Med. Hist. (2016), *vol.* 60(1), *pp.* 126–131.
doi:10.1017/mdh.2015.82

Media Reviews

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

Introduction

Digital worlds are producing ever-increasing amounts of information across databases and such born-digital resources as blogs, websites, social media and digitised physical materials. Such 'big data' joins a longstanding world that is deeply rich with a variety of persistent material objects that contain records of the human condition and the human past. As these analogue and virtual worlds collide and co-exist, opportunities abound for scholars to advance interdisciplinary collaborations and expand co-operation throughout institutions and organisations that preserve history and support historical research.

On 1 May 2015, in a panel at the annual meeting of the American Association for the History of Medicine, scholars from the United Kingdom and the United States, as well as leaders of the Wellcome Trust and the National Library of Medicine's History of Medicine Division, addressed key philosophical and practical issues impacting the application of digital humanities techniques to the history of medicine. Offering perspectives practised by institutions that are producing digitised and born-digital resources, and from individuals who are using them, this panel engaged audiences associated with both enterprises and challenged them with a wide and sustained reflection on the processes of digitisation and the meaning of 'big data' for the future of the medical humanities.

Jeffrey S. Reznick
History of Medicine Division, U.S. National Library of Medicine,
National Institutes of Health, USA

doi:10.1017/mdh.2015.83

Why Creating a Digital Library for the History of Medicine is Harder than You'd Think!

It has never been easier for a historian of medicine to access medical books and journals. There are very few journals nowadays, especially in medicine and bioscience, that rely solely on paper as a medium for publishing. With books, the transition to digital has been slower but is gathering pace. Digitisation has also opened up access to historical collections. Of course, this process is far from complete, despite the stated ambitions of organisations such as Google, and this in turn creates challenges for the generation of students who cannot conceive of texts that exist only in a physical world (especially one as remote and alien as a research library). Nonetheless, the rapid growth of digital access does present huge opportunities for the researcher. These range from the simple – the ability to find or check a reference online, or to run a free text search across large bodies of material – to the more complex, such as image matching or pattern recognition, the extraction

The Intramural Research Program of the U.S. National Library of Medicine, National Institutes of Health, supported the research and writing of this introduction, and the editing of its associated articles.

of numerical data or the use of semantic mapping technologies to recognise patterns of language rather than just words (processes collectively referred to as content mining).

Libraries have not allowed the transition to digital to pass them by, but it has changed the relationship between institutions and audiences, eroding the status of the library as privileged point of access to historical collections. A significant (and growing) slice of many libraries' nineteenth-century printed book holdings are duplicated in Google books,[1] or in the Hathi Trust[2] or in the Internet Archive.[3] The digital back-files of many journals, along with other digital resources, are now hosted by publishers, and the content simply licensed to libraries. Most university librarians would admit that the need to manage access to online content while still maintaining print collections and physical spaces has placed library budgets under additional stress. Nevertheless, libraries remain necessary and valued components of the research ecosystem. Moreover content mining encourages a different way of thinking about the ways in which libraries holding historical medical collections might work in future.

There are notable exceptions to the idea that libraries are becoming marginal to the digital information environment and the creation of resources for content mining. For example the US National Library of Medicine has overseen the evolution of the *Index Medicus* from print to the database *Medline*, to its online interface, *PubMed*, and to its full-text subsets, *PubMed Central* and *Europe PubMed Central*, both which have become major resources for content mining.[4] In France, BiuSanté, the combined medical and pharmaceutical libraries of the Universities of Paris Descartes, Paris Diderot and Paris Sud, hosts *Medic@*, with over 122 000 digitised items;[5] while in London the Wellcome Library has embarked on its own digitisation programme, with the aim of adding fifty million pages of books, journals, archives and manuscript material to its digital repository by 2020.[6] Wellcome has also contributed to the Medical Heritage Library (MHL),[7] a major collaborative endeavour on the part of libraries in the USA, Canada and now the UK to create a comprehensive digital collection of books and journals. But while it has stemmed from the activities of medical libraries, the MHL has only been realised through partnership with the Internet Archive, a wholly virtual repository that now encompasses many millions of books, audio-visual files and pieces of software and hundreds of billions of web pages – digital hosting on a scale that even the largest traditional physical libraries would struggle to match. As historians of medicine and librarians grapple with the implications of the digital turn, issues of scale and access loom ever larger.

In fact, in some ways the sheer proliferation of digital content, the so-called 'literature deluge', has a created challenge that can only be met through the application of content mining against otherwise brute force labour. But, while mining might depend on being able to access content on a large scale, and without regard to the physical location of the original item, there remains a need to be able to define relevant content. *Medline* (the database that underpins *PubMed*) does this for biomedicine, defining – primarily through the selection of journal titles, rather than individual articles – the content that 'counts'. When it comes to older material, however, the limitations of retrospective application of

[1] http://books.google.com.
[2] http://www.hathitrust.org.
[3] http://archive.org.
[4] http://www.nlm.nih.gov.
[5] http://www.biusante.parisdescartes.fr.
[6] http://wellcomelibrary.org.
[7] http://www.medicalheritage.org.

https://doi.org/10.1017/mdh.2015.83 Published online by Cambridge University Press

current categorisations are self-evident. Instead the definition of content sets is an integral part of the research process, demanding the same kind of diligent attention to sources which has traditionally underpinned historiographical practice.

