Collection (English Department Uppsala University n.d.; Varieng 2014). Statistical analysis of the results showed that VARD yielded very few improvements and created new errors to the extent that the corrected files were essentially as different from the gold-standard files as the originals were. With OverProof, however, there was a significant improvement such that it may well be worth applying this correction method to the entire corpus. The difficulty with this is that OverProof is commercial software and paying to improve this volume of material is beyond the scope of an individual research project while, unfortunately, the freely available academic software was not deemed successful enough to be worth attempting. While it is encouraging that there are potential ways of improving the OCR in large corpora, we currently have to use uncorrected, noisy text.

Given that the OCR resulted in significant errors a further question then becomes the extent to which this error affects the results of any analysis conducted on a single source known to contain significant OCR errors. In conventional analyses conducted through a web interface the implications will be low for browsing, as long as the reader is able to interpret the text, but potentially serious for key-word searching as the search will miss any instances of a search-term containing an error. Fuzzy searching may help reduce this. Our analytic work makes extensive use of techniques from corpus linguistics, particularly *collocation* which measures the extent to which two words are found near each other within the text (Adolphs 2006; McEnery & Hardie 2012). In a typical collocation analysis the aim will be to identify the words that collocate with a particular search-term, in other words, are found within a set number of words of the search-term more frequently than would be expected given how often each word occurs in the text. A range of statistical tests can be used to measure the extent to which two words collocate. In this analysis we explored two: log-likelihood statistics and MI scores. Log-likelihood statistics quantify the probability that the observed pattern of two words tending to occur together is due to chance with high scores suggesting that the degree to which they co-occur is unlikely to have occurred by chance. MI (mutual information) scores tend to be used to rank words that collocate, the higher the MI score the more strongly the two words are observed to collocate. The two measures have different characteristics: common words tend to attract higher log-likelihood scores; unusual words attract higher MI scores (Oakes 1998). For this reason they are often used together. The results were encouraging – they showed that both log-likelihood and MI statistics were broadly

reliable. We thus feel that collocation statistics can be used with the uncorrected Nineteenth Century Newspaper Collection and that OCR errors do not seriously undermine their effectiveness. An implication of this result is that it suggests that the line filtering approach, described above, is unnecessary as it removes potentially useful information for little gain.

A further way of enhancing the text is to geoparse it, a process that identifies place-names in the text and allocates them to coordinates so that their locations can be mapped. There are two strategies to geoparsing. The first is to geoparse the entire corpus before loading it into analytic software such as CQPweb. This is the approach used by the Edinburgh Geoparser for corpora such as Histpop and Bopcris (Grover et al 2010). We have not done this with the Nineteenth Century Newspaper Collection for two reasons: first, geoparsing is prohibitively slow for very large corpora, and secondly it is error prone and correcting the results would be difficult. Instead, we modified the Edinburgh Geoparser to implement what we term *concordance geoparsing* (Rupp et al 2014). This involves starting with a search-term associated with a theme of interest and extracting 50 words of text from before and after each occurrence of this term. These *concordances*, as they are termed, are then geoparsed giving a relatively small number of words that can be geoparsed relatively quickly. The results of this geoparsing can then be easily explored to check for errors using both concordances and maps. Where errors are spotted they can be corrected by adding them to an updates file that can be progressively built up as more search-terms are geoparsed. The advantage of this approach is that it allows the user to progressively geoparse a corpus in a way that enables them to be confident of how accurate the results are and what the potential errors may be. The obvious disadvantage is that the entire corpus is never geoparsed by this approach, however, it would be possible to do so using the additional information from the updates file once a user is confident in the accuracy of the geoparsing.

We have therefore adopted several stages of manipulation and analysis concerned with enhancing the dataset and understanding its strengths and limitations. This type of work takes considerable time and effort and provides valuable methodological lessons that may be of interest to the wider research community, even if it is not traditional humanities research.

