12
Coda: Web archives for humanities research – some reflections

Jane Winters

Introduction

For historians, and researchers in many other humanities disciplines, web archives remain largely an unknown, and certainly underused, primary source. Even within digital humanities, web archives as a focus for study have remained on the fringe, much more likely to be represented on the programme at events such as the ACM Web Science conference than the Annual Conference of the Alliance of Digital Humanities Organisations (ADHO).[1] There are many possible reasons for this – the particular focus of digital humanities, for example on textual editing; the difficulties of gaining access to web archives within national libraries and archives; the real and perceived technical barriers to working with this material; the paucity of digital skills training in the humanities generally; and simply the natural length of time it takes for new ways of researching to emerge and be recognized – but it is nevertheless a problem which needs to be overcome.

It is hard to imagine how one might study the history of the developed world[2] in the late twentieth and early twenty-first century without recourse to the archived web.[3] The traditional tools of the historian's trade – newspapers, letters, diaries, the records of government and business – are commonly, and in some instances now solely, online.[4] Some of these have been transformed – think of the relationship between, and intended audience for, a paper diary and a blog – while others are broadly similar in form and purpose but the method of delivery and consumption has changed. Our primary sources are increasingly on the web, whether we like it or not, and this is a trend which is unlikely to be reversed any time soon.

And time is important in another sense. The web was 25 years old in 2014, and an archiving process has been in place for almost 20 years, when the Internet Archive in the USA began its invaluable work, acknowledged elsewhere in this volume. For contemporary historians at least, this is beginning to look like a reasonable chronological span. The UK has traditionally adopted a 30 Year Rule in relation to the public release of non-sensitive government records, but in 2013 The National Archives began a move towards releasing records when they are just 20 years old, that is, the same age as the earliest instances of archived websites. It is becoming increasingly hard to argue that this is not material worthy of historical study.

So why, then, do web archives remain so persistently underutilized, so hidden from the mainstream of historical and digital humanities research? It was this mismatch between the clear value of web archives – for modern cultural, economic, political, social and technical history – and low levels of usage and awareness that we set out to address in the Big UK Domain Data for the Arts and Humanities (BUDDAH) project which has led to this present collection of essays.[5] The substantive research conducted during the project is described by Josh Cowls in his chapter on 'Cultures of the UK web', but the generation of this series of case studies was just one element of the project. We were concerned with the incubation of a community of humanities researchers who would move on from the project to advocate for the importance of web archives within their host institutions and among their disciplinary peers. The intention was not to transform them into 'web researchers' but to equip them to use web archives, and to encourage others to do the same. For most humanities scholars it will be a very long time before they transition to using solely digital sources, let alone solely born-digital sources, and for many this will never be the case. They will continue to mix and match, to compare and contrast, and to work with overlapping sets of material which contain subtly different information and are designed for subtly different audiences. Their research, however, will be impoverished if they are unaware of what web archives may contain – even if it is only to discount that information as unhelpful or unreliable.

Combining old and new approaches?

This is not, of course, to understate the challenges posed by web archives, as highlighted by many of the chapters in this volume. There are challenges arising simply from scale, or from the nature of the archiving

process, but there are also new conceptual challenges that will require innovative approaches and ways of thinking. Some of the problems are very familiar. For example, it is difficult to ascribe a clear date of publication to an archived web page. Even if all of the elements on a web page were captured at the same time, the date associated with them marks the point at which they were archived rather than the point of their formal publication (however we might think of this in an age of limitless editing possibilities and multiple versions). This seems far removed from the publication of a modern printed book for example, which will have an apparently clear date listed in the preliminary pages, or prelims.[6] But it is not uncommon even today to see bibliographic citations for serial publications along the lines of '2013 (really 2015)'. This indicates a discrepancy between the scheduled or official publication date and the date on which the book actually appeared in print. Which is definitive? And will the answer be the same in 50 years' time? Medieval manuscripts may be datable to, for example, only a rough 25-year period. Uncertainty about date is something with which historians have always had to deal. Not so is the presence in the archive of a 'web page' which never actually existed. The memento protocol pieces together the 'best' view of a page, bringing together elements from different archives captured at different points in time. The British Library home page on 20 July 2009, for example, may be assembled using 14 mementos from four separate archives, spanning four months. Despite the superficial similarity of the process to critical or scholarly editing,[7] this is a new phenomenon which is embedded within the archive itself and not imposed subsequently by one or more human editors.