For example, historian Colin Jones has posited that smiling became an increasingly common mode of emotional and social behaviour over the course of the eighteenth century in France, a significant cultural shift connected to the application of techniques in cosmetic dentistry.[8] Jones' argument is based on many sources, including analysis of portraiture and deep reading, but it is substantiated by a fairly simple yet effective piece of text-mining which looks at the frequency of use of the word *sourire* – smile – in French literature from the relevant period, and the adjectives or adverbs applied to it. Jones's analysis was based in part on the ARTFL-FRANTEXT corpus,[9] a digital collection of canonical literary texts. In this instance, the value of the analysis is strengthened by the fact that it draws upon a select collection of works, allowing a conclusion to be drawn about the cultural significance of the changing semantic patterns revealed by content mining.

The ability to apply boundaries to sets of digital content does not depend on the duplication of content. The Medical Heritage Library is built upon the Internet Archive's much larger collection of digitised texts. It contains over 91 000 individual works but none are identified exclusively as being part of the MHL. Rather each item belongs to a number of non-exclusive and overlapping sets – books digitised from a particular library, from a region or as part of a particular project. Some are items that originated from general library collections, and have been subsequently identified and tagged as being 'medical'.

This pattern raises the obvious question of what 'medical' might mean as a category. There is no straightforward answer. Nor should there be. Indeed, such blunt forms of categorisation would probably have little value in a traditional model of scholarship in which the historian's reasoned judgement prevails. But for the application of content mining it is both necessary and useful to be able to define boundaries for large content sets. There are of course lists which can be taken to indicate specific historically located definitions of what might be considered canonical, such as *Morton's Medical Bibliography*, which has since been updated by Jeremy Norman and is now online.[10] Morton's list was itself derived from Fielding Garrison's earlier 'Texts Illustrating the History of Medicine', published in the *Index Catalogue of the Surgeon-General's Office* in 1912, and the latter represents another, rather more expansive, set, or rather, given that the *Index Catalogue* was published periodically, a number of overlapping sets.

The *Index Catalogue's* nearest modern equivalent would be *LocatorPlus*, the current catalogue of the National Library of Medicine – probably as comprehensive a definition of what is, or has been, counted as 'medical'. But comprehensiveness is not the only quality of value to the historian. What about using the catalogues of the Royal College of Surgeons of England, the College of Physicians of Philadelphia or of the Medical Society of London as proxies in the present for what certain communities of practitioners might have considered as falling within their field of interest in the past? Reconceptualising such catalogues not just as inventories of specific repositories, but as filters to be applied to increasingly universal digital collections (universal in scale as well as scope) may help historians make more effective use of content mining in the future.

Both content mining and the application of catalogues as filters for content sets raise further questions for libraries' management of data. Unlike museum catalogues,

[8] Colin Jones, *The Smile Revolution in Eighteenth Century Paris* (Oxford: Oxford University Press, 2014).
[9] http://artfl-project.uchicago.edu.
[10] http://historyofmedicine.com.

https://doi.org/10.1017/mdh.2015.83 Published online by Cambridge University Press

library catalogues have not commonly been used to record historical information about provenance, especially date of acquisition, which would be of particular value for the retrospective construction of content sets. Mining also demands machine-readable text, which is problematic with certain typefaces for printed material and near to impossible with manuscript. Mining of tabular data such as that included in the published reports of the London Medical Officers of Health, recently digitised by the Wellcome Library,[11] is only possible because the tables themselves have been separately re-keyed and presented in appropriate formats. For historians interested in large-scale analysis of images, there is also the need to separate illustrations from text, while retaining some sense of the original context of the image.

None of these are insurmountable obstacles. Librarians have traditionally managed data about the items they hold as adeptly as they have cared for the physical objects: the transition to digital content sets and to the application of content mining simply requires that these skills be applied a little differently, and without preciousness about the correspondence between physical holdings and virtual repository. Building a digital library for the history of medicine may be hard, but then again being a librarian has never been easy either!

Simon Chaplin
Wellcome Trust, UK

doi:10.1017/mdh.2015.84

Look Out for 'La Grippe': Using Digital Humanities Tools to Interpret Information Dissemination during the Russian Flu, 1889–90

On 28 December 1889, and at the height of global anxiety about a spreading epidemic, the American journal *Medical News* published a lengthy article by Dr Roberts Bartholow about 'The Causes and Treatment of Influenza'.[1] Noting that the 'reappearance of influenza in one of its cyclical manifestations, or epidemics, is an interesting event', Bartholow offered a sweeping statement about the impact of the disease:

> Influenza comes suddenly; goes as quickly. The least robust, at any age, and women seem to be the first victims. It is here a question of bodily condition, not of the sex. The large numbers simultaneously attacked attracts general attention, and thus those most impressionable are seized, the onset being facilitated by any depressing emotion like fear or illness.

To treat influenza, Bartholow recommended cures such as sulphurous acid, iodoform, tannin, resorcin, chinoidin, calomel, antipyrin, acetanilide, phenacetin, and more.

This article resembled many contemporary reports about an epidemic already referred to in late 1889 as 'Russian influenza' that combined specific descriptions of symptoms with prognostication about the course of disease. Bartholow's recommended treatments were clearly intended for doctors and druggists rather than the general public, yet his sage

[11] http://wellcomelibrary.org/londons-pulse.

Research funding was provided by the Virginia Tech Department of History. For comments on earlier versions of this project, the authors wish to thank Madhav Marathe, Stephen Eubank, Samarth Swarup, Meredith Wilson, Ed Fox, Aditya Prakash, Bryan Lewis, and Daniel Sullivan, all from the Virginia Bioinformatics Institute at Virginia Tech, and Jeffrey S. Reznick, from the US National Library of Medicine.

[1] Roberts Bartholow, 'The Causes and Treatment of Influenza', *Medical News*, 55, 26 (1889), 710–4.