Exploratory analysis of the digital source

Most historical research starts from the perspective of attempting to answer specific research questions. With digital projects that combine complex sources and new methods it is often desirable to conduct more exploratory analyses to evaluate the effectiveness of applying the methods to be used and to explore the source without preconceived questions. We applied these to the *Era*, a newspaper published in London whose main concern was theatre and sport but which also included national and international news (Brake & Demoor 2009).

The first stage of this was to explore how corpus techniques could be used to evaluate the way that two countries – France and Russia – were represented in this newspaper. The first step in doing this was simply to establish frequency counts of how often the names of these two countries were found in the *Era* in each year. The results of this are shown in Fig. 2. The most striking feature of this graph is the pattern of peaks and troughs. A preliminary hypothesis suggested by the graph is that the peaks represent discussions of major military and political events associated with the countries. Russia is frequently mentioned in the mid-1850s, the time of the Crimean War, and the late 1870s during the Russo-Turkish War. France

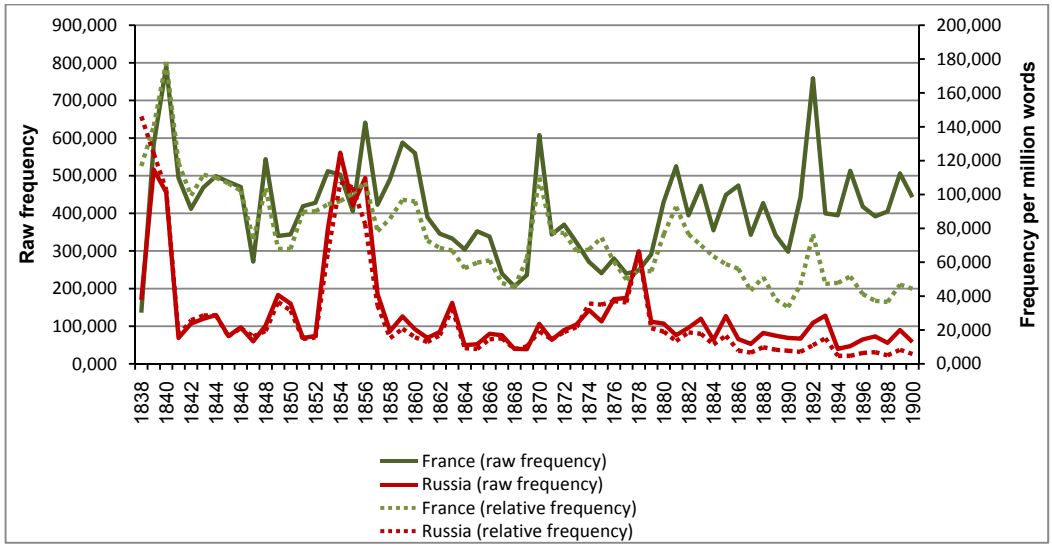
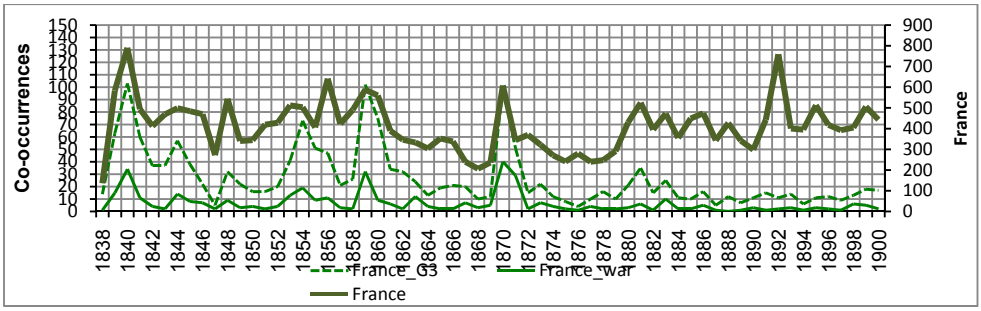


Fig. 2. Frequencies of instances of the terms 'France' and 'Russia' from the *Era*