It is clear that humanities researchers need to acquire new skills and develop new methodologies if they are to get to grips with web archives as a source, but much can be achieved either by repurposing and adapting existing analytical frameworks or simply by approaching digital data with the same critical eye that one might bring to incunabula or to early modern newsletters. Eric Ketelaar, for example, has argued persuasively that diplomatic, traditionally applied to medieval documents, may also be useful for the analysis of digital materials like web archives: 'The principle of provenance and other basic tenets of archival science can be put to new uses in the digital age'[8] (Ketelaar, 2007: 167–91). Existing methodologies may be adapted to accommodate different data structures and different signifiers of purpose, authority and authenticity, in combination with new tools, approaches and theoretical frameworks.

This, however, is to take a primarily micro-historical approach to the study of web archives, to search for stories about particular individuals, institutions or events. There is scope for the macro-historical too, as championed recently by Jo Guldi and David Armitage in their call to arms, *The History Manifesto*. This deliberately provocative book, which includes a chapter titled 'Big questions, big data', argues that

> Together, micro-historical work in archives and macro-historical frameworks can offer a new horizon for historical researchers who want to hone their talents of judging the flow of events and institutions across centuries and around the globe as well as a new opportunity to engage with the public. (Guldi and Armitage, 2014: Conclusion)

The key point here, and one which has been overlooked by some commentators, is the combination of approaches – there is room for what Tim Hitchcock has described as 'beautiful histories of small things' (Hitchcock, 2014) but also for the historian's macroscope (Graham et al., 2015). The data in which humanities researchers are most interested is characterized by complexity and mess because it reflects and records complex and messy human interactions. Hitherto unsuspected patterns emerge when it is analysed at scale, but these can only be tested by digging in to the data and understanding the individual elements which make up the whole.[9]

Nowhere is this approach more apposite than when working with web archives, as evidenced by the research presented in this volume. The histories of individuals and organizations, at least as they played out online, can be traced over the past 20 years. Conclusions may be drawn about how the culture of an institution has evolved; how a government department has interacted with the public (and what information it has deemed to be most important to communicate at particular points in time); how a small business has expanded and/or contracted; how an individual has reflected on their journey through illness or on their family life. Alternatively, wider social and cultural changes may be traced through the online development of a single organization. How have changes in design and technology influenced a company's web presence? What has been the effect of developments in e-commerce on its online services? How has the increased penetration of the web into everyday life affected the language used to communicate with users and consumers? How, if at all, has it accommodated social media and the

growing customer expectation of increased interaction, sometimes in real time?

However, it is also possible to study wider patterns and trends, for example to attempt to trace developments in language, to undertake complex network analysis or to track the movements of peoples and political ideas. There is no need to rehearse again here the many difficulties posed by web archives for this kind of research, but the fact that it is challenging does not undermine its enormous potential value. Even a very simple n-gram approach can produce immediately suggestive results, for example when identifying neologisms and the point at which they become widely adopted. In the UK, the *Oxford English Dictionary* (*OED*) produces an annual 'word of the year', chosen because it has risen to prominence in the previous 12 months, or might in some way be said to characterize that period. A comparison between some recent *OED* choices and instances of those terms in the archive of UK web space for 1996–2013 reveals that the selections of the dictionary's experts are mirrored (driven?) by online trends. In 2004, the chosen word was 'chav',[10] and the trends graph developed for the British Library's 'Shine' interface reveals a clear spike in mentions of the term in that year. In 2003 it appeared just 923 times, but in 2004 this figure jumped to 60,467 (an increase of 6,551%). In 2008 'credit crunch' was nominated and the pattern in the web archive was very similar (even if the order of magnitude differed): in 2007 there were 128,152 instances of the phrase, while in 2008 this rose to 1,555,960 (an increase of 1,214%). Interestingly, the web archive indicates a rather different fate for these two 'words of the year': 'chav', perhaps rather unfortunately, persists, but there is a sharp drop in instances of 'credit crunch'[11] relative to the archive as a whole after 2009. It would seem that it was specific to a particular moment and set of circumstances, or at least as it was used on the web. These are, of course, very simple examples found using a simplistic methodology, but nonetheless interesting.[12]

Moving beyond text (and search)

The digital humanities embrace a wide range of methods and sources, but much of the most innovative work to date has been concerned with the analysis of text.[13] Web archives contain a great deal of text, from formal publications and newspapers to material verging on direct speech (some social media), but the data is distinguished by its variety.

