Negotiating the archives of UK web space

Jane Winters, School of Advanced Study, University of London

Introduction

Establishing the size and shape of a national web sphere poses enormous technical challenges, as other chapters in this volume have demonstrated (and see, for example, Brügger, 2017a), but these are not the only hurdles facing a researcher who wishes to study the archived web from a primarily national perspective. Web archives, like their more traditional counterparts, reflect the social, political, cultural and economic contexts within which they are formed. And these contexts are neither stable nor purely national; they are influenced over time by changes in legislation, by shifting organisational priorities, by fluctuations in funding, by new governments, even by the interests of individuals. This chapter will discuss what this has meant, and continues to mean, for the archives of UK web space.

As the title of this chapter suggests, there is no single archive of the .uk country code Top Level Domain (ccTLD). Rather there are many archives, which overlap and diverge in multiple and still largely unexplored ways. Data collected by the Internet Archive (IA) from December 1996 to the present day is accessible via the Wayback Machine, where both URL and limited keyword searching are available;[1] and a substantial subset of this data, from 1996 to April 2013, is fully searchable through the Shine interface developed and hosted by the British Library.[2] There are two other collections at the British Library: material crawled[3] since April 2013 in accordance with legal deposit legislation, to which at present there is only limited on-site access; and the open, selective UK Web Archive.[4] The UK Government Web Archive at The National Archives (TNA) is also open but is concerned solely with the online presence of government, while the UK Parliament Web Archive takes a similarly selective approach. Data for the .uk ccTLD made available by the Common Crawl

Foundation has a different profile again.[5] Finally, and perhaps counter-intuitively, substantial elements of the UK web have found their way in to other national web archives, as a result of the idiosyncrasies of the crawl processes.[6]

This is both a dauntingly complex landscape for a researcher and a significant problem for archiving institutions trying to promote usage of their web archives. Any discussion of what is available has to start with caveats and exceptions, with apologies for the greater access restrictions placed on one type of content but not another, with explanations of selective versus broad crawling. Defining the relationships within and between these differing archives is an essential step towards any kind of quantitative research using web archives and to encouraging greater use of this fascinating new primary source. Quite simply, where does the new researcher – even one with the requisite degree of technical knowledge – start?

Temporality and the archived web

The first problem to be faced is that of temporal coverage, although this is also deeply connected to data provenance. The invaluable work of the IA underpins many of these UK collections, since its activities pre-date all British web archiving programmes.[7] Where collections extend back to December 1996, this data will always have been derived from the IA. It will not, however, always – indeed ever – be the same. Data made available by the IA to organisations such as the British Library and TNA was prepared at different times, using different criteria and methods, and patching different gaps in institutional holdings. In both of these cases temporal coverage has been retrospectively enhanced and the data re-presented in what are effectively new content silos. By contrast, the earliest content in the UK Parliament Web Archive dates only from July 2009; and the special collections in the open UK Web Archive document nothing before 2005.

Attempts to extend temporal coverage have produced what might be described as patchwork collections: different datasets are stitched together in order to produce something which is greater than the sum of the parts. This stitching is largely hidden from the user, but it remains an important characteristic of the larger archives of the UK web. Inconsistency of selection and capture is thus not accidental but central to the nature of these collections.[8] This is nicely illustrated by the history of the UKGWA. TNA's archiving of government digital information began in 2003, when it partnered with the IA to harvest around 50 websites (it gained access to the IA's back catalogue from 1996 at the same time). Between 2004 and 2009, some government sites were selectively archived using infrastructure developed by the UK Web Archiving Consortium (UKWAC),[9] but since 2005 the main work of archiving has been undertaken on TNA's behalf by a third-party supplier, the Internet Memory Foundation (The National Archives of the UK, n.d. (a)).[10] From July 2017, a new supplier, MirrorWeb, will take over responsibility. Complexity is layered on complexity. Transparency, admirable though it may be, only gets you so far in unpicking the implications of all of this.

Issues of scope and size

The researcher next has to contend with the scope of these archives, which in turn determines their size. The largest UK archive of all is the Web Domain Crawl at the British Library, which aims to capture all UK websites at least once a year. The first domain crawl, which began on 8 April 2013, took almost eleven weeks to complete. Starting from a seed-list of 3.8 million URLs, it resulted in the capture of 1.9 billion web pages and other digital assets, amounting to 31TB of data (Webster, 2013; Hartmann, 2015). Just a year later, the annual crawl captured 2.5 billion web pages and generated 56TB of data (Hartmann, 2015), a rate of growth which is likely at least to be sustained in subsequent years. This is big data collected at the level of the nation. The UKGWA, by contrast, consists of more than 3,000 websites, as well as Twitter and video archives (The National Archives, 2016). At this scale it is possible to offer an A-Z browse list of sites, albeit a long and rather unwieldy