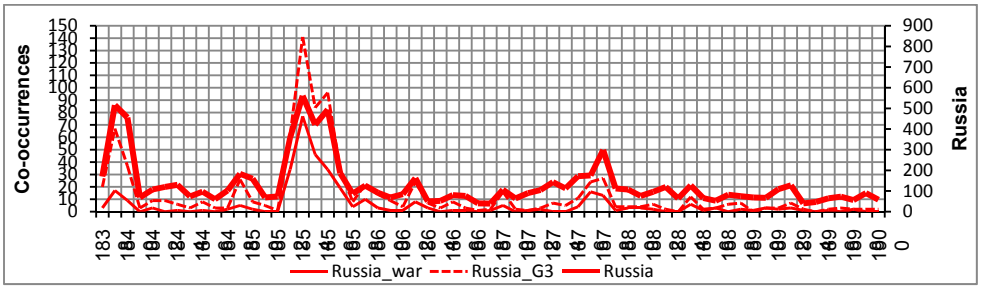
has spikes that coincide with the Crimean War, the Second Italian War of Independence in 1859, the Franco-Prussian War in the early 1870s, and the 1892 Conquest of Dahomey.

This might suggest that these major events are what drove interest in these countries. To investigate this further collocation can be used. In this case a relatively simple method was used, simply counting the number of times that 'France' and 'Russia' collocate with either the word 'war' or with words allocated by the semantic tagger to class 'G3', words associated with war. The results of this are graphed in Fig. 3. The graph shows that the proportion of instances which collocate with 'war' is so small that it is difficult to conclude that major military events are what was driving changes in the amount of attention dedicated to the countries. Equally, however, upsurges in 'war' and G3 do seem to be important contributors to most, but not all, of the spikes. One thing that is clear, however, is that there are also significant amounts of discussion of other things that are not military events. This is particularly true for France after the Franco-Prussian War.

This poses the question of what is thus of interest beyond war. Fig. 4 explores the nouns that collocate with each country and how these change over time. The graph suggests that for France politics and political power are of decreasing importance in the period leading up to the Franco-Prussian War. After the Franco-Prussian War there is a major rise in mentions of named individuals, including their initials and terms of address. There are also major rises in nouns associated with places, the performing arts and advertisements. This might be thought to suggest that following the Franco-Prussian War there was a rise in interest in French culture and particularly the theatre. However, confirming this would require close reading of some of the relevant articles to check that this is in fact the case. An alternative hypothesis is that the rise in advertising that occurred in the *Era* in this period leads to an increase in the count of mentions of France. For Russia there are obvious changes in the way that the country is represented but the pattern is less clear than for France. One particular question that this generates is the extent to which these patterns are reflective of broad discourses