There are varieties of textual information – html pages, MS Word documents, PDF files – but there are other media too – videos, image files, sound clips, animated gifs. The web is becoming an ever more visual medium, with the dominance of services like YouTube and Instagram and the ease with which photographs and video can be captured and uploaded to the web from smartphones. Much of this data is beyond the reach of web archives as they currently exist. This is either because it falls outside a nationally-harvested country code Top-Level Domain (ccTLD) or because it is the property of a commercial service provider like Facebook. There is, however, a great deal that falls within the scope of the archives. A British Library visualization of popular image formats in the archive of UK web space for 1996–2010, for example, reveals that JPEGs alone account for 10% of the total crawl in 2010 (the figure is roughly consistent across the whole period). The analysis of non-textual big data at scale is a significant challenge that will only become more pressing as born-digital data becomes a focus of research. Traditional image databases, like the John Johnson Collection of Political Ephemera or the Warburg Institute Iconographic Database, rely on the generation of exhaustive metadata to support discovery, but this is not present for the bulk of the films, images and sound clips in web archives. The problem is particularly acute for platforms and services where the addition of metadata is largely optional and almost entirely uncontrolled. A very simple example serves to illustrate the problem. It is a truism that the web is overrun with pictures of cats. Searching the 'Shine' interface for 'cat' and limiting results to the content type 'image' produces 340,453 instances of the term. This would, of course, not be a sensible search to conduct, as is clear from an investigation of the first few images listed: the initial four are blocked because of robots.txt; the fifth is indeed a photograph of cat; but the sixth is a pair of Caterpillar boots, the seventh a 'music catalogue' gif, and so on. It is here that existing methods of interrogating data begin to break down.

The dominance of (a particular type of) search as a digital research method very quickly becomes problematic for web archives where, quite apart from difficulties arising from scale, the scope of a particular archive is unknown and the process of creation largely undocumented. Discovering what might be in the archive is often the primary objective – and this is not well served by keyword searching which produces a list of results unordered by anything other than date. For sound and image, moving or still, there is the extra limitation of poor or non-existent metadata for even crude keyword searching. The absence of metadata is a limitation here in another way too. Images are one of the elements of

a web page which the crawl process is more likely to fail to capture, and the absence of metadata or alternative text confers invisibility. This may be seen, for example, in the capture of the home page of the Institute of Historical Research (IHR), University of London in the Internet Archive from 1 December 2003. There are three broken image links at the top left of the page, but the associated text makes it clear that they are two logos – one for the IHR itself and one for the 'History' website – and a picture of the building that hosts the institute. For other missing images on the page, however, there is no associated text so it can only be guessed what might have been used to illustrate, in this instance, training courses for Latin and for Palaeography and Diplomatic.

(In)completeness and loss

Web archives raise questions of (in)completeness. Should we be trying to keep everything, particularly as existing methods of selection and cataloguing are not scalable? If we do not know what future scholars will be interested in, should we simply collect it all? And what do we mean by 'everything', when the web archiving process is marked by patchy data collection and loss? Web archives are, after all, only an often partial snapshot in time. Notions of comprehensiveness exist simultaneously in our consciousness with the counter-narrative that we are about to enter or have already entered a 'digital dark age' which will see the historical record lost for future generations. Of course, neither of these is true, but questions of survival and loss do seem to loom particularly large in relation to born-digital data, including web archives.

