Web archives and (digital) history: a troubled past and a promising future?

Jane Winters, School of Advanced Study, University of London

Introduction

'For more than four decades, the Internet has grown and spread to an extent where today it is an indispensable element in the communication and media environment of many countries, and indeed of everyday life, culture and society' (Brügger, Goggin, Milligan and Schafer, 2017: 1). So begins the introduction to the journal *Internet Histories: Digital Technology, Culture and Society*, launched in 2017. The World Wide Web, which unlocked the full potential of the Internet, has been with us for nearly 30 years; and in October 2016 the Internet Archive celebrated 20 years of capturing, preserving and republishing the Web (Hanamura, 2016). These are pleasingly round figures, indicating the passage of substantial time and the relative maturity both of the Web itself and the processes that have evolved to ensure that it is archived for the benefit of researchers. But those same researchers, and historians in particular, remain largely oblivious to the richness of the archived Web as a primary source for the study of the recent past, if not oblivious to the very existence of Web archives.[1] This chapter will examine the reasons for historians' relative failure to engage with the archived Web, and suggest why it is critical for contemporary, political and digital historians at least to do so. It will go on to explore the changing relationship between archivists, librarians and historians, which is beginning to break down researchers' reluctance to work with born-digital materials and big data. Finally, it will propose an exciting future for (digital) historical research, which employs a combination of quantitative and qualitative approaches to recover the lives and voices of ordinary people.

Historians and web archives

The Web, like the newspapers that it now incorporates, contains material of interest for every sub-discipline of history – politics, sport, finance, culture, food, fashion, conflict are all present in infinite variety. Web archives, imperfect though they may be,[2] reflect this range and diversity; there is something for everyone. But there are three (overlapping) groups for whom Web archives might be expected to hold immediate and particular interest: contemporary, political and digital historians. Why is there so little evidence that they are engaging with this new primary source, or indeed with a whole range of born digital archives?

Contemporary history

According to Kandiah (2008), 'the aim of contemporary history is to conceptualise, contextualise and historicise – to explain – some aspect of the recent past or to provide a historical understanding of current trends or developments'. Web archives are an invaluable lens through which to study life in the developed West in the late twentieth and early twenty-first century, but perhaps that past is still too recent, the digital apparently still too new. Weber (2017: 26) reports that 'When I told people I was researching the history of the Web in early 1995, about half of them were amused: "But it's too young to have a history!"' The strong connection between the Web and journalism may also be a problem here. As Kandiah notes, 'Critics of the discipline feared that contemporary history could … at best be nothing more than a form of journalism because its concerns were so closely rooted to the present'. In relation to Web archives, and the history of the Web, it is the desire to historicise that seems to be most dominant. A conference on 'History and the Internet' organised by History and Policy[3] in December 2016, for example, included presentations on Domesday Book as big data and parliament and print culture in the seventeenth century, but only limited treatment of the Internet itself. For historians who approach the Web in this way it is not necessary to engage with its archives, although arguably to do so would enrich their understandings of our most recent

twentieth-century media revolution. It is to be hoped that this engagement will come with greater chronological distance, as the early technologies of the Internet become as unfamiliar as those of the printing press or the scriptorium.

Political history

The question of how to work with born digital data is more pressing for political historians, some of whom, of course, would also think of themselves as working in the field of contemporary history. Governments have taken to the Web with marked enthusiasm and no little skill, as they seek to engage with, provide support for and learn more about their citizens. A 2016 United Nations report (p. 82), for example, identified the UK as world-leading in e-government, noting 'a Whole-of-Government approach in online service delivery'. Data portals like data.gov.uk, open.canada.ca and data.gouv.fr are increasingly making the workings of government transparent, but they serve as an early warning to political historians that they will have to change how they work. In the sphere of government, the adoption of digital means of communication, both internally and externally, has been definitive and startlingly quick. Political historians will soon have little choice but to seek information from Web archives because government plays out on the Web. They 'will need to transpose their long-established disciplinary skills and instincts into a digital register: asking the usual critical questions about their source material – how it was produced and why it has survived – and establishing a deep and rich set of contexts through which to interpret it' (McCarthy, 2016).

For historians in both these and other fields, however, the key reason for failing to use Web archives is the requirement to develop new skills, or to refresh old ones.[4] This was an important, if not unexpected, finding of the Big UK Domain Data for the Arts and Humanities (BUDDAH) research

project.[5] The project case studies are revealing of the problems faced by researchers: 'we do not have enough case studies or methodological literature to help us design this research (Millward, 2015: 10); 'Keyword full-text searching as the standard methodology needs to be critically reconsidered' (Deswarte, 2015: 9); archived Web pages bring 'the challenge of defining the analytical object itself' (Huc-Hepher, 2015a: 8). Some of the missing skills are technical ones – manipulating and cleaning large quantities of data is much easier if you have an understanding of Regular Expressions,[6] for example – but others relate to historians' (in)ability to work with statistics, to undertake even the most basic quantitative analysis. A lack of statistical understanding impedes both analysis at scale and the sampling that might facilitate closer reading and micro-analysis. The turn away from quantitative methods and approaches that has characterised much recent historical research and training has left historians singularly ill-equipped to deal with increasingly vast Web and other born digital archives.[7]

