Exploring Big Historical Data: The Historian's Macroscope

Review Number: 1889  
Publish date: Thursday, 11 February, 2016  
Author: Shawn Graham  
Ian Milligan  
Scott Weingart  
ISBN: 9781783266371  
Date of Publication: 2015  
Price: £65.00  
Pages: 282pp.  
Publisher: Imperial College Press  
Publisher url: http://www.worldscientific.com/worldscibooks/10.1142/p981  
Place of Publication: London  
Reviewer: Adam Crymble

For the past decade, digital history students have really only had one book upon which to draw to introduce them to the field: Dan Cohen and Roy Rosenzweig’s 2005 Digital History: A Guide to Gathering, Preserving, and Presenting the Past on the Web. The book continues to appear on nearly every ‘digital history’ syllabus in the English-speaking world. Despite the clear value of this work for a generation of scholars (myself included) the world of digital history has moved on. No more does ‘digital history’ (DH) mean ‘putting stuff online’. Instead, a decade later, we’re interested in the potential of historical data and what it can tell us about the past.

Within history, in many respects, this emphasis on data is a continuation of the pursuits of economic and demographic history that had roots in the 1960s and 1970s with mainframe computers. The current digital methodology has been approached with new enthusiasm by a new generation of scholars who don’t always recognise their intellectual connections to the statisticians of the history world. This might be because the emphasis has frequently shifted away from the census or the traditional stalwart sources of the quantifiers, and instead these new scholars have begun to apply statistical approaches to newspapers, web archives, and literary texts.

Scholars in the field of literary studies have already begun publishing in earnest on this new data-centric view of the humanities that goes beyond merely publishing online. One of the best examples comes from Matthew Jockers in his book Macroanalysis: Digital Methods & Literary History, which attempts to demonstrate how digital approaches to literature studies can change the types of questions and answers available to scholars.

Until Exploring Big Historical Data, historians had no such bridge between the Web 1.0 era of Cohen and Rosenzweig and the data-centric view of today. Exploring Big Historical Data is perhaps the first printed book, targeted specifically at historians, to fill this niche in the new wave of digital history, and will play an important part in the digital history classrooms of the next decade, including my own.

I say first ‘printed book’ because, since 2008 at least, scholarly blogging by digital historians and digital
humanities scholars more broadly has been filling the hole while we awaited a text such as this. The authors draw heavily on academic blogs in their footnotes, as well as digital scholarly initiatives such as Digital Humanities Now and The Journal of Digital Humanities, both out of the American DH community. This acceptance of blogging and non-peer reviewed or non-traditional contributions to the field is as yet unusual in historical monographs and signals, for this reviewer, a welcome change of practice that formalises the belief in DH that a good idea doesn’t need to be vetted by a competitor to be a good idea.

The book also relies very heavily on the intellectual foundations of the open access digital history textbook, The Programming Historian – perhaps too much so. The Programming Historian (3) offers technical tutorials aimed at historians looking to learn specific digital skills to aid in their research processes. Both myself and Milligan are editors of the project and all three authors of this book contributed a popular lesson on topic modeling in 2012, which is the basis of chapter four: ‘Topic modeling: a hands-on adventure in big data’.

While flattering that the authors hold The Programming Historian in such high esteem, I did notice more than a few examples plucked directly from its digital pages. These included simple things, such as using the same Old Bailey Online trial of Benjamin Bowsey (p. 68) in their own examples (originally suggested to me by Tim Hitchcock and used extensively in Programming Historian tutorials), despite having more than 197,000 trials to choose from. I would suggest that this was a missed opportunity to highlight the breadth and diversity of digitised historical resources.

