DIGITAL CONTEMPORARY HISTORY

SOURCES, TOOLS, METHODS, ISSUES[1]

PETER WEBSTER

I was for several years a founding convener of a seminar in digital history at the Institute of Historical Research in London. The seminar was intended to showcase new projects but also to hear substantive conclusions on historical matters reached using digital methods. As the convener with research interests after 1900, it fell to me to identify and invite speakers for that period. And we heard fascinating papers: on text mining for the history of medicine in the UK; on mapping Harlem in New York, USA, in the 1920s; on using Twitter to track language change. However, it was necessary to range across continents to find those speakers; in the UK at least, there are many more scholars working digitally on the 18th and 19th centuries than on the 20th. In this article I want to explore why this might be, why the situation might be changing (and changing soon), and some shifts in method and approach, which that changed situation is likely to entail.

Part of the difficulty is the sheer weight of source material with which contemporary historians must deal. The eminent historian of English religion Christopher Hill was reputed, even by his critics, to have known more about the whole 17th century than any other scholar, based on a prodigious reading of printed sources. By contrast, a historian of contemporary religion is faced with a challenge on an altogether different scale. Data from the British Library shows that in the five years to 1954 there were nearly 5,000 titles published in the UK relating to religion and theology; between 2000 and 2004 the number was 13,714.[2] The situation is similar when we consider unpublished sources. My own 2015 book on Michael Ramsey, archbishop of Canterbury and leader of the established church in the UK in the 1960s, is part of a series on the holders of the same office. Three of his predecessors in the 12th century are dealt with in a single volume covering half a century, so little is there that can be known about them.[3] Ramsey's 13 years in office, in contrast, generated an archive of some 338 volumes, each containing

1 An earlier version of this paper was given as a lecture to the Danish Association for Research in Contemporary History in Copenhagen, 29 January 2016. I am grateful to Nils Arne Sørensen for the invitation, and to the audience for their attention and questions. I am also indebted to Professor Sørensen, Ian Milligan, Jane Winters and the anonymous peer reviewer for their comments on a draft version.

2 A subset of the British National Bibliography for theology monographs and serials, available from http://www.bl.uk/bibliographic/download.html, (24.03.2017) updated annually.

3 Truax, *Archbishops*, passim; Webster, *Archbishop Ramsey*, passim.

300 numbered folios – perhaps 100,000 individual pieces of paper to select from, quite apart from his considerable published work and the traces left in newspaper reports and the like. The historian of the 20th century faces a vast volume of sources.

This sheer volume has the effect that vast swathes of the archival record for the 20th century are untouched by the hands of any researcher. My PhD thesis (completed in 2001) was on a topic in the English 17th century, but by no means one of the most hotly debated ones. Even for the Church of England in the 1630s, a community of scholars could come to know the same sources as each other, and well. Working as I now do on the 1960s, one might go from year to year without ever using an unpublished source that any scholar had previously seen.

Of course, we knew this before the advent of the digital, and it might be argued that this sheer bulk is a spur to the use of digital approaches. Given so much to read, are techniques such as text mining not the perfect solution to the problem? By and large, 20th century materials are also often technically easier to digitise than those from the medieval period, and tools are available to make this possible for an individual researcher. However, my point is that the sheer volume and diversity of the sources makes it harder to achieve the kind of breakthrough in large-scale digitisation that is needed to take the whole profession on to a new level. An early pioneer in historical digitisation in the UK suggested that "early modern and 18th century British history may well be the most digitized place and time on the planet".[4] It would be inconceivable to say the same of any aspect of the 20th century, and this is in large part due to the difficulties of knowing what should be digitised. To illustrate: I imagine that few scholars are able to describe themselves simply as a 'twentieth century historian'. I often describe myself as a historian of 20th century British religion; in reality, my particular specialism is the 1960s, and in England (rather than Scotland or Wales), and of religious ideas (rather than practice), and at a national rather than a local level. The result of the specialisation which is forced upon us by the volume of sources is that, if asked which 10 primary sources we would most like to see digitised, there would be few that appeared on more than one person's list. But digitisation is expensive, and to persuade research funders to pay, we need always to show that there is a large community of scholars who stand to benefit.