Digital history

If a dearth of appropriate skills, and skills training, is hindering many historians from studying Web history, or from adding Web archives to their basket of primary sources, one might assume that this would not be the case for our third group: digital historians. But until very recently, the focus of even digital history has lain elsewhere. This lack of attention is both striking and, in my view, surprising. Digital humanities 'has its origins in the research carried out … in textually focused computing'; it has diversified admirably in the second decade of the 21st century, but as recently as 2004 there could be no argument that 'it remains deeply interested in text' (Schreibman, Siemens and Unsworth, 2004). Web archives contain vast quantities of text, but they are far removed from a digital scholarly edition or a corpus prepared for linguistic analysis.[8] Digital history, by contrast, has no such unifying thread. It encompasses GIS and approaches drawn from historical geography; significant elements of

public history, as exemplified by the work of the Roy Rosenzweig Center for History and New Media; scholarly editing and textual scholarship; economic and social research, influenced by social science methodologies; prosopographical, biographical and genealogical investigation; and so on. Its proponents are interested in digital pedagogy, in scholarly communication, in big and small data, micro- and macro-analysis. All of these approaches and interests may be brought to bear on the historical Web, yet digital historians have generally displayed the same 'tepid interest' in Web archives and Internet histories as other humanities researchers (Weber, 2017: 27).

One explanation is that while digital history has embraced a range of historical sub-disciplines, and borrowed readily from cognate subjects like archaeology and historical geography, it has largely failed to take account of developments in two crucial areas: library, archive and information studies; and digital preservation. Libraries and archives have necessarily been at the forefront of Web archives research and practice: it is they who have been responsible for developing the tools and protocols to harvest the Web, for running Web crawls, for devising preservations tools and standards, for exploring how to document and search Web archives of varying size and scale. This is work that is discussed among the members of the International Internet Preservation Consortium (IIPC), but not among historians, digital or otherwise. Webster (2017) rightly notes that much of the debate about the impact of 'the transition from paper to digital in records management and archiving … is to be found in the journals of the archival profession, into which historians rarely look'. The boundaries between digital history and digital preservation are even more clearly delineated; as with the conservation of books and manuscripts, digital preservation may only be noticed when it has failed in some way.[9] The first reaction of a historian on seeing an archived Web page is more likely to be 'Why are those images missing?' than 'How has so much of this page been successfully preserved?' The narrative of the 'digital dark age', which sometimes seems ubiquitous in the

mainstream media, only persists because of a general lack of awareness of, and appreciation for, the scope of existing digital preservation work and expertise (Winters, 2017a: 45).

Changing times?

But there are signs that this is beginning to change. In the UK, for example, the Arts and Humanities Research Council (AHRC) has funded two separate research networks which bring together historians, archivists, librarians and digital preservation specialists, among others, to discuss the challenges posed by collecting, preserving, publishing and using born-digital data of all kinds.[10] The much larger European network RESAW (A Research infrastructure for the Study of Archive Web Materials)[11] similarly includes both humanities researchers, including several historians, and representatives of memory institutions with a responsibility for archiving the Web. More events are being organised which offer something to multiple sectors. There has generally been very little exchange of ideas and personnel between the 'Digital History' and 'Archives and society' seminars hosted by the Institute of Historical Research in London, for example, but in January 2017 a joint seminar was organised on 'Sensitivity review and digital records' (Seles, 2017). In June 2017, RESAW and the IIPC collaborated to run a conference which considered 'Researchers, practitioners and their use of the archived Web', highlighting the value and importance of cross-sectoral conversations. The strict separation of spheres that has obtained for so long is beginning to break down in the face of the challenges posed by born-digital data, and nowhere is this more apparent than in relation to Web archives.

Mediating access: historians, librarians, archivists and the archived Web

If some historians are still prone to confuse Web archives with archives of historical data which happen to have been published on the Web – 'archive' is not a particularly helpful term here, as Brügger (2016) has discussed – it does seem as though a turning-point has been reached. It no longer seems entirely fanciful to argue that we are moving towards the promising future of this chapter's title; a future in which many different types of historian, not just those with an interest in contemporary politics or digital methods, can integrate Web archives into their research. For most of those historians, early encounters with Web archives are likely to be mediated by archivists and librarians (another reason to ensure that disciplinary silos are breached). Given the scale of most Web archives,[12] and the consequent limitations of keyword searching, curated special collections provide an easy and obvious route in to the data. They are also more likely to be openly available than the broad national crawls undertaken by libraries and archives on a statutory basis.