This seems to me to be an old problem of the survival of evidence filtered through a new expectation that it is somehow possible, even desirable, to keep everything. This is to ignore the fact that the primary sources we value so much from earlier periods have in large part survived through historical accident. One particular monastic library burnt to the ground while another did not; one individual was more diligent at keeping her correspondence than another; one national archive was bombed during the Second World War while another was spared when an incendiary device failed to go off. Medieval historians, for example, become used to working with and around gaps, to speculating about the representative nature of a particular set of records, to trying to reconstruct a legal code from vague references to it in other documents. Perhaps the difference in focus comes from our ability to know precisely what we do not have when we are dealing

with web archives. A missing image confronts us with a blank square on the web page; a broken link produces an error. To take one example, the first capture of the IHR website in the Internet Archive dates from 27 December 1996, but the website went live on 9 August 1993 (Segell, 1993: 4). We are immediately confronted with the fact that more than three years' worth of data no longer exists. Data loss is also a very real presence in our daily lives, whether it is the disappearance of whole services which once seemed essential or the failure to back up a much-used computer.

The susceptibility of the web to archiving may, however, lead to other kinds of gaps. It is noted on the British Library website, for example, that 'Where […] web crawling software encounters a login facility, it cannot access any material behind the login facility without the appropriate password or access credentials'. In practice this means that data of this kind is not captured, so openly published information is privileged in the archive. This has potentially fascinating implications for what will remain available to researchers in 10, 15 or 50 years' time. In an interesting reversal of previous patterns of data survival, might open data be more likely to persist than commercially managed and published digital material?[14] Publishing companies are, of course, taking steps to ensure the long-term availability of their outputs, but they are often working outside the national infrastructures that underpin web archiving. Which is the more likely to last, if we accept that the digital presents a sustainability challenge? And what of apps,[15] which are largely closed systems unsusceptible to archiving by national institutions, let alone the wealth of data published via social media platforms such as Facebook. This is truly vulnerable information, reliant on the self-interest of corporations for its maintenance (Webster, 2015). Might we be forced to rethink what we consider to be ephemeral?

Unlocking value

The sheer variety of information contained in web archives poses huge difficulties for researchers, but this mixture of formats and types, of the personal and official, of the public and private is precisely why they are such an important primary source for humanities researchers. It is at once possible to compare, for example, the official announcement of a government policy with its subsequent coverage in newspapers and other online media, and then with its discussion in online forums and selected

social media. Perhaps the policy is a controversial one which results in the creation of an online petition, which in turn triggers a debate in parliament.[16] All of this information may be found in the archive, even if it is not in any sense comprehensive or indeed easy to locate. Our stories and our histories are increasingly online, but the inherent ephemerality of the live web means that they only achieve any degree of permanence in web archives.[17]

There are clear technical and methodological hurdles facing researchers who wish to study these histories, but simply gaining access to web archives introduces an additional layer of complexity. Over the past two decades and more, researchers have become used to the widening of access through digitization and the increased availability of digital materials online. Hierarchies, of course, remain – notably between those within and outside well-resourced universities – but nevertheless more people have greater access to the historical materials held in our national memory institutions than at any point in our history. And that access is often international – the selective open UK Web Archive, for example, can be viewed from anywhere in the world. But legal frameworks which have failed to keep up with changing technologies and modes of communication mean that artificial barriers are being erected around web archives which are preventing the integration of their study into the mainstream of humanities research. In the UK, access to archived websites and electronic publications is severely restricted by legal deposit regulations, with the result that 'deposited works may not be made available online externally, including for readers logging in remotely. They can only be viewed on the premises of the six deposit libraries'.[18] Moreover, 'the 2013 Regulations stipulate that "A deposit library must ensure that only one computer terminal is available to readers to access the same relevant material at any one time"' (Netarkivet, n.d.); in other words, two people may not look at the same instance of an archived web page concurrently. In other countries which archive their ccTLD, access may be even more restrictive.[19]

All of this is against a background of increased expectation not just of open access to data but that there will be APIs which allow researchers to download and take away the material with which they choose to work. It is the portability of data, its separability from an easy-to-use but necessarily limiting interface, which underpins much of the most exciting work in the digital humanities. Web-based tools such as Voyant have brought quite sophisticated textual analysis within the reach of anyone who has access to data, but data from web archives can be very hard to come by. When it is available, as with the host link graph derived

from the UK domain dataset 1996–2010 used by Meyer and colleagues to examine 15 years of UK universities on the web in this volume, the results are fascinating and suggestive of numerous avenues for research. Initiatives like the Common Crawl, which provides 'an open repository of web crawl data that can be accessed and analyzed by anyone', are doing important work here too. Problems of access are not, of course, unique to web archives, but if it is made too difficult for researchers to engage with the data, they will turn elsewhere or simply rule out using web archives as a source.