a. France

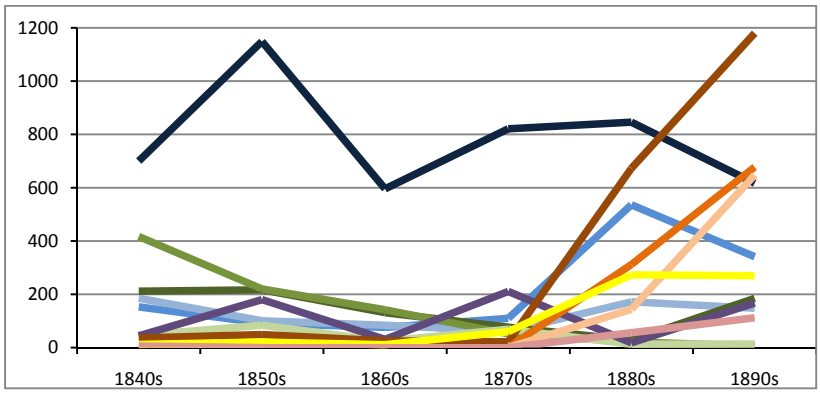


b. Russia

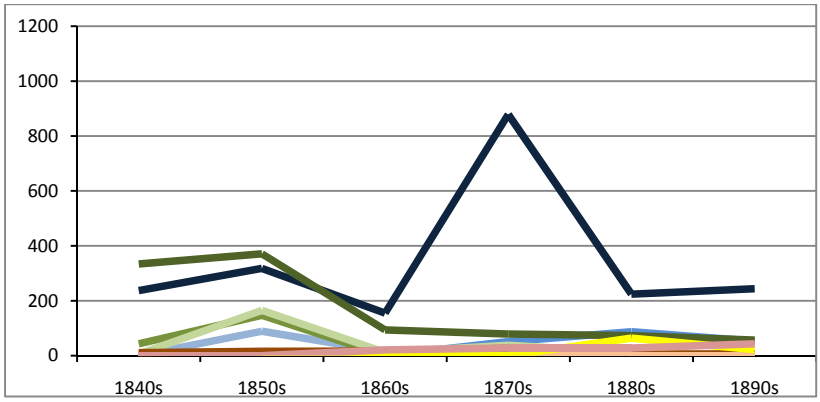
Fig. 3. Collocations between (a) 'France' and (b) 'Russia' and the word 'war' and words tagged as semantic class G3, words associated with war.

around the countries or simply reflections of changes in publication practices. It should also be noted that these findings are preliminary, however they do begin to suggest patterns that should be explored in future analyses.

In another type of exploratory analysis we explore the geography of infectious diseases as described by the *Era*. To do this a range of common nineteenth century infectious diseases were identified using a nosology described by Woods & Shelton (1997) that includes: food and water-borne diseases such as cholera and diarrhoea; diseases associated with overcrowding such as measles and typhus; and respiratory diseases such as bronchitis and influenza. The text around the names of these diseases was concordance geoparsed to identify the place-names that occur within ten words of the disease names in the texts and can thus be assumed to be places that are associated with the diseases. A global map of place-names that collocate with the diseases is shown in Fig. 5. Perhaps unsurprisingly, the pattern is clearly strongly associated with places in Britain and neighbouring countries. Beyond this, the places that stand out are eastern North America, the West Indies, Egypt, India and the Philippines. Exploring this pattern suggests that much of it is associated with patterns of global trade and the British Empire, with reports of deaths on ships between ports and the deaths of colonial officials being common-place. The very high number of instances in Egypt is, however, an exception to this. These are actually driven by advertisements for a medicine called 'fruit-salt' that contained testimonies from a man who claimed to have used it successfully while living there.



a. France



b. Russia

Legend: Country, Location (not country), Place (misc.), Position of power, Political, War, Newspaper production, Name, Initial, Terms of address, Performing arts, Adverts.

Fig. 4. Categories of nouns that collocate with (a) France and (b) Russia

Fig. 6 shows the pattern for the same diseases but focusing in on England and Wales. London and its surroundings show the highest concentration of disease mentions. Other places that are mentioned include some of England's other major cities such as Liverpool, Manchester, Newcastle, Leeds and Sheffield but other smaller places show similar amounts of attention to these including Preston, Lincoln and Brighton. The pattern actually derives from three different types of newspaper reports: those that are actual reports of deaths and disease, those that are reproductions of the Registrar General's Reports, and those that are advertisements for medicines.

The graphs and maps in this section all provide the beginnings of new analyses. They illustrate the potential for techniques, allow the researcher to refine and critically evaluate the techniques that they want to use, and to begin to define their research questions which might be concerned with how different countries were represented in the British media at different times or to what extent there was a relationship between the geography of reporting on diseases and the actual deaths from diseases in those locations.