To date, the British Library has published 45 special collections around themes which have been deemed 'useful and interesting' by curators.[13] They immediately showcase both the chronological span of the archive and the range of human activity represented within it. One of the earliest special collections is concerned with the terrorist attacks that took place in London on 7 July 2005, which killed 52 people and injured more than 700; the most recent deals with the UK General Election of 2015. A special collection capturing the 2016 EU Referendum debate is in preparation, and will soon be made available to researchers (Kunze, 2016). There is plenty of material here for political historians: a series of UK general elections from 2005 to 2015; the Scottish parliamentary election of 2007 and the Scottish Independence Referendum of 2014; the London Mayoral election of 2008 (although not those of 2012 or 2016); the European Parliament election of 2009; the Credit Crunch, 2008-2010. Other areas of strength include sport (the Commonwealth Games held in Glasgow in 2014, the London Olympic and Paralympic Games of 2012), anniversaries of national significance (the 200th anniversary of the birth of Charles Darwin, Queen Elizabeth II's Diamond Jubilee in 2012,

the centenary of the Easter Rising in 2016), health (collections dedicated to mental health, personal experiences of illness, the 2012 Health and Social Care Act and even pandemic influenza outbreaks since 2005) and religion (a general collection on religion, politics and law since 2005 and more specific ones concerned with the Quakers and the Free Churches).

These special collections are enormously rich and diverse, but that very eclecticism poses something of a problem for Web history and historians. Why, for example, are there collections relating to Cornwall and Hampshire but to no other counties in the UK? What was it about the Cambridge Network, 'a membership organisation based in the vibrant high technology cluster of Cambridge', which led to its being singled out in this way? Why have 26 websites concerned with e-publishing trends been given special status alongside those dealing with national politics? Clearly, certain types of predictable event generate special collections in the Web archive – elections, anniversaries, major sporting occasions; other clusters are responses to the unexpected, to natural disasters like the Indian Ocean Tsunami in December 2004 or to terrorism. This latter trend is also apparent in the collections developed through Archive-It, which is described as 'The leading web archiving service for collecting and accessing cultural heritage on the web'.[14] Archive-It involves more than 400 institutions in 16 countries who between them have curated more than 4000 special collections. Of these, 178 (4.25 per cent) are categorised as arising from 'Spontaneous events', and while the first collection on the list is an archive of the 100,000 Poets for Change website,[15] many concern shocking and more or less unpredictable events, from the 2013 Boston Marathon bombing to Hurricane Katrina.[16] Politics looms large too: for example, 412 collections (9.83 per cent) are categorised as relating to 'Government', and there are numerous smaller and perhaps overlapping clusters concerned with particular elections or states in the US.[17]

In both of these instances, the special collections serve an important role in illuminating the wider Web archives from which they are derived, respectively those of the British Library and the Internet Archive. They act as a shop window for archives that are challenging to encounter at scale, encouraging initial browsing which might then lead on to more in-depth analysis and research. This necessarily imbues them with an importance that may not always have been considered by those responsible for their creation. A particular collection almost certainly has enormous value for the curator(s), and for the many others who will explore it in years to come, but what does it say about the shape and significance of the wider Web archive? The British Library's remit to preserve and make available the UK's intellectual and cultural heritage is apparent in the Web archive collections that deal with significant and/or traumatic events, but others are suggestive of personal interest and enthusiasm or a serendipitous partnership.[18] This is even more the case for Archive-It, where some collections have been curated in specific teaching contexts, for example.[19] This is still an evolving landscape, and experimentation is entirely appropriate, but there is a risk that these early experiments may begin to 'fix' a particular view of Web archives and the kinds of historical research for which they are most suitable. This is particularly true if access to the larger archives remains restricted, for legal, technical or other reasons. The choices that are made now could resonate for decades to come, and some of the consequences might be unintended: as Schwartz and Cook (2002) note, 'Archives – as records – wield power over the shape and direction of historical scholarship, collective memory, and national identity, over how we know ourselves as individuals, groups and societies' (p. 2), and archivists have the power to shape how we access those records.

The historian as archivist

If librarians and archivists will play an important role in determining not just what is included in Web archives but how that archived material is used by historians, even the kinds of questions that they

will ask, there is also considerable scope for historians to take personal responsibility. Web archiving at scale is a highly technical process, requiring investment in expertise and equipment, but researchers can build their own collections. This tendency towards personal archiving has been present from the very earliest days of Web history. Brügger, for example, noted in 2010 that his personal Web archive 'contains a substantial part of the Danish web activity in relation to the Olympic Games in 2000' (p. 351). There are a number of options available to researchers who would like to assume some control over the archiving process. They might, for example, choose to collaborate with archivists and programmers, as described in Milligan, Ruest and St. Onge (2016); or they might investigate the various open-source Web archiving tools that have been developed, like Warcbase and Wget.[20] Milligan in particular has shown what can be achieved when historians embrace these approaches and participate in the creation as well as the analysis of Web archives (2012; 2017). Both of these approaches to Web archiving, however, require a degree of technical expertise with which many historians are, and are likely to remain, uncomfortable. Fortunately, there are alternatives that do not involve such a steep technical learning curve, notably Webrecorder.[21]