one, and even to organise the archived websites into broad thematic categories: business, industry, economics and finance; central and regional government; culture and leisure; environment, and so on. This is not, however, infinitely scalable. Even with a collection of around 3,000 sites, browsing is already of most use to the researcher who has a clear sense of what they might be looking for, supplemented by an understanding of the structure of government. At the other end of the spectrum is the UK Parliament Web Archive, which takes three snapshots per year of around 30 websites (the archive includes 37 sites in total, some of which are no longer on the live web) (Parliamentary Archives, n.d.). This is an eminently browseable collection, welcoming to the casual user and easy to comprehend. It is a very different proposition from the vast UK Domain Crawl, which overwhelms with volume and offers no simple point of entry. Browsing may begin to break down with more than 3,000 archived websites, but familiar search options start to fail when the numbers reach the billions (or even millions).

Access to web archives

This leads on to the question of access, a deceptively simple word which encompasses a range of different expectations and perceptions. Four distinct types of access will be considered here: first, the ability merely to read an archived web page; second, the opportunity not just to study individual pages or even whole websites, but to download and manipulate data; third, the availability of appropriate archival frameworks and tools to facilitate navigation and research; and finally the access facilitated by publication of research findings and data. The distributed nature of the archives of the UK web has already been discussed, but this is not the only factor determining ease of access (or otherwise).

<u>Accessing the archived web page</u>

The example of the British Library reveals that differing physical and virtual access restrictions may apply even within the collections of a single organisation. The British Library's annual domain crawl is subject to Non-Print Legal Deposit legislation,[11] which effectively considers an archived web page to be equivalent to a printed book, in that it must be read in the library building by a single user. This leads to a situation where each of the UK's six legal deposit libraries 'must ensure that only one computer terminal is available to readers to access the same relevant material at any one time' (The Legal Deposit Libraries (Non-Print Works) Regulations 2013, sect. 23).[12] The frustrations inherent in this type of restricted access are well described by Milligan (2015), who also notes, for example, the disabling of dynamic content and the necessity of printing out rather than photographing pages (or indeed copying and pasting text). Of course, the ability to access archival material in any form within six months or so of its deposit is a huge boon to researchers,[13] particularly given the mayfly-like existence of much web content.[14] One is, however, necessarily confronted with the reduction in functionality, in breadth and quality of access. This problem is not unique to the UK, but it is exacerbated by the fragmented nature of its national web archiving activity, whether that fragmentation is the result of unwieldy legislative frameworks or institutional histories.

Analytical access

The various 'open' web archives in the UK do not present this problem to the researcher – a single web page may easily be read by a researcher based anywhere in the world – but they do share some of the closed nature of the legal-deposit collection. Reading a web page is one thing, analysing a corpus of web pages is quite another. At present, none of the organisations responsible for archiving the web in the UK allows the bulk download of data: qualitative research approaches are built in to the existing discovery and delivery systems. This acts as a brake on truly innovative research use, and places the archiving institutions themselves under increasing pressure to develop tools that can be used on-site or in-browser. It is a source of frustration to a handful of researchers currently, but as

interest in born-digital data develops, current means of access are likely to come under increasing pressure.

Archival discovery: tools and methods

Appropriate tools, whether provided by libraries and archives themselves or developed independently, are clearly vital if researchers are to interact with web archives in more sophisticated ways than at present. There is, though, another stepping stone which receives much less attention – that is, how do researchers find relevant content, or indeed become aware of web archives at all, through their routine use of an archival catalogue? For institutions which undertake web archiving on any kind of scale, the volume of data collected is such as to overwhelm other forms of catalogued material. The British Library's main collections may include over 150 million items (British Library, n.d. (a)), but there are billions of URLs in just a single annual domain crawl. It is possible to provide a full-text search of this vast quantity of data, but it is not susceptible to traditional forms of cataloguing. How, then, to bring together the library's established finding mechanisms with this enormous full-text index, or even with a database of page titles? How to expose the casual researcher to the wealth of information the web archive contains? As Jackson (2017b) notes, for the British Library, 'The question is not "how [do] we collect these documents?" rather "how [do] we find the documents we've already collected?"'

Websites do appear in British Library catalogue search results, but currently in only a very limited way. To take one example, a search for 'terrorism' produces 37,754 results, of which just 27 are described as 'archived websites'.[15] The open UKWA alone has a special collection of 65 archived sites relating to the London terrorist attacks of 7 July 2005, while a search of the Shine dataset produces more than 11 million results. This is obviously not comparing like with like – the catalogue search is simply returning 'titles' which contain a particular word – but it does reveal the problems faced by both researchers and archivists/librarians in finding and exposing archived websites respectively. A

website ‘title’ does not perform the same function as the title of a book or journal article,[16] but as yet there is no intermediate step between this highly limited search and the overwhelming full-text option. In order to meet this challenge, cataloguing processes are having to be reimagined: ‘instead of thinking in terms of chains of events, we’re thinking in terms of layers of information’ (Jackson, 2017b). Connecting these ‘layers of information’ holds out huge promise for the integration of web archives and other born-digital data into library catalogues, and thus into the everyday business of the library or archive. In this scenario, web archives are no longer ‘other’, hidden from researchers and lurking in the peripheral vision of the librarian; they are just another type of primary source.