I also have concerns about the combination of excellent background information on digital techniques, such as textual and network analysis, and technical tutorials that rely heavily on specific versions of specific software. While the authors noted in their introduction that they had done their best to future-proof the book to technological changes, they repeatedly make software suggestions: Notepad++ or TextWrangler (p. 92), offer reviews of specific software packages: UCINET, Pajek, Network Workbench, Sci2, NodeXL, Gephi (p. 237–239), and even offer a suggestion to manage a conflict between Mac OS X Mavericks' operating system and a Java compatibility issue that causes problems with Gephi (p. 253). While helpful for the reader working in 2016, I am doubtful that the student of 2026 will find such suggestions useful. Given the fact that the authors had to remove from the book an example using the tool Paper Machines because an update to the software on their computers had broken the tutorial in the time it took to write the book (p. 156), they should have been more cautious in their selection of specific examples. A greater focus on core principles would have been helpful in future-proofing the text, while the tutorial-style elements of the book would have perhaps fit better in a digital resource, such as the Programming Historian, which can be continually updated as needed.

The work was written and peer reviewed openly online, and the pre-print draft will still be available openly (http://www.themacrooscope.org/?page_id=584 [2]) as long as the publisher remains satisfied that its access does not cannibalise sales of the volume (p. xix). The online version is considerably less user-friendly than the printed book, which demonstrates the added value of the publisher, and demonstrates the adage that you get what you pay for. This is, I think, a fair compromise between a group of authors intent on promoting open access and a publisher with costs to cover and profits to make. This open access approach also follows in the footsteps of Cohen and Rosenzweig’s volume, which has always been available openly as well as in print. We can only hope that the press does indeed keep the open version available, but if they don’t, the authors have promised (threatened?) to archive it in the Internet Archive.

The real strengths of this book are in their straightforward explanations of what can at first seem like fairly daunting topics, as well as the excellent background information on the development of the field of digital history. Chapters one and two give readers an excellent understanding of ‘big data’ from the perspective of a historian – as opposed to a physicist or biologist – as well as a broad set of skills that will help the next generation of scholars avoid common problems. These include being critical of digital tools, and not blindly trusting search algorithms or black box software without first understanding how they work and how they affect the conclusions a researcher will draw through using them.
The authors include a range of good examples that provide readers with the skills to be critical of digital tools and sources. For example, readers are informed about the trouble with the ‘medial S’ (?) and its effects on search results. They are also challenged to think about one-click tools such as the Google N-Gram Viewer, which allows users to plot trends in language use over time. For that tool, the authors highlight the importance of thinking about what is or is not in the corpus upon which the analysis is based. This critical-first approach to digital tools is a valuable lesson for the target audience, and makes the book useful for anyone designing a syllabus.

In chapter two, ‘The DH moment’, the authors highlight important copyright implications for historians working with big data, before putting forth a convincing case that we are all already digital historians, suggesting that many ‘traditionalists’ may be turning a blind eye to this fact. Throughout the book the authors rightly argued that a digital resource is fundamentally different from its paper predecessor. As they note, the current practice of most historians of pretending that they consulted the paper version of a source damages our ability to test the conclusions of their research because the digitisation process is not objective and involves decisions by the digitiser that influence what information is preserved, transferred, or lost. For us to be critically engaged with the historical scholarship of the future, we must know, for example, if a historian read through newspaper content systematically on microfilm, or dived in through keyword searches that relied upon questionable transcriptions. If hidden by dishonest footnoting, our field’s very scholarly integrity comes under threat. With this in mind, I was pleased to see that the authors practiced what they preached. Uniform Resource Locators (URLs) have been used extensively in the footnotes, acknowledging the digital nature of the intellectual contributions of others.

Chapter three, ‘Text mining tools’, represents a shift towards practical matters, first introducing readers to the strengths and weaknesses of word clouds as a way of distant reading a text or corpus, before taking them beyond and into concepts such as concordance or keywords in context. The chapter includes a useful overview of regular expressions, including some examples as to how to use them to solve common problems. For historians who spend most of their time close reading individual documents, the chapter does little to convince the reader of the need for regular expressions, but for those who find themselves cleaning or sorting historical data, the uses are obvious. This unobtrusive approach is probably the right one, allowing readers to decide for themselves if there is value in a tool, instead of clubbing them over the head with it, and it’s an approach the authors have taken throughout.