Webrecorder allows anyone 'to create high fidelity, context rich and interactive archives of the dynamic web' (Espenschied, 2016). The resulting collections of WARC files are not just personal but personalised: the pages are captured as the researcher moves through a website, keeping a record of her chosen pathway. This personalisation even extends to the faithful recording of the 'logged-in' experience on online social media. A shallow hierarchical structure is present in these archives, with 'sessions' organised into collections, and simple descriptions can be added to aid future navigation and discoverability. All of this functionality is available in the browser, but the archived files can also be downloaded and replayed offline using a desktop app, the Webrecorder Player.[22] This is a very different approach to Web archiving from the comprehensive full domain crawl undertaken by large

memory institutions, one which supports a micro-level approach both to the harvesting and study of the Web. Archiving a website of any size, for example that of a major public broadcaster, would be difficult and time-consuming using this method, with all links having to be followed to ensure their successful capture. The collections that can be created in this way will remain relatively small, and focused narrowly on a researcher's interests and experiences. In some instances they will record and reflect her personal social and research networks online, providing some of the ethnographic context that is missing from the larger automated Web crawls. There are layers of interest and value in such archives, which can only be fully realised if they are shared, with other researchers and perhaps ultimately with libraries and archives.

Capturing the voice of the individual

This focus on the small, on the personal, is one possible future for Web history; and one which reflects growing interest in the value of digital storytelling (see, for example, Burgess, 2007; Coleborne and Bliss, 2011). Web archives, which record so many different voices, hold out the promise of a new golden age of history from below.[23] It is not true to say that anyone can create online content which might find its way into a Web archive – certain groups are still privileged, depending on education, age, social class, geographical location, and so on[24] – but there are unprecedented opportunities to self-publish, to comment on the publications of others. The sheer diversity of authors who may be represented in Web archives is highlighted by a Blogs special collection in the UK Web Archive: among the 763 blogs included, alongside those of politicians and protest groups, are 'Alan in Belfast', who writes about 'cinema, books, technology, and the occasional rant about life'; 'Nelly's Garden', which presents the thoughts of Nelly Culleybackey from County Antrim; and 'Lizzy's Literary Life', 'celebrating the pleasures of a 21st century bookworm'.[25] Ian Milligan's work on the GeoCities archive has begun to recover the voices of the children who

were active participants in the early web, and specifically in the Enchanted Forest 'neighbourhood' (Milligan, 2017); Megan Dougherty has shone a spotlight on the brief flourishing of the subcultural Islamic punk movement in North America from the early 2000s (Dougherty, 2017); and Saskia Huc-Hepher has explored the histories of French communities in London through their archived blogs (Huc-Hepher, 2015b). But this volume and variety poses challenges for historians. How do we make sure that we find individual voices among all the noise? How can we judge the significance of what they are saying when any and all points of view will have been captured, but we have little or no data indicative of circulation or popularity?

Big history and the macroscope

Alongside this renewed emphasis on close reading, the current abundance of digital data has led to calls for a return to 'big' history and the *longue durée* (Guldi and Armitage, 2014), for the adoption of the distant reading approach first proposed by Franco Moretti for the study of digitised literary corpora (Moretti, 2013). Graham, Milligan and Weingart, 2015 argues for historians to embrace big data – one chapter is even titled 'The joys of big data for historians'. The authors are among an increasing number of historians to apply the concept of the macroscope (see, for example, Jockers, 2013; Hitchcock, 2014), which 'instead of allowing you to see things that are small or far away … makes it easier to grasp the incredibly large. It does so through a process of compression, by selectively reducing complexity until once-obscure patterns and relationships become clear' (Graham, Milligan and Weingart, 2015: 15). This is a far more nuanced approach than Culturomics, which more or less completely failed to take account of humanities research practices and concerns (Michel *et al.*, 2011). The proponents of the macro-historical are not suggesting that the micro-historical should be abandoned in the face of the data deluge, rather that there is value in

considering both; and indeed the interactions between the two can produce new insights and interpretations.

The small and the large: a question of scale

As is so often the case, the most promising approaches for historical research in this field bring together the small and the large, 'the general and the particular' (Manovich, 2016). The challenge is for historians to find new ways of working without losing the emphasis on the individual that has long distinguished humanities research. Using the framework of cultural analytics, Lev Manovich proposes that 'we may combine the concern of social science, and sciences in general, with the *general* and the *regular*, and the concern of humanities with *individual* and *particular* … analysing massive datasets to zoom in on the unique items' (Manovich, 2016). Tim Hitchcock argues that, in contrast with 'the categories of knowing that dominated the nineteenth and twentieth centuries', born digital data, and Web archives in particular, provide 'an opportunity to re-think what is possible, and to re-think what it is we are asking; how we might ask it, and to what purpose'; but while historians must 'be able to wield the tools of large-scale visualization … we need to do so at the same time as we preserve the values and practices that underpin traditional academic history, while going beyond the standards of scholarship we have inherited' (Hitchcock, 2015 and 2013: 20).

A promising future?

These are provocative calls to action, which speak to an exciting future for historical research – one which is predicated on the ever-increasing availability of digital sources and the development, and widespread adoption, of innovative digital methods that build on the best traditions of humanistic

exploration. It is, however, an exciting future that has been invoked before in the current life-cycle of digital history. A report on the impact of digital resources published in 2011, for example, considered whether 'they simply make research easier, cheaper (to the researcher), more convenient, and less time consuming, or whether there is evidence that they open up *new avenues for* research'. The story here is one of enormous potential: researchers are perceived to be developing new methods and even theories, but 'How these perceived changes result in new research questions across the humanities in the long run is still emerging' (Meyer, 2011: 39). Perhaps it is unfair to consider sixteen years as 'the long run', but this description of humanities research, and digital history, would not seem out of place today. The promise of transformation is still tantalisingly there, but how close is it to being realised?