Again, an illustration from an earlier period: the *Acts and Monuments* by John Foxe, usually known as Foxe's Book of Martyrs. After the Bible, the Book of Martyrs was possibly the single most important religious text in English in the 16th century, and was still read by Protestants well into the 19th and indeed 20th centuries. Unsurprisingly, Foxe was the subject of one of the earliest research grants in the UK for digitisation, and the project ran from 1992 to 2011 with successive

4 Shoemaker, *Digital transformations*.

tranches of funding, producing first a CD-ROM edition and then an online version.[5] Although it cannot be proven, it is likely that the reason why there are relatively few publicly funded digitisation projects for the last 100 years is that there are few Foxes; few sources the value of which it would be so easy to agree upon.[6] To be clear, there are of course many digitised sources for the 20th century, but few indeed that have been funded by public money in conjunction with public libraries and/or scholars.

The second major brake on the digitisation of sources for the 20th century is copyright. In Denmark, copyright persists in images for 50 years after their creation, in books for 70 years after the death of the author, a regime not dissimilar to that in the UK. So digitisation efforts are either in the hands of the owners of the copyright, or some other party with the funding to obtain licences for that material. That a 'black hole' between perhaps 1940 and the millennium has been the result is evident from the holdings of the pan-European Europeana service.[7]

The digitisation of large single resources then tends to fall into three categories other than those made possible by public funding (as I leave aside those resources created to meet demand from genealogists outside the historical profession, although historians have cause to be grateful for them as we should also be to the same genealogists who help local archives justify their public funding). Some resources are funded philanthropically with free access, although rather few; one such is the Margaret Thatcher Archive (http://www.margaretthatcher.org/). Others attract public funding for political reasons, such as in the case of resources like Hansard, the record of proceedings in the Houses of Parliament of the UK. These are in line with a more general impetus among governments towards openness for official documents. Commercial digitisation in the UK has often concentrated on newspapers: large in scale, covering long time periods and of general use to many people. In the UK, the Times Digital Archive runs from 1785 to 2009, a product of Gale Cengage Learning. In all of these, the kinds of services that have resulted tend to serve only the most generic use and thus to have only limited functionality. In particular, access to the raw data is highly unusual, ruling out more advanced use by scholars with the skills to handle data without resort to the online user interface. There is also often a lack of transparency about the quality of underlying data created using Optical Character Recognition, which

5 www.johnfoxe.org. (24.03.2017)

6 I leave aside the question of whether the digitisation of particular sources that are already well used cements their position, leading to their increased use and the neglect of others. The issue has been noted by several scholars. One of the earliest was Adrian Bingham in relation to digitised newspapers: Bingham, 'The digitization of newspaoper archives', 229; more recent is the important contribution of Lara Putnam, 'The trans-national and the text-searchable'.

7 Gomez and Keller, 'The missing decades'.

makes interpretation of search results more difficult, as does the deployment of 'black box' search algorithms, the designs of which are not transparent to users.[8]

BUT THERE IS DATA!

So far, I have concentrated almost entirely on the retrospective digitisation of print and manuscript, and contemporary historians need to reconcile themselves to the fact that the combination of sheer volume, copyright law and restricted funding means that access to digitised primary sources is unlikely ever to be as comprehensive as it is for earlier periods.

But there is data, vast amounts of it, but rarely used for contemporary history. What is this data? It is the data collected by social scientists – political scientists, economists, sociologists and others – about almost every aspect of social and economic life, for the needs of their own research. It dates from the earliest days in academic computing, giving perhaps some forty years of useful data. The Consortium of European Social Science Data Archives (cessda.net) comprises some 15 national research data services including those for Denmark and the UK. In early 2016 the UK Data Service contributed nearly 600 datasets classified as 'historic': Canadian census and election data from 1908 to 1968; the speed of trains in the UK from 1910 to 2008; infant mortality statistics from Georgian London. Some of them are produced by historians, particularly for more distant periods, but as the temporal coverage comes closer to the present, they are the product of social scientists. Of the nearly 6,000 data collections in the UK Data Archive, 90% were not classified as historic. But for contemporary historians, very soon *all* this data will be historic.

