DIGITAL HUMANITIES

doi: 10.3389/fdigh.2014.00001

Challenges and opportunities for digital history

Ian Gregory*

Lancaster University, Lancaster, UK

**Correspondence: i.gregory@lancaster.ac.uk*

Edited and reviewed by:

Robert C. H. Sweeny, Memorial University of Newfoundland, Canada

Keywords: digital history, digital humanities, digitization, digital methods, digital sources

The challenge for digital historians is deceptively simple: it is to do good history that combines the computer's ability to search and summarize, with the researcher's ability to interpret and argue. This involves both developing an understanding of how to use digital sources appropriately, and more importantly, using digital sources and methods to deliver new scholarship that enhances our understanding of the past. There are plenty of sources available; the challenge is to make use of them to deliver on their potential.

There have been false dawns for digital history, or "history and computing," in the past (Boonstra et al. 2004). Until very recently, computers were primarily associated with performing calculations on numbers. This has resulted in them becoming fundamental tools in fields such as economic history, historical demography and, through the use of geographical information systems (GIS)[1], historical geography. These are, however, relatively small fields within the discipline as a whole and much of the work that has been done in them has taken place outside of History departments in, for example, Economics, Sociology, and Geography. As most historians work with texts, it is hardly surprising that this style of computing has made little impact on the wider discipline. Within the last few years, however, there has been a fundamental shift in computing in which, put simply, computers have moved from being number crunching machines to become an information technology where much of the information that they contain is in textual form. This has been associated with the creation of truly massive amounts of digital textual content. This ranges from social media and the internet, to private sector digitization projects such as Google Books and the Gale/Cengage collections, to the more limited investment from the academic and charitable sectors (Thomas and Johnson 2013). Thus, computers are now inextricably concerned with texts – exactly the type of source that is central to the study of history.

As a consequence, many historians have become "digital historians" almost without realizing it through making use of the vast number of sources that are now available from their desktop. So is everything in the garden that is digital history currently rosy? The answer, judging by work such as Hitchcock (2013) and the responses to it (Knights 2013; Prescott 2013), seems to be a resounding no. Many criticisms are centered on the digital sources themselves, whose quality is lower than that might be hoped. Digitizing a document is usually a two-stage process: first a digital image of the document is created as a bitmap, then the textual content is encoded as machine readable text. The two are then often brought together such that a user can type a search term, this is located in the text, and then the user can be shown the appropriate image of the page. The first of the two stages is relatively simple using a scanner or camera and, if done properly, only results in relatively minor abstractions from the original as the result is a facsimile copy. The second stage, however, is hugely problematic involving either the text being manually typed, or optical character recognition (OCR) software being used to automatically identify letters from the bitmap image. Both of these are slow, expensive, and error-prone. OCR tends to be used on large-scale projects: it is faster and cheaper but tends to result in far more errors. Whatever approach is used, checking the results is very difficult. Common approaches involve carefully typing up "gold standard" samples of parts of the source and comparing these with bulk-entered material to give a percentage of words or letters that have errors. Understanding what the consequences of these scores mean in practice is difficult. Even without error, if the text is removed from the page scans then they are heavily abstracted from the original and much potentially useful information is lost.

Once created, digital sources are often interrogated using techniques that are not properly understood but are nevertheless used uncritically. The classic example that combines both the data capture and uncritical use problems is typing a keyword search into a web interface, which returns a list of hits sorted by "relevance." As Hitchcock (2013) points out, most historians using digital sources do this without having any idea of the implications either of the data capture that created the digital copy of the source, and thus whether the search will miss words as a result of spelling variations derived from digitization errors, or of how the search engines decides what is – and, more importantly, is not – "relevant." While using search engines may be problematic, in reality they are the only digital tool that most historians use, indeed there is a lack of widely used techniques that can be used to interrogate, summarize, and understand the large volumes of material that are available.