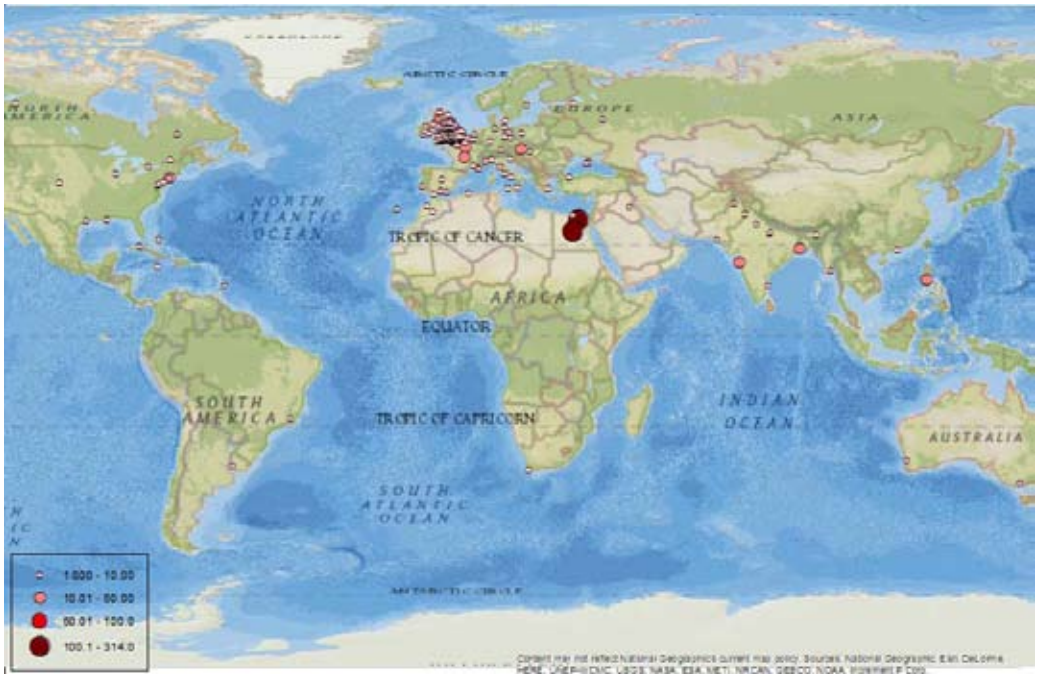


Fig. 5. Global place-names associated with major nineteenth century diseases in the *Era*.

Fig. 6. Places within England and Wales associated with major nineteenth century diseases in the *Era*.

Conclusions

A by-product of research such as this is that it enhances the source that it works with. Within the processes described here we have produced versions of the Nineteenth Century Newspapers Collection that have: reformatted the text to take out extraneous tags; removed sections of text to reduce OCR errors; automatically corrected the OCR to reduce errors (a process that will inevitably add new errors); added a range of tags including part-of-speech tags, semantic tags and geo-tags, adding new information on grammar, meaning and geography respectively; and combinations of these. These processes potentially create several new versions of the source that enhance the original digital copy but also add new errors and further abstract it from the original paper source. Although created for a specific research project, the resulting versions are potentially valuable to other researchers. This poses the question of how to disseminate these enhanced versions in ways that allow other researchers to evaluate whether they are suitable for their purposes and, if so, what the strengths and limitations of the enhanced versions are. This is not as easy as it sounds. Obviously enhanced sources can be placed in digital repositories or given to digital libraries for dissemination (subject to intellectual property rights), however the issue of how to document these enhanced versions in a way that allows subsequent researchers to evaluate them without posing an unrealistic burden in terms of creating documentation is difficult. There is also the question of how to give academic credit for the creation of such enhanced resources.