Chapter four, ‘Topic modeling’, introduces topic modeling as a means of discerning what is in a series of texts – letting bodies of texts tell us what is in them, rather than a historian approaching a corpus looking for specific documents that match a research question. The authors astutely point out that ‘the topic model generates hypotheses, new perspectives, and new questions: not simple answers’ (p. 157), which is a common conclusion about a variety of digital methods, including those highlighted in this book. The chapter includes an amazing ‘topic modeling by hand’ exercise that the authors have used previously in class. The exercise is so good that readers are left wanting many more of these throughout the book, as a way of delineating the technical task from the core skills behind it.

Chapter five, ‘Making your data legible’, again includes a core skill, outlining basic principles of visualisation and explaining a number of different graph types, rather than focusing on specific software for implementing each. This emphasis on core skills is valuable, and data visualisation literacy levels amongst historians are very low at the moment, so the principles behind this chapter are well considered. Unfortunately, in the printed book, the visualisations themselves have been printed in grayscale, despite being designed for colour viewing. The text frequently refers to subtle colour variations that are invisible to a reader. In general, the visualisations in the book are not of sufficient quality and many are ambiguous. This is an irony, given the fact that the book itself contains a chapter on the importance of getting it right.

These ambiguous figures include:
Many of these could have been fixed with a few hours in an image-editing program. Figure 5.28 is especially bad. It is a well-known image that demonstrates the difference between hue, value, and saturation as they relate to the colour wheel. However the famous image is utterly useless in a dark grayscale, which undermines the message entirely. These figures are available in colour on the book’s website for free, but that hardly makes for a fluid reading experience.

The final two chapters (six, ‘Network analysis’, and seven, ‘Networks in practice’) shift to an introduction to the concepts behind network analysis and then to a series of examples of the same. As is the case throughout the book, the core skills highlighted in chapter six are more useful than the software-dependent tutorials of chapter seven. For anyone considering work on network analysis, chapter six is well worth reading and keeping to hand, as it includes a number of definitions and clear examples for the uninitiated. Some of the examples used are hypothetical, including a mythical trade network between ‘Netland’, ‘Connectia’, ‘Graphville’, and ‘Nodopolis’ (p. 240), which is disappointing given the opportunities to highlight an actual historical use case of the approach.

Despite a few rough patches, the book certainly achieves its aim, providing a scholar who is new to the field with a number of ways to understand new approaches to textual analysis. It is a gentle introduction and accessibly written throughout. Used as recommended by the authors, in conjunction with Cohen and Rosenzweig’s Digital History, and with the Programming Historian, it would certainly make a valuable addition to any course reading list. Time will tell how technically sustainable its example problems are. I have my doubts, but I remain certain that the fully explained concepts that form the bulk of the book will remain valuable for the next generation of digital history scholars.

The book is not a ‘history’. There are few historiographical arguments that undergo scrutiny herein, which is a shame because it will do little to win over ‘traditional’ historians. However, as that is not its purpose, it really cannot be counted against it. Instead, this is a textbook of digital methodology aimed at ‘an advanced undergraduate looking for guidance as they encounter big data for the first time’ (p. xx). The paperback version is fairly priced for such use at £26, and the book would be equally valuable for a postgraduate or academic looking to get up to speed on several approaches to textual analysis – for there is rightly little attempt here to engage with materials beyond text, for fear of overextending the scope of the book. I would heartily recommend it for anyone designing a course on digital textual analysis, and students can expect a solid background in a range of useful concepts, including distant reading, topic modeling, and network analysis, as well as data visualisation.

Notes

Source URL: https://reviews.history.ac.uk/review/1889

Links
[1] https://reviews.history.ac.uk/item/151123