Why then is this data used by relatively few contemporary historians? Firstly, historians (at least in the UK) until very recently have not been routinely trained in handling and analysing data. This situation is changing gradually, but the training is itself often given by supervisors who themselves are not wholly sure of the tools and methods involved. For others, the issue is not so much that it is data, but that it was prepared within the intellectual frameworks of other disciplines. There is perhaps a concern that historians do not understand the governing assumptions behind the preparation of the data: assumptions common in sociology or economics such that they are nowhere stated, leaving the historian exposed to the risk of significant misunderstanding. The terminology is also often unfamiliar, as are the systems of classification being used. Since these datasets are usually prepared to answer a research question which is not *ours*, perhaps they are also somehow flawed (it is thought); perhaps the nature of the originating research question resulted in a dataset that includes and excludes the wrong things: not too many data, but the *wrong data.* Despite all these concerns, this kind of data

8 For a useful survey of the issues in quality control for digital libraries, see J. York and K. Hagedorn, 'Quality in Hathi Trust'.

created by researchers in other disciplines constitutes a rich and increasingly important class of source.

THE BORN-DIGITAL REVOLUTION

So far I have dealt with sources that originate on paper and are subsequently digitised, and with secondary datasets compiled as part of the research of others. However, a seismic shift in the nature of our sources is under way, in two parts, one of which is already visible in the archives available to us, and the other which we will begin to see in the next 5 to 10 years.

Firstly, the Web. It can have escaped no-one's notice that there has been a massive transition in the last 20 years, as public communication which had previously been primarily either face-to-face, or in print form – publications, journalism, ephemera such as leaflets – has migrated to the Web. To begin with, material published online was often the duplicate of a printed object, but relatively soon the Web became the sole point of publication for many kinds of material. Although not many historians are yet aware of it, there is already a massive body of source material captured by web archives, but yet hardly exploited. In early 2016, the Internet Archive, the largest Web archive, held some 280 billion archived pages (more than 500 billion individual digital files). Several nations, including Denmark, have legal frameworks in place (often known as non-print legal deposit) that allow national libraries to archive the whole of the national Web, subject to various restrictions on access to the archive.[9] For much of the first 20 years of Web archiving, the archive grew faster than historians' understanding of how to use it, but the last few years have seen the growth of a new sub-discipline which might be termed 'Web history', as historians have begun to grapple with issues of provenance and interpretation.[10] It is already the case, however, that a history of the 1990s is near impossible to write without engagement with the archived Web.

The other most significant shift to the digital which is yet to have its impact on historians' working practices concerns the private archives of organisations and individuals. The shift towards digital record keeping predates the Web and includes: digital documents held on shared drives, intranets and private wikis; financial accounting systems; systems that record staff attendance; systems that record movements of people on public transport networks; archives of internal and external email, and more besides. All of these are of potential use to historians, but because they are not yet available to scholars, the impact is yet to be felt.

Until very recently, the National Archives of the UK imposed a delay of 30 years (as do many other archives) before unpublished materials may be seen by scholars. As a result archives are beginning to address the issue of how to provide

9 For the history of web archiving, see Webster, "Towards a cultural history of world Web archiving", passim.

10 Brügger (ed.), *Web history*; Brugger and Schroeder (eds.), *The Web as history*.

access to digital material created in those thirty years between 1987 and the present. While Web archives have been with us for nearly 20 years, we have yet really to see this other kind of material on our desks. The material being released to the public from the late 1980s is still overwhelmingly on paper. But, in a very short time, electronic records will start to appear, and as a community of scholars we have not yet thought through how *exactly* we will wish to use an archive of emails, or the log of changes made to a government department intranet. We are at a crucial point in time, then, and I hope to see a period of engagement between archivists and scholars in how to deal with this material.[11]

By now, readers may be feeling rather overwhelmed by the scale and complexity of the challenge as it will present itself to us in the next few years. Though the challenge is significant, historians have adjusted to new kinds of sources in the past and will do so again. But the digital turn suggests some wider changes in the institutional circumstances in which historians are formed and then do their work.