Web archives, and other kinds of born-digital data, do bring the possibility of, and perhaps even necessitate, a radical reframing of digital history – through their scale, their heterogeneity, their complexity, their fragility. Some historians will undoubtedly continue to focus solely on the textual elements of the archived web, abstracting words from their rich digital context. But others will work with text, sound, and still and moving image in the round. In doing so they might engage with the history of art and design, media and communication studies, the history of technology, linguistics, film studies – and with other researchers in those fields. They might move across the boundaries of the (social) sciences, arts and humanities, learning new skills themselves or building partnerships. They might, as Marc Weber urges, turn their attention to the materiality of the web, learning from the work of museums.[26] They might do some or all of these things in combination. And they might, by combining big data approaches with humanistic understandings, at last begin to develop genuinely new research questions and generate new knowledge.

It is vitally important that historians, digital or otherwise, should carve out a space for themselves in the study of both the live and the historical web, especially where these are conceived first and foremost as big data. The individual, the human, risks becoming lost in the face of arguments that 'Causality … is being knocked off its pedestal as the primary foundation of meaning. Big data turbocharges non-causal analyses, often replacing causal investigations' (Mayer-Schönberger and Cukier, 2017: 66); or claims that 'Culturomic results are a new type of evidence in the humanities' (Michel *et al*., 2011: 181). Large-scale trends are, of course, enormously important to understand, as Mayer-Schönberger and Cukier demonstrate, but historians are very well placed to combine an appreciation of broader patterns and movements with a forensic understanding of the small-scale and the human – of ordinary lives not just data points. And the archives of the Web are a unique record of many millions of ordinary lives, alongside histories of celebrities, institutions and nations. As yet those archives remain largely unexplored, but there are many Web histories to be uncovered, many pathways to be explored, and many new questions to be asked.

Reference list

Belovari, S. (2017) 'Historians and Web archives', *Archivaria*, 83: 59-79.

Blank, G. (2013) 'Who creates content? Stratification and content creation on the Internet', *Information, Communication and Society*, 16(4): 590-612. DOI: 10.1080/1369118X.2013.777758.

Brügger, N. (2016) 'Webraries and Web archives – the Web between public and private', in D. Baker and W. Evans (eds), *The End of Wisdom? The Future of Libraries in a Digital Age*. Oxford: Chandos Publishing. pp. 185-90.

Brügger, N. (2012a) 'Web history and the Web as a historical source', *Zeithistorische Forschungen*, 9(2): 316-25.

Brügger, N. (2012b) 'When the present Web is the later past: Web historiography, digital history, and Internet studies', *Historical Social Research/Historische Sozialforschung*, 37(4): 102-17.

Brügger, N. (ed) (2010) *Web History*. Bern, Switzerland: Peter Lang US.

Brügger, N. and Finnemann, N. O. (2013) 'The Web and digital humanities: theoretical and methodological concerns', *Journal of Broadcasting and Electronic Media*, 57(1): 66-80. DOI: 10.1080/08838151.2012.761699.

Brügger, N., Goggin, G., Milligan, I. and Schafer, V. (2017) 'Introduction: Internet histories', *Internet Histories: Digital Technology, Culture and Society*, 1(1-2): 1-7. DOI: 10.1080/24701475.2017.1317128.

Burgess, J. (2006) 'Hearing ordinary voices: cultural studies, vernacular creativity and digital storytelling', *Continuum: Journal of Media and Cultural Studies*, 20(2): 201-14. DOI: 10.1080/10304310600641737.

Coleborne, C. and Bliss, E. (2011) 'Emotions, digital tools and public histories: digital storytelling using Windows Movie Maker in the history tertiary classroom', *History Compass*, 9(9): 674-85. DOI: 10.1111/j.1478-0542.2011.00797.x.

Cowls, J. (2016) 'Cultures of the UK Web', in N. Brügger and R. Schroeder (eds), *The Web as History: Using Web Archives to Understand the Past and the Present*. London: UCL Press. pp. 220-37.

Davies, M. (2014) Corpus of Global Web-Based English (GloWbE) (https://corpus.byu.edu/glowbe/).

Department of Economic and Social Affairs (2016) *United Nations E-Government Survey 2016: E-Government in Support of Sustainable Development*. New York: United Nations (http://workspace.unpan.org/sites/Internet/Documents/UNPAN97453.pdf).

Deswarte, R. (2015) Revealing British Euroscepticism in the UK Web domain and archive. London: School of Advanced Study. pp. 1-10 (http://sas-space.sas.ac.uk/6103/1/Deswarte%20case%20study.pdf).

Dougherty, M. (2017) '"Taqwacore is dead. Long live Taqwacore" or punk's not dead? Studying the online evolution of the Islamic punk scene', in N. Brügger and R. Schroeder (eds), *The Web as History: Using Web Archives to Understand the Past and the Present*. London: UCL Press. pp. 204-19.