So what do digital historians need to do? The answer, I would argue, is to remember that they are first and foremost historians and that historians fundamentally are in the business of taking complex, incomplete sources that are full of biases and errors, and interpreting them critically to develop an argument that answers a research question. Digital sources do not change this;

[1]The 'maps' (spatial data) used by GIS are created from coordinates and are thus a specialised form of numeric data.

however, they do increase both the opportunity and the complexity. As identified above, there are complexities centered on the implications of digitized sources and on the tools that we use to interrogate and analyze them. The opportunities are enormous centering on the ability to search, visualize, and analyze historical sources be they very large or much more modest.

Thus, doing good digital history requires a number of things: firstly, historians need to critically evaluate digital sources (whether born digital or digitized) in much the same way as they critically evaluated other sources and consider these implications in their arguments. At the moment, this is more difficult than it should be because debates around the benefits and problems of digital sources have generated much heat but little light. A more nuanced understanding of these implications is required. Secondly, there is a pressing need (and opportunity) to develop and understand new techniques to deal with the errors in these sources. One possible approach is to automatically correct errors [see, for example, Evershed and Fitch (2014)], although this will inevitably introduce new errors and adds further abstraction from the original source. An alternative is to conduct studies that help us to understand the implications of the digitizing errors.

Thirdly, there is the need (and opportunity) to develop and use methods to exploit digital sources. Close reading will always be an important component of the historian's toolbox and simple keyword searches provide a way of assisting this with very large volumes of material by allowing the passages worthy of close reading to be quickly identified; however, other approaches are also required. This presents challenges. When large quantitative databases came available, statistical techniques were available to help analyze them; however, there is no obvious equivalent for large textual sources. There are, however, a wide range of promising opportunities as diverse as: corpus linguistics, distant reading, network analysis, GIS, and so on. This presents the discipline with an opportunity – historians have always been concerned with the analysis and understanding of texts, if they can develop techniques to help with this then these should be much more broadly applicable in a world that is increasingly awash with digital texts.

Fourthly, while work on sources and methodologies is important, it is a means to an end rather than an end in itself. The work that will ultimately prove the relevance and importance of digital resources and methods will not stress the digital, it will stress the applied and make a contribution to knowledge on particular topics within history that "non-digital" historians will be interested in. This is not easy as it means effectively handling the interpretive challenges faced by traditional historians, as well as the technical and interpretive challenges presented by the digital. It also means that the discipline as a whole needs to be better at, and more receptive to, work that stresses methodological developments that help us better understand digital sources. Ultimately though, the combination of the computer's ability to manipulate and summarize large volumes of material and the human brain's ability to interpret this appropriately will provide major advances in our understanding of the past. The opportunities for those who are skilled enough, adventurous enough, and imaginative enough to do this are enormous.

ACKNOWLEDGMENTS

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

Boonstra, O., Breure, L., and Doorn, P. 2004. *Past, present and future of historical information science.* Amsterdam: NIWI-KNAW.

Evershed, J., and Fitch, K. 2014. Correcting noisy OCR: Context beats confusion. In *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage 2014*, 45–51. doi:10.1145/2595188.2595200

Hitchcock, T. 2013. Confronting the digital or how academic history writing lost the plot. *Culture and Social History* 10:9–23. doi:10.2752/147800413X13515292098070

Knights, M. 2013. The implications of social media. *Culture and Social History* 11:329–33. doi:10.2752/147800414X13983595303156

Prescott, A. 2013. I'd rather be a librarian: a response to Tim Hitchcock, 'Confronting the digital'. *Culture and Social History* 11:335–41. doi:10.2752/147800414X13983595303192

Thomas, D., and Johnson, V. 2013. New universes or black holes? Does digital change anything? In *History in the digital age.* Edited by T. Weller, 173–94. Routledge: Abingdon.

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 29 October 2014; accepted: 03 December 2014; published online: 17 December 2014.

*Citation: Gregory I (2014) Challenges and opportunities for digital history. Front. Digit. Humanit. **1**:1. doi:10.3389/fdigh.2014.00001*

This article was submitted to Digital History, a section of the journal Frontiers in Digital Humanities.