Firstly, there is a pressing need to build digital method training into graduate research training – and indeed earlier, arguably, in undergraduate programmes. As Lara Putnam has shown, this ought not to be a matter simply for the minority who may themselves go on to exploit the most advanced techniques; a much greater depth of critical thinking about the implications of the whole digital turn is urgently needed.[12] However, we may be in a phase where methodological change is unprecedentedly fast and will remain so, assuming that the technologies involved continue to develop at least as quickly as they now do. As such, new research tools are put into use by small communities of scholars whilst still in a process of development and improvement in response to their needs. This is very different to the more familiar model, characteristic of Microsoft or Apple, where software is 'finished' and only then released. Given this, I expect that the best source of training will remain other historians. In this, social media have become vital, in that they allow the formation of specialist communities around particular tools and kinds of sources that would not form in any one locality. Historians have also taken it into their own hands to provide training and methodological reflection in a collaborative way, in services such as The Programming Historian and Web Archives for Historians.[13]

The digital turn also has implications for historians' way of writing. Social scientists are accustomed to documenting the theoretical basis of their work and the precise research methods used. In history theses of the generation of my own,

11 In the UK, the Arts and Humanities Research Council funded a network in 2016-17 on born-digital data for history to begin to address just these issues. More details may be found on the project website at http://www.history.ac.uk/projects/digital/born-digital-big-data-and-methods-history-and-humanities (24.03.2017).

12 Putnam, 'The trans-national and the text-searchable', 378-9.

13 http://programminghistorian.org/ ; https://webarchivehistorians.org/ (24.03.2017).

this would have been strange, since much of the historians' craft was taken as given; the implicit assumption was that the reader knew by which means the work was carried out and conclusions reached. In sub-fields of history where there is still a particular value placed on the quality of our writing *as writing* – in terms of the elegance of the style – there is, I think, a resistance to the explicit documentation of method, lest it scare or bore the reader (and particularly the non-specialist reader). This may have been possible when our methods were more settled. If, however, I am correct in arguing that we are in a time of faster methodological change than ever before, it will not be sustainable. When only a minority of readers can be assumed to know the particular tool or method in play, such transparency cannot be avoided.

Historians trained as I was tend to think in terms of unique things – texts, artefacts, images – the interest of which lies precisely in what makes them different from other things. Historical research has often been closer to the status of an art than of a science, in the detailed and indeed imaginative recreation of the particular significance of a single thing in a single time and place. Concern is sometimes expressed that the rush to the digital, and the focus on 'big data' in particular will result in this being lost. I am not exercised by that concern to any great extent, but there is an opportunity in learning to think in both ways: both in terms of sources as unique things, but also about bodies of source material as a whole, and of which characteristics of a source can be extracted and handled as data. Biographical narratives such as appear in monumental dictionaries of national biography are in one sense the very epitome of the particularity, the irreducible uniqueness of the individual life. But viewed another way, they are also rich datasets for the study of lifespan, education, or the patterns that careers take. Bodies of text can often function both as unstructured text and structured data. There is no reason why it should be necessary to choose between traditional historical method and the so-called 'distant reading' of our sources; there is simply now an opportunity to be grasped to use both methods as part of the same enquiry.[14]

I earlier noted the need to acquire data handling skills, and familiarity with particular software tools, in order to navigate these new seas. However, when working with data at a certain scale, we will often need help. In the UK and elsewhere, there has been a noticeable trend towards interdisciplinary working in research projects, in which humanities scholars work together with computer scientists in order to understand data of a scale and complexity that requires such specialist help.[15] But this is not the only cluster of relationships that is changing. Librarians and archivists are in the midst of a very significant change in ways of