As stated in the introduction, the ultimate aim of digital history research must be to use digital sources to conduct research whose results are of interest to historians because of their contributions to our knowledge of history. With complex digital sources this is a far from rapid process and there are intermediate stages that can and should be navigated. These enable the researcher to explore, evaluate and enhance the opportunities and limitations offered by the digital version of the source and the methods to be applied to it. They also allow methods to be applied in an exploratory manner to get an overview of what they reveal about the particular source. Exploratory approaches are primarily concerned with using macro-analysis techniques to identify broad trends and patterns within the source. These are useful in part because they help us to understand the strength and weaknesses of the methods when applied to the particular sources under study. They are also of use because they help to frame subsequent research questions by potentially removing preconceived ideas of what the source is likely to tell us and which parts of it are important. In this way they can help remove the criticism that historians tend to plough well-ploughed furrows. These intermediate analysis stages take time but are rewarding in themselves as they generate new knowledge about sources and methods that can be applied by other researchers. They can and should also lead to publishable research in their own right. The one caveat to this is that they must be focussed on eventually leading to applied research as there is a risk that methodological research ends up becoming detached from the actual requirements of researchers.

This type of research is also highly interdisciplinary. The work described here is conducted as part of a research project based in a history department but, of the authors on the paper, arguably only one is a historian, the rest come from linguistics, computer science and geography. All bring their own distinct expertise and no one individual or even collection of individuals from a single discipline could hope to have the skills required to conduct such a project. An advantage of this is that research conducted under the auspices of history may be of interest

to researchers well beyond the discipline's traditional concerns.

Acknowledgements:

We are very grateful to the British Library for making the Nineteenth Century Newspapers Corpus available to us and particularly to Dr James Baker for his assistance with this. We are also grateful to: Claire Grover (University of Edinburgh) for her assistance with the Edinburgh Geoparser, Erik Smitterberg (Uppsala University) for allowing us to conduct analytic work on CNNE and Kent Fitch for correcting a sample of data using the OverProof software. The research leading to these results has received funding from the European Research Council (ERC) under the European Union's Seventh Framework Programme (FP7/2007-2013) / ERC grant "Spatial Humanities: Texts, GIS, places" (agreement number 283850).

References

Adolphs, S. (2006). *Introducing Electronic Text Analysis*. London, Routledge.

Baron, A., & Rayson, P. (2008). *'VARD 2: A tool for dealing with spelling variation in historical corpora.' Proceedings of the Postgraduate Conference in Corpus Linguistics. Aston University, Birmingham, UK*.

Brake, L., & Demoor, M. (2009). *Dictionary of nineteenth-century journalism in Great Britain and Ireland*. Gent, Academia Press.

British Library (n.d.). 19th Century British Library Newspapers Database. Available at: http://www.bl.uk/reshelp/findhelprestype/news/newspdigproj/database (accessed 4 January 2016).

Daðason, J.F., Bjarnadóttir, K., & Rúnarsson, K. (2014). 'The Journal Fjölnir for Everyone: The post-processing of historical OCR texts.' In *Language Resources and Technologies for Processing and Linking Historical Documents and Archives-Deploying Linked Open Data in Cultural Heritage–LRT4HDA Workshop.*

English Department, Uppsala University. (n.d.). The Corpus of Nineteenth-Century Newspaper English (CNNE). Available at: http://www.engelska.uu.se/Forskning/engelsk_sprakvetenskap/Forskningsomraden/Electronic_Resource_Projects/Nineteenth-century_Newspaper (accessed 4 January 2016).

Evershed, J., & Fitch, K. (2014). Correcting noisy OCR: Context beats confusion. *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*, Madrid, 45-51.

Garside, R., Leech, G., & Sampson, G. (1987). *The Computational Analysis of English: A corpus-based approach*. Harlow, Longman.

Gregory, I.N. (2014). 'Challenges and opportunities for Digital History,' *Frontiers in Digital Humanities* (1), 1-2.

Grover, C., Tobin, R., Byrne, K., Woollard, M., Reid, J., Dunn, S., & Ball, J. (2010). 'Use of the Edinburgh geoparser for georeferencing digitized historical collections,' *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 368 (1925), 3875-3889.

Hardie, A. (2012). 'CQPweb – combining power, flexibility and usability in a corpus analysis tool,' *International Journal of Corpus Linguistics*, 17, 380-409.

Hitchcock, T. (2013). 'Confronting the digital or how academic history writing lost the plot,' *Cultural and Social History,* 10, 9-23.