A perception of difficulty is most damaging for those who might study web archives as just one of a number of primary sources, including printed newspapers, the paper records of government, film and television, and other kinds of digitized data. They are not concerned with the history of communication or technology, but with what the archived web can reveal about the development of a popular political movement, health scare or terror attack. These are the researchers whose work on the BUDDAH project has been admirably synthesized by Josh Cowls and, as noted above, it is typified by a mixture of both methods and sources. They do not have the time, or indeed the willingness, to develop the full range of skills that might be expected from a specialist; nor do they commonly have access to the high performance computing facilities that working with web archives may require. They are likely, however, to be the key to increasing familiarity with and usage of the growing volumes of data that archiving institutions are collecting and storing, often at considerable expense in a time of generally straitened finances. That is where a volume of this kind, which showcases innovative research using web archives and presents a range of use cases for different humanities disciplines, is so useful. If the BUDDAH project is any indication, it often only takes dipping a toe in the water for researchers to discover the value of web archives.

If web archives need to be integrated into established processes and workflows in order to become widely consulted, they also need to be considered in debates about approaches to working with born digital big data more generally. National libraries and archives are not simply responsible for archiving the web; they are increasingly having to deal with email archives, with institutional and departmental file systems, and with personal digital data. The difficulties of storing, preserving and making available these different types of data vary, as do the problems facing researchers who wish to study them, but there are commonalities too, which go beyond mere scale. One such is the question of how you protect individuals in this mass of information. It is not just individual

pages or documents which may be sensitive but the combination of those pages and documents, or even very small snippets of data. These may reveal a larger picture or more information about someone than they would either anticipate or be comfortable with. The reverse is a problem too – how do you securely identify persons of interest in large-scale and complex data where diversity in naming is almost systemic? This is to return to the requirement for new theoretical and methodological frameworks identified above, which these chapters, and the explicit connections between them, are helping to advance. Other interdisciplinary and international forums and networks are developing to consider these questions and, as a crucial first step, to articulate which problems are common to many forms of born digital data and which relate only or primarily to web archives.[20] It is a thriving and vital (in both senses) field of research.

At present, and necessarily, scholarly debate has tended to focus on the impediments to working with web archives, and on the sheer effort involved in making sure that this data is captured effectively. This edited volume has sought to move the discussions on, to make available the first fruits of research – and in an open access form which introduces them to the widest possible audience. It is a starting point, a signpost to future interesting locations which may be reached by more or less circuitous routes. As the roadmap becomes clearer, and the data begins to be better understood, it is to be hoped that the enormous richness of the archived web will come increasingly to the fore. Like any new area of investigation, any new type of primary source, it takes time before its full potential is realized. These chapters are a first, and fascinating, indication of what we might expect.

Chapter 12

Big UK Domain Data for the Arts and Humanities (n.d.). Retrieved 9 May 2016, from http://buddah.projects.history.ac.uk/

Born Digital Big Data and Approaches for History and the Humanities (n.d.). Retrieved 25 May 2016, from https://borndigitaldata.blogs.sas.ac.uk/

British Library home page (20 July 2009). Retrieved 10 July 2015, from http://timetravel.mementoweb.org/reconstruct/20090720093000/http://www.bl.uk

Common Crawl (n.d.). Retrieved 13 May 2016, from http://commoncrawl.org/

Davies, M. (2013). Corpus of Global Web-based English (GloWbE). Retrieved 13 May 2016, from http://corpus.byu.edu/glowbe/

Depositing Websites and Web Pages (n.d.). Retrieved 25 May 2016, from http://www.bl.uk/aboutus/legaldeposit/websites/websites/

Digging into Data (n.d.). Retrieved 13 May 2016, from http://diggingintodata.org/

Digital Humanities 2016, Kraków 11–16 July (n.d.). Retrieved 9 May 2016, from http://dh2016.adho.org/program/

Diplomatics (n.d.). Retrieved 9 May 2016, from https://en.wikipedia.org/wiki/Diplomatics