14 One example of such a hybrid method using Web archives is Webster, 'Religious discourse in the archived Web'.

15 One such project is the Illustration Archive from Cardiff University: http://www.cardiff.ac.uk/news/view/89413-the-illustration-archive (21.02.2017).

working, in relation to their users. It was previously enough to take a thing – a printed volume, or an archival box – and place it upon a scholar's desk; there was no need to know what was being done with it in order to deliver it correctly. Now, as material is delivered digitally, every design decision taken when building new user interfaces allows some kinds of use but may exclude others. The more far-sighted archivists have recognised that this means building a new kind of relationship with the user, at the very beginning of that process.[16] This is then a call to historians to be there at the beginning of that process, to help design those systems to meet our needs.

BIBLIOGRAPHY

Bingham, Adrian: 'The digitization of newspaper archives: opportunities and challenges for historians', *Twentieth Century British History* 21 (2), 2010, 225-31.

Brügger, Niels (ed.): *Web history*. New York: Peter Lang, 2010.

Brügger, Niels and Schroeder, Ralph (eds.): *The Web as history*, London: UCL Press, 2017. Also available Open Access at https://www.ucl.ac.uk/ucl-press/browse-books/the-web-as-history (24.03.2017).

Gomez, Pablo Uceda and Keller, Paul: 'The missing decades: the 20th century black hole in Europeana' (2015): http://pro.europeana.eu/blogpost/the-missing-decades-the-20th-century-black-hole-in-europeana (20.02.2017).

Putnam, Lara: 'The trans-national and the text-searchable: digitized sources and the shadows they cast', *American Historical Review* 121 (2), 2016, 376-402.

Shoemaker, Robert: *Digital transformations: the Old Bailey Online and historical research* (unpublished lecture, 2012), http://anglais.u-paris10.fr/spip.php?article2092 (20.02.2017).

Truax, Jean: *Archbishops Ralph d'Escures, William of Corbeil and Theobald of Bec. Heirs of Anselm and Ancestors of Becket*, Abingdon: Routledge, 2012.

Webster, Peter: *Archbishop Ramsey. The shape of the church*, Abingdon: Routledge, 2015.

Webster, Peter: 'Religious discourse in the archived web: Rowan Williams, archbishop of Canterbury, and the sharia law controversy of 2008'. In Niels Brügger and Ralph Schroeder (eds), The Web as History: the first two decades, London: UCL Press, 2017.

Webster, Peter: 'Users, technologies, organisations: towards a cultural history of world web archiving'. In Niels Brügger (ed.), Web 25: histories from the first 25 years of the World Wide Web, New York: Peter Lang (forthcoming, 2017).

York, Jeremy and Hagedorn, Kat (2015): 'Quality in Hathi Trust': https://www.hathitrust.org/quality-in-hathitrust (21.02.2017).

PETER WEBSTER

PH.D.

DIREKTØR, WEBSTER RESEARCH AND CONSULTING LTD.

HTTP://WEBSTERRESEARCHCONSULTING.COM/

EMAIL:PETER@WEBSTERRESEARCHCONSULTING.COM

16 Such an approach was taken by the BUDDAH project in the UK, bringing together archivists and humanities scholars to develop new access tools for web archives. http://buddah.projects.history.ac.uk/ (24.03.2017).

ABSTRACT

Peter Webster:

Digital contemporary history: sources, tools, methods, issues

This essay suggests that there has been a relative lack of digitally enabled historical research on the recent past, when compared to earlier periods of history. It explores why this might be the case, focussing in particular on both the obstacles and some missing drivers to mass digitisation of primary sources for the 20th century. It suggests that the situation is likely to change, and relatively soon, as a result of the increasing availability of sources that were born digital, and of Web archives in particular. The article ends with some reflections on several shifts in method and approach, which that changed situation is likely to entail.