Dougherty, M., Meyer, E. T., Madsen, C. M., van den Heuvel, C., Thomas, A. and Wyatt, S. (2010) *Researcher Engagement with Web Archives: State of the Art*. Jisc Report (https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1714997).

Espenschied, D. (2016) Introduction to Webrecorder (https://www.youtube.com/watch?v=n3SqusABXEk&feature=youtu.be).

Graham, S., Milligan, I. and Weingart, S. (2015) *Exploring Big Historical Data: the Historian's Macroscope*. London: Imperial College Press (http://www.themacroscope.org/2.0/).

Guldi, J. and Armitage, D. (2014) *The History Manifesto*. Cambridge: Cambridge University Press (https://www.cambridge.org/core/books/the-history-manifesto/AC1A1EC711AE91A4F9004E7582D79AFD#).

Hanamura, W. (2016) The Internet Archive turns 20! Internet Archive Blogs (https://blog.archive.org/2016/09/19/the-internet-archive-turns-20/).

Hargittai, E. and Walejko, G. (2008) 'The participation divide: content creation and sharing in the digital age', *Information, Communication and Society*, 11(2): 239-256. DOI: 10.1080/13691180801946150.

Hartmann, S. (2015) 2015 UK domain crawl has started. UK Web Archive blog (http://blogs.bl.uk/webarchive/2015/09/2015-uk-domain-crawl-has-started.html).

Hitchcock, T. (2015) The UK Web Archive, born-digital sources, and rethinking the future of research. Web Archives for Historians (https://webarchivehistorians.org/tag/tim-hitchcock/).

Hitchcock, T. (2014) Big data, small data and meaning. Historyonics blog (http://historyonics.blogspot.co.uk/2014/11/big-data-small-data-and-meaning_9.html).

Hitchcock, T. (2013) 'Confronting the digital: or how academic history writing lost the plot', *Cultural and Social History*, 10(1): 9-23. DOI: 10.2752/147800413X13515292098070.

Huc-Hepher, S. (2015a) Searching for home in the historic Web: an ethnosemiotic study of the London-French habitus as displayed in blogs. London: School of Advanced Study. pp. 1-27. (http://sas-space.sas.ac.uk/6252/1/Huc-Hepher%20case%20study.pdf).

Huc-Hepher, S. (2015b) 'Big Web data, small focus: an ethnosemiotic approach to culturally themed selective Web archiving', *Big Data & Society*, 2(2) (http://journals.sagepub.com/doi/abs/10.1177/2053951715595823).

Jockers, M. L. (2013) *Macroanalysis: Digital Methods and Literary History*. Urbana-Champaign: University of Illinois Press.

Kandiah, M. D. (2008) Contemporary history. Making History (http://www.history.ac.uk/makinghistory/resources/articles/contemporary_history.html).

Knox, D. (2013) Understanding Regular Expressions. *The Programming Historian* (https://programminghistorian.org/lessons/understanding-regular-expressions).

Kunze, S. (2016) Capturing and preserving the EU Referendum debate (Brexit). UK Web Archive blog (http://blogs.bl.uk/webarchive/2016/06/capturing-and-preserving-the-eu-referendum-debate-brexit.html).

McCarthy, H. (2016) Political history in the digital age: the challenges of archiving and analysing born digital sources. Impact of Social Sciences – LSE Blogs (http://eprints.lse.ac.uk/66690/).

McKie, R. and Thorpe, V. (2002) Digital Domesday Book lasts 15 years not 1000. *The Guardian* (https://www.theguardian.com/uk/2002/mar/03/research.elearning).

Manovich, L. (2016) 'The science of culture? Social computing, digital humanities and cultural analytics', *Journal of Cultural Analytics*. DOI: 10.22148/16.004 (http://culturalanalytics.org/2016/05/the-science-of-culture-social-computing-digital-humanities-and-cultural-analytics/).

Mayer-Schönberger, V. and Cukier, K. (2017) *Big Data: the Essential Guide to Work, Life and Learning in the Age of Insight*. London: John Murray (Publishers).

Meyer, E. T. (2011) *Splashes and Ripples: Synthesizing the Evidence on the Impacts of Digital Resources*. Jisc Report (https://ssrn.com/abstract=1846535).

Michel, J. B. *et al*. (2011) 'Quantitative analysis of culture using millions of digitized books', *Science* 331(6014): 176-182. DOI: 10.1126/science.1199644.

Milligan, I. (2017) 'Welcome to the Web: the online community of GeoCities during the early years of the World Wide Web', in N. Brügger and R. Schroeder (eds), *The Web as History: Using Web Archives to Understand the Past and the Present*. London: UCL Press. pp. 137–58.

Milligan, I. (2017) 'Pages by kids, for kids': unlocking childhood and youth history through the GeoCities Web archive. Researchers, practitioners and their use of the archived Web, International Internet Preservation Consortium/RESAW Conference, London.

Milligan, I. (2016) 'Lost in the infinite archive: the promise and pitfalls of web archives', *International Journal of Humanities and Arts Computing*, 10(1): 78-94. DOI: 10.3366/ijhac.2016.0161.