Graham, S., Milligan, I. and Weingart, S. (2015). *Exploring Big Historical Data: The Historian's Macroscope*. London: Imperial College Press. Retrieved 13 May 2016, from http://www.themacroscope.org/2.0/

Guldi, J. and Armitage, D. (2014). *The History Manifesto*. Cambridge: Cambridge University Press. Retrieved 9 May 2016, from http://historymanifesto.cambridge.org/read/conclusion-public-future-past/

Hitchcock, T. (9 November 2014). Big data, small data and meaning. Retrieved 13 May 2016, from http://historyonics.blogspot.co.uk/2014/11/big-data-small-data-and-meaning_9.html

Host Link Graph: Jisc UK Web Domain Dataset (1996–2010) (n.d.). DOI: 10.5259/ukwa.ds.2/host.linkage/1

IHR-Info. Hypertext Internet Server (n.d.). Retrieved 13 May 2016, from http://web.archive.org/web/19961227133909/http://ihr.sas.ac.uk/

Internet Archive – About (n.d.). Retrieved 9 May 2016, from https://archive.org/about/

Jackson, A. (27 April 2015). Ten years of the UK web archive: what have we saved? Retrieved 13 May 2016, from http://www.slideshare.net/andrewnjackson/ten-years-of-the-uk-web-archive-what-have-we-saved

John Johnson Collection of Printed Ephemera (n.d.). Retrieved 13 May 2016, from http://www.bodleian.ox.ac.uk/johnson

Ketelaar, E. (2007). Archives in the digital age: new uses for an old science. *Archives & Social Studies: A Journal of Interdisciplinary Research* 1: 167–91.

The National Archives (n.d.). *20-year rule*. Retrieved 9 May 2016, from http://www.nationalarchives.gov.uk/about/our-role/plans-policies-performance-and-projects/our-projects/20-year-rule/

The National Archives (2016). *The Digital Landscape in Government 2014–15*. Kew: The National Archives. Retrieved 25 May 2016, from http://www.nationalarchives.gov.uk/documents/digital-landscape-in-government-2014-15.pdf

Netarkivet (n.d.). Retrieved 13 May 2016, from http://netarkivet.dk/in-english/

Newton, C. (n.d.). Life and death in the app store. *The Verge*. Retrieved 9 May 2016, from http://www.theverge.com/2016/3/2/11140928/app-store-economy-apple-android-pixite-bankruptcy

Petitions: UK Government and Parliament (n.d.). Retrieved 13 May 2016, from https://petition.parliament.uk/

Research Infrastructure for the Study of Archived Web materials (n.d.). Retrieved 25 May 2016, from http://resaw.eu/

Segell, G. (1993). *Guide to IHR-Info: Hypertext Internet Server*. London: Institute of Historical Research.

'Shine' image search for 'cat' (n.d.). Retrieved 13 May 2016, from https://www.webarchive.org.uk/shine/search?query=cat&tab=results&action=search&facet.in.content_type_norm=%22image%22

'Shine' trend graph (n.d.). Retrieved 13 May 2016, from https://www.webarchive.org.uk/shine/graph

Sweney, M. (5 May 2016). *The New Day newspaper to shut just two months after launch*. Retrieved 9 May 2016, from http://www.theguardian.com/media/2016/may/04/new-day-newspaper-shut-two-months-launch-trinity-mirror
Text Encoding Initiative (n.d.). Retrieved 13 May 2016, from http://www.tei-c.org/index.xml
UK Web Archive (n.d.). Retrieved 13 May 2016, from http://www.webarchive.org.uk/ukwa/
UK Web Format Profile 1996–2010 (n.d.). Retrieved 13 May 2016, from http://www.webarchive.org.uk/ukwa/visualisation/ukwa.ds.2/fmt
Voyant Tools (n.d.). Retrieved 13 May 2016, from http://voyant-tools.org/
Warburg Institute Iconographic Database (n.d.). Retrieved 13 May 2016, from http://warburg.sas.ac.uk/vpc/VPC_search/main_page.php
Webster, P. (29 June 2015). Why hoping private companies will just do the Right Thing doesn't work. Retrieved 25 May 2016, from https://peterwebster.me/2015/06/29/why-hoping-private-companies-will-just-do-the-right-thing-doesnt-work/