What this means for researchers is already apparent with regard to the UKGWA. This is an order of magnitude smaller than the British Library legal-deposit collection, and consequently it has been possible fully to integrate websites within TNA’s Discovery service. A description for a website like ‘About my vote’ appears alongside entries for nineteenth-century census returns or a thirteenth-century court roll. This catalogue equivalence is significant. For example, the ninth result in a search for the ‘Ministry of Defence’ is for the archived MOD website; it is recognisably related to paper records generated by the MOD Chiefs of Staff Committee and others.[17] Here, the catalogue is beginning to erase the artificial divide between print and born-digital, and to make the work of navigation easier. Accommodating hybridity is likely to be one of the key challenges for libraries and archives for decades to come, as paper ‘business as usual’ continues alongside the deposit of ever more born-digital data. And it will need to be accommodated, if researchers are successfully to study histories of people and institutions which exist in digital and paper forms, on the archived web and in archive boxes.

Communication, sharing and reuse

The final type of access to be considered is that promoted by publication, of both research findings and data. One of the main ways in which researchers will begin to negotiate the archives of UK web

space is by following the paths illuminated by others. The literature is finally beginning to emerge, with the recent launch of the journal *Internet Histories*, for example, providing a much-needed outlet for humanities researchers keen to explore the recent digital past. The history of the UK web, among others, is addressed in Brügger and Schroeder (2017), which seeks to reach the widest possible audience through open-access publication. Yet this openness is in marked contrast to the options for publishing and re-publishing data derived from web archives, whether a single screenshot or a curated dataset. And again, it is legal frameworks and nested legislation which hinder our ability to explore and comprehend the histories of the UK (and other) webs. In the UK, it is the venerable Crown copyright which offers hope to the researcher,[18] although even here the position is not as clear-cut as it might first appear. Guidance provided to users of the UKGWA states that:

> Most, *but not all*, of the websites accessible through the UKGWA were created by Crown bodies and are Crown copyright. Most of the archived content of these websites and services is also Crown copyright. Unless otherwise stated, you may re-use Crown copyright material obtained from the UKGWA freely under the terms of the Open Government Licence. (my italics)

This is a huge benefit to the researcher, and more or less as good as it gets when dealing with web archives, but there is still a warning that some third-party content may exist in the archive, that it may not be clearly identified as such, and that 'It is your [the researcher's] responsibility to ensure that you have any necessary permission for the re-use of copyright material obtained from the UK Government Web Archive' (The National Archives, n.d. (a)). There remains some uncertainty, but in reality, the reasonably diligent researcher should have nothing to fear about publishing and re-using content from the UKGWA.

But what about the collections at the British Library? Sites accessible via the open UK Web Archive have been included with the express permission of the content-owner or copyright-holder, but it is

not clear if this permissiveness extends to re-publication by a third party. In the absence of public guidance, the cautious researcher has to assume that they are not entitled, for example, to reproduce images of archived web pages without first seeking additional permission. This is certainly the case for the data collected as part of the annual domain crawl. As with any primary source, selective quotation is possible so long as the source is duly acknowledged, but nothing more than that. This problem is not, of course, unique to web archives – art historians, for example, struggle equally with image reproduction – but it is compounded by the nature of the data. Printed books may be 'orphan works', with unknown or untraceable rights holders, but generally there will be some indication of an author and/or publisher. Trying to find the owner of a 20-year-old archived blog is more of a challenge, and one which is multiplied across the bulk of any web archive corpus. If the true potential of these vast archives is to be unlocked, researchers will want to create and share their own datasets, to reproduce multiple images of web pages, to work freely with the data. The inability to do this easily may become more and more of a disincentive to work with web archives, to take time to navigate and learn about them, particularly as open data and the reproducibility of research increasingly come on to the agenda in UK higher education. There is a danger, too, that barriers to certain kinds of use will distort the type of research that is undertaken, away from small-scale storytelling towards large-scale analysis, from the micro decisively to the macro. It is in the combination of the two that the humanities have the most to contribute to our understandings of the digital world.

Providing context

The British Library is all too aware of these problems, and has done admirable work to assist researchers by providing them with some context for the web archives that it hosts. While the content itself cannot be made fully open and reusable, data derived from that content certainly can. The British Library's open data hub, data.bl.uk, contains five datasets related to the UK Web Archive,

here used as an umbrella term to describe all of the library's web archive materials. One is derived from the open selective web archive, and indicates the subject classification associated with each archived website. Far more interesting, for the purposes of defining the shape of the historical .uk ccTLD, are the four datasets that relate to post-1996 data supplied by the Internet Archive. Three – a format profile, a geoindex and a