Milligan, I. (2012) 'Mining the "Internet graveyard": rethinking the historians' toolkit', *Journal of the Canadian Historical Association/Revue de la Société historique du Canada*, 23(2): 21–64. DOI: 10.7202/1015788ar.

Milligan, I., Ruest, N. and St. Onge, A. (2016) 'The great WARC adventure: using SIPS, AIPS, and DIPS to document SLAPPS', *Digital Studies/Le champ numérique*, 2015-16 open issue (http://www.digitalstudies.org/ojs/index.php/digital_studies/article/view/325/412).

Millward, G. (2015) Digital barriers and the accessible web: disabled people, information and the internet. London: School of Advanced Study. pp. 1-12 (http://sas-space.sas.ac.uk/6104/1/Millward%20case%20study.pdf).

Moretti, F. (2013) *Distant Reading*. London: Verso.

Schreibman, S., Siemens, R. and Unsworth, J. (2004) 'The digital humanities and humanities computing: an introduction', in S. Schreibman, R. Siemens and J. Unsworth (eds), *A Companion to Digital Humanities*. Oxford: Blackwell, 2004 (http://www.digitalhumanities.org/companion/).

Schwartz, J. M. and Cook, T. (2002) 'Archives, records, and power: the making of modern memory', *Archival Science*, 2(1-2): 1-19. DOI: 10.1007/BF02435628.

Seles, M. (2017) 'It always seems impossible until it is done': sensitivity review and digital records. Archives & Society and Digital History seminars (https://www.youtube.com/watch?v=ncdbLdshheI).

Taylor, M. (1997) 'The beginnings of modern social British history?', *History Workshop Journal*, 43: 155-76.

Tech Partnership (2017) Basic digital skills UK 2017: summary of findings (https://www.thetechpartnership.com/globalassets/pdfs/basic-digital-skills-standards/basicdigitalskills2016_findingssummary.pdf).

Weber, M. (2017) 'A common language', *Internet Histories: Digital Technology, Culture and Society*, 1(1-2): 26-38. DOI: 10.1080/24701475.2017.1317118.

Webster, P. (2017) Book review: The Silence of the Archive by David Thomas, Simon Fowler and Valerie Johnson. Review of Books – LSE Blogs (http://blogs.lse.ac.uk/lsereviewofbooks/2017/08/11/book-review-the-silence-of-the-archive-by-david-thomas-simon-fowler-and-valerie-johnson/).

Winters, J. (2018) 'Digital history', in P. Burke and M. Tamm (eds), *Debating New Approaches in History*. London: Bloomsbury Publishing.

Winters, J. (2017a) 'Will history survive the digital age?', *BBC History Magazine*, 3: 41-5.

Winters, J. (2017b) 'Coda: Web archives for humanities research – some reflections', in N. Brügger and R. Schroeder (eds), *The Web as History: Using Web Archives to Understand the Past and the Present*. London: UCL Press. pp. 238-48.

[1] A 2010 study into the 'state of the art' noted the gap 'between the *potential* community of researchers who have good reason to engage with creating, using, analysing and sharing web archives, and the *actual* (generally still small) community of researchers currently doing so' (Dougherty *et al.*, 2010: 5).

[2] The problems of working with web archives have been well rehearsed by, among others, Brügger (2012a and 2012b), Brügger and Finnemann (2013), Milligan (2016) and Winters (2017b).

[3] History and Policy is a UK network of historians which 'promotes better public policy through a greater understanding of history' (http://www.historyandpolicy.org/).

[4] I am unconvinced by the argument that historians are reluctant to use Web archives because they are concerned 'that they will not be able to replicate their historical research process when using web archives, and may not find essential and authoritative records' (Belovari, 2017: 60). This seems to me simultaneously to underestimate the variety of humanities research methodologies, the fuzziness of concepts like 'essential' and 'authoritative', and the ability of historians to navigate uncertainty and devise new approaches.

[5] BUDDAH was funded by the Arts and Humanities Research Council (AHRC) as part of its Digital Transformations in the Arts and Humanities theme (https://buddah.projects.history.ac.uk/; grant reference AH/L009854/1). Cowls (2016) usefully summarises the progress made and challenges faced by the 10 researchers who were awarded bursaries to use Web archives in their research.

[6] 'Regular expressions … are a way of defining patterns that can apply to sequences of things … and they are incorporated into most general programming languages' (Knox, 2013).

[7] There are many reasons for the current lack of interest in, even distrust of, quantitative methods among historians, but at least some of it may be traced back to the overblown claims made for the 'Cliometrics Revolution' of the 1960s and 1970s. The fall from grace experienced by Cliometrics was dramatic, and echoes of it may be seen in the heated debates that followed the publication of Michel *et al*. (2011), and reactions to 'Culturomics'. For a fuller discussion of these issues, see Winters (2018).

[8] Compare, for example, the order and structure of the Corpus of Global Web-Based English (Davies, 2014) with the disorder and relative chaos of the UK Web Archive.

