Med. Hist. (2016), *vol.* 60(2), *pp.* 294–300.
doi:10.1017/mdh.2016.18

Media Reviews

Digitisation, Big Data, and the Future of the Medical Humanities

Text-Mining and the History of Medicine: Big Data, Big Questions?

Most of us have heard about 'Big Data', often as part of discussions about the information collected about us as consumers and citizens, and the increasingly sophisticated tools that analyse such information. But can we as historians of medicine benefit from thinking about our historical sources as 'Big Data', and 'mine' this data by adapting the tools used by commerce, computing science and intelligence? What possibilities for historical scholarship would such tools open up; what challenges do they present? These questions motivated our involvement in a collaborative project using text mining tools with medical history sources. In early 2014 we joined University of Manchester colleagues from the National Centre for Text Mining (NaCTeM) to work on a project funded by the UK Arts and Humanities Research Council, under their Digital Transformations theme.[1] As their name suggests, NaCTeM[2] develops text mining tools, mostly for academic use.[3] Our team set out to create a semantic search engine, one that would go beyond finding a specific keyword or string of text in a document. Semantic searches consider the context of use in order to locate terms and their variants representing particular concepts. We wanted to explore how such a search could provide new ways of working with series of medical texts covering a long period of major change in medical knowledge, practice and language. We chose two sources to form our *corpus*, as large-scale collections of structured text are known in digital humanities: the digitised run of the *British Medical Journal* from 1840 onwards, and the more recently digitised London-area Medical Officer of Health Reports that form the Wellcome Library's *London's Pulse* collection.

Text mining (TM) uses digital tools to detect the structure of textual information, then find and recognise patterns and relationships in the structured data.[4] For instance, one TM task is to find and compare the number of instances of particular terms over time in a defined corpus, as Google's N-Gram viewer does using the 'millions and millions of books' that Google has digitised or has access to in digital form.[5] Another common TM application is finding the relative frequency of the words in a text and then visualising these in a way that makes the different frequencies apparent, for example, by size and position in word clouds. Other TM tools track and compare the relative locations of terms and their variants in texts, or, in the case of *topic modelling*, identify groups of terms that tend to be representative of a given topic. As Tom Ewing's use of these approaches demonstrates, they can provide insights that are not readily apparent in traditional reading, however

[1] Our great thanks go to our NaCTeM colleagues and collaborators Sophia Ananiadou, John McNaught and Paul Thompson, and also to Nick Duvall, now at Warwick, who made important contributions to the historical side of the project. We also thank NaCTeM's Riza Batista-Navarro, Jacob Carter, Georgios Kontonatsios and Claudiu Mihăilă.

[2] http://www.nactem.ac.uk.

[3] For instance, the Evidence Finder tool available on Europe PubMed Central (labs.europepmc.org).

[4] A good introductory discussion of text mining for historians is S. Graham, I. Milligan and S. Weingart, *Exploring Big Historical Data: The Historian's Macroscope* (London, 2015). An earlier version of this work is available online at www.themacroscope.org.

[5] books.google.com/ngrams.

intensive and analytical.[6] Collaborating with NaCTeM allowed us to take advantage of even more complex approaches to TM, where systems can be 'taught' to recognise textual data as representing *entities* of different types, such as place names, medical conditions, etc., as well as specific types of relationships between these entities, for example, which symptoms are presented in the text as being caused by a condition. When combined with other tools and approaches used by digital humanities scholars, such as visualisation tools and GIS mapping, TM allows sources to be interrogated in ways that build upon and complement our traditional reading and analysis. Its proponents claim that automated technologies can do this not only faster and more thoroughly with very large data sets, but in ways that reveal new and interesting historical findings. Is this the case with big medical history data?

Before we applied TM tools, we needed to make sure that our digitised corpus was sufficiently correct to be effectively mined, and this was no small task. As Tim Hitchcock has pointed out, many historians do not recognise the extent of OCR errors in the digitised texts that our existing search systems query.[7] The recently created *London's Pulse* is relatively error free, but in the *BMJ* files, which were digitised and OCRed several years ago, up to thirty per cent of the words have errors. Our NaCTeM colleagues devised a customised approach to correcting OCR errors in medical historical texts,[8] which means our system provides a significant improvement on full-text *BMJ* searches.