Project Computing (n.d.). OverProof: Automatic correction of OCR. Available at: http://overproof.projectcomputing.com (accessed 4 January 2016).

McEnery, A.M., & Hardie, A. (2012). *Corpus Linguistics: Method, theory and practice*. Cambridge, Cambridge University Press.

Oakes, M. (1998). *Statistics for Corpus Linguistics.* Edinburgh, Edinburgh University Press.

Rayson, P., Archer, D., Piao, S., & McEnery, T. (2004). 'The UCREL semantic analysis system.' *Proceedings of the workshop on Beyond Named Entity Recognition: Semantic labelling for NLP tasks in association with the LREC 2004*. Lisbon, 7-12.

Reynaert, M. (2008). 'Non-interactive OCR post-correction for giga-scale digitization projects.' In *Computational Linguistics and Intelligent Text Processing.* New York, 617-630.

Rupp, C.J., Rayson, P., Gregory, I., Hardie, A., Joulain, A., & Hartmann, D. (2014). 'Dealing with heterogeneous big data when geoparsing historical corpora.' *Proceedings of the 2014 IEEE Conference on Big Data.* Washington, 80-83.

Tanner, S., Munoz, T., & Ros, P.H. (2009). 'Measuring mass text digitization quality and usefulness: Lessons learned from assessing the OCR accuracy of the British Library's 19th Century Online Newspaper Archive,' *In D-Lib Magazine*, 15, available at: http://www.dlib.org/dlib/july09/munoz/07munoz.html

Varieng. The Corpus of Nineteenth-Century Newspaper English (CNNE) (2014). Avaolable at: http://www.helsinki.fi/varieng/CoRD/corpora/CNNE/index.html (accessed 4 January 2016).

Volk, M., Furrer, L., & Sennrich, R. (2011). 'Strategies for reducing and correcting OCR errors.' In *Language Technology for Cultural Heritage*. New York, 3-22.

Wattam, S., Rayson, P., Alexander, M., & Anderson, J. (2014). 'Experiences with parallelisation of an existing NLP pipeline: Tagging Hansard.' *LREC 2014, Ninth International Conference on Language Resources and Evaluation*. Reykjavik, 4093-4096.

Wick, M.L., Ross, M.G., & Learned-Miller, E.G. (2007). 'Context-sensitive error correction: Using topic models to improve OCR.' *Ninth International Conference on Document Analysis and Recognition.* Curitiba, 1168-1172.

Woods, R.I., & Shelton, N. (1997). *Atlas of Victorian Mortality.* Liverpool, Liverpool University Press.

От цифровых ресурсов к историческим знаниям: исследование на материале коллекции газет XIX века Британской библиотеки

Иан Грегори, Пол Аткинсон,
Эндрю Харди, Амелия Жулен-Джей, Дэниел Кершоу,
Кэтрин Портер, Пол Рейсон и Си Джей Рапп
Университет Ланкастера
Ланкастер LA1 4YT Великобритания

В последнее время становится все очевиднее, что цифровые гуманитарные науки чрезмерно ориентированы на создание новых данных, в то время как их главным приоритетом должно быть использование цифровых ресурсов для обогащения знаний. Хотя это звучит относительно просто, для достижения данной цели необходимо провести исследования промежуточной стадии для расширения цифровых ресурсов, развития новых методов и изучения их потенциала для получения новых знаний из имеющихся ресурсов. И хотя эта стадия исследования уже известна в области социальных наук, гуманитарные науки знакомы с ней в меньшей степени. Данная работа представляет собой изучение такой промежуточной стадии на примере исследования коллекции газет XIX века Британской библиотеки – корпуса текста объемом в несколько миллиардов слов, который предоставляет собой ценный материал о жизни XIX века, но в силу своего объема и сложной структуры является неудобным в обращении.

Ключевые слова: корпуса данных, GIS, увеличение ресурсов, методы исследования, качество распознавания сканированных данных.

Научная специальность: 24.00.00 – культурология.