[9] In the UK, the most resonant example is still that of the BBC Domesday project. This ambitious initiative to create a comprehensive digital record of life in Britain, 900 years after the compilation of the original Domesday Book, was an early and very high profile digital preservation failure. It cost £2.5 million, but the special BBC Micro computers required to read the 12-inch video disks rapidly became obsolete and the data was inaccessible just 15 years after it was collected (McKie and Thorpe, 2002). The contrast with its medieval – analogue – predecessor could not have been greater. What is much less remarked is that there were a number

of efforts to recover the data, and indeed in 2011 the contents of the 'Community' disk were published online as part of Domesday Reloaded (http://www.bbc.co.uk/history/domesday).

[10] The two interdisciplinary networks are: Born-digital Big Data and Approaches for History and the Humanities (https://borndigitaldata.blogs.sas.ac.uk/; grant reference AH/N006178/1) and Record DNA (https://recorddna.wordpress.com/; grant reference AH/P006868/1).

[11] RESAW (http://resaw.eu/).

[12] There are some exceptions to this. At the time of writing, the UK Parliament Web Archive (http://webarchive.parliament.uk/) only includes around 37 websites. Individually, some of these might be very large (the Hansard Archive from 1803 is one example), but the collection as a whole remains manageable and susceptible to qualitative analysis. By contrast, the annual domain crawl of the .uk country code Top Level Domain (ccTLD) undertaken by the British Library in 2014 captured 2.5 billion web pages and other assets (Hartmann, 2015).

[13] UK Web Archive: Browse by Special Collection (https://www.webarchive.org.uk/ukwa/collection/).

[14] Archive-It (https://archive-it.org/).

[15] 100,000 Poets (https://archive-it.org/collections/2845).

[16] There are, in fact, three special collections dealing with the Boston Marathon bombing: Blasts in Boston Marathon (https://archive-it.org/collections/3752); 2013 Boston Marathon Bombing (https://archive-it.org/collections/3649); and Boston Marathon Bombing: Twitter and RSS Feeds (https://archive-it.org/collections/3645). Hurricane Katrina warrants two: Hurricane Katrina (https://archive-it.org/collections/174); and Hurricane Katrina Blogs Web Collection (https://archive-it.org/collections/7625). This multiplicity of collections relating to single events is an additional problem for the historian, as the relationship between them is unclear. In the Boston example, two have been curated by the Virginia Tech: Crisis, Tragedy, and Recovery Network so there is presumably no overlap, but what of the collection curated by Internet Archive Global Events? Detailed comparison is required to establish the ways in which these three collections differ from or are similar to each other.

[17] Government special collection (https://archive-it.org/explore?show=Collections&fc=meta_Subject%3AGovernment). Examples of smaller collections on a similar theme are Government – US Federal (https://archive-

it.org/explore?show=Collections&fc=meta_Subject%3AGovernment-usfederal) and SF [San Francisco] Government (https://archive-it.org/explore?show=Collections&fc=meta_Subject%3ASF+Government).

[18] This is explicitly the case with the Live Art collection, which has been produced in partnership with the Live Art Development Agency in London (https://www.webarchive.org.uk/ukwa/collection/26312782/page/1).

[19] See, for example, the K-12 Web Archiving Program in the US, which at the time of writing has generated more than 300 student collections (https://archive-it.org/k12/). I owe this reference to Ian Milligan.

[20] Warcbase (https://github.com/lintool/warcbase) has been developed by Jimmy Lin at the University of Waterloo; Wget (https://www.gnu.org/software/wget/) is part of the GNU Operating System.

[21] Webrecorder (https://webrecorder.io/). Webrecorder has been developed by Ilya Kramer at Rhizome, with funding from the Andrew W. Mellon Foundation.

[22] Webrecorder Player 1.0.5 (https://github.com/webrecorder/webrecorderplayer-electron/releases/tag/v1.0.5).

[23] See Taylor, 2017, for a discussion of history from below and modern British social history.

[24] Just in the UK, a 2017 survey of digital skills published by the Tech Partnership, in association with Lloyds Banking Group, found that '21% (11.5m) of the UK are classified as not having Basic Digital Skills'. A heatmap of digital exclusion derived from the survey data reveals significant geographical variation (http://heatmap.thetechpartnership.com/). Even where relatively advanced digital skills are present, there are a range of other factors influencing whether or not people choose to participate in the creation of online content (see, e.g., Hargittai and Walejko, 2008; Blank, 2013).

[25] Alan in Belfast (https://www.webarchive.org.uk/ukwa/target/18710548/collection/100698/source/collection); Nelly's Garden (https://www.webarchive.org.uk/ukwa/target/7176221/collection/100698/source/collection); Lizzy's Literary Life (https://www.webarchive.org.uk/ukwa/target/65208425/collection/100698/source/collection).

[26] Weber sees 'the history of the online world as not just about the Web itself, or networks like the Internet, or computers, but as all of these within the long tradition of tools we have created for sharing and refining information: books and clay tablets and talking drums and more' (Weber, 2017: 34). Web90 – Patrimoine, mémoires et histoire du Web dans les années 1990, at L'Institut des sciences de la communication du CNRS, is an exemplary project in this respect (http://web90.hypotheses.org/).