We then worked with NaCTeM colleagues to analyse sample text, identifying entities and relationships so we could teach our system how to carry out that identification on its own. We began by considering the kinds of entities and relationships historians might want to search for in this corpus. After experimenting with a very large, complex scheme with many subcategories, we decided on a streamlined scheme with seven entity categories (*Anatomical*; *Biological Entity*; *Condition*; *Environmental*; *Sign or Symptom*; *Subject*; and *Therapeutic or Investigational*) and two relationship categories (*Affect* and *Cause*). A team marked up a large sample of text, highlighting where these entities and relationships occurred. We then submitted this sample to a system equipped to 'learn' how to recognise annotations of different types, based on language patterns in the text. The 'trained' system was able to use these learned patterns to recognise entities and relationships in the un-annotated remainder of the corpus – more than 150 years' worth of weekly issues of the *BMJ*, and more than 5000 reports by London-area Medical Officers of Health. Teaching the system to discriminate between the entities historians consider important in historical medical texts proved much more difficult than teaching it to identify simpler entities like named locations. First, terms such as disease names that have been used to describe similar phenomena have changed over time, but using TM techniques the system was able to learn, for instance, that 'infantile paralysis' and 'poliomyelitis' were different terms used in overlapping time periods for a reasonably similar phenomenon. However, some terms have multiple and changing meanings and uses, depending not only on temporal but also textual context, reflecting the very changes in medical thinking we want to examine. One example is the term 'inflammation': as an entity, is it a *Condition*? a *Sign or Symptom*? Or is it a characteristic of a body part and thus *Anatomical*? Any categorisation decision

[6] E. Thomas Ewing, *et al.*, 'Look Out for "La Grippe": Using Digital Humanities Tools to Interpret Information Dissemination during the Russian Flu, 1889–90', *Medical History*, 60 (2016), 129–31.

[7] T. Hitchcock, 'Confronting the Digital, or, How Academic History Writing Lost the Plot', *Cultural and Social History*, 10 (2013), 12–14.

[8] P. Thompson, J. McNaught and S. Ananiadou, 'Customised OCR Correction for Historical Medical Text', *Proceedings of Digital Heritage 2015*, Granada (2015).

we made, and the rules we devised for making sense of the context, would have to be clear enough for the search system to 'learn' and apply. Yet that decision would still need to reflect the term's changing and indeterminate use in a way that would satisfy historian users of the search system. These tricky and problematic decisions have been built into our system, and we are of course anxious that our users be aware that, thanks to such decisions and to the complexity of the overall task, they need to be critical as they approach the results our system gives. In fact, the system facilitates critical engagement by the ease and speed of making alternative and cross-checking interrogations.

As we trial our 'beta-version' with our Advisory Group, we are excited by the possibilities that this system, and that TM and indeed digital humanities tools as a whole, open up. First, our system speeds up searches dramatically, and allows more focused searches than would be possible even with fairly sophisticated Boolean searching. By searching for *Condition*: 'tuberculosis', for example, the user gets results where the system has recognised the term as referring to tuberculosis as a condition, rather than finding every instance of the word 'tuberculosis' in the text (in phrases like 'National Tuberculosis Association', or 'tuberculosis nurse'). But semantic searching is about much more than convenience. The user can find all instances of a particular entity category: one can, for example, locate all articles published in 1892 where a *Biological Entity* (including non-human animals and microorganisms) is mentioned, and find the frequency with which each *Biological Entity* is mentioned. Combining entity searches and relationship searches enables the user to find instances where one entity is said to cause another: by asking what *Condition* entities are said to *cause* the entity *Sign or Symptom*: 'swelling' in the entity *Anatomical*: 'feet', the user can find case reports and reviews that discuss which ailments were understood to cause the feet to swell. (By contrast, consider the overwhelming flood of results the searcher would get by searching for the terms 'feet' and 'swelling'.) This capacity is particularly useful for those who want to investigate relatively common, everyday phenomena that would stymie the best intentions of researchers because they are difficult to find in text, too numerous to manage easily, or easily overlooked by the all-too human researchers. We thus expect this tool not only to speed up searching and make it more precise, but also to help us see things that would otherwise be too difficult to see or too easy to miss, or that we might not even have known we were looking for. It will never provide easy and obvious answers to big questions, and it requires that the user know something about how it works. Nevertheless, we hope that as a tool that can facilitate exploration and new ways of encountering existing resources, it will be valuable both as a resource in its own right, and as a means of introducing our colleagues to TM tools and some of the possibilities of digital humanities.

Elizabeth Toon, Carsten Timmermann and Michael Worboys
University of Manchester, UK

doi:10.1017/mdh.2016.19

Reflections

'Grand challenges' abound in every discipline.[1] The grandest challenges are interdisciplinary. They hold the potential to change theory, practice, and the very shape of research,

[1] American Academy of Arts & Sciences, Commission on the Humanities and Social Sciences, The Heart of the Matter: The Humanities and Social Sciences for a Vibrant, Competitive, and Secure Nation, available online at http://www.humanitiescommission.org/_pdf/hss_report.pdf (accessed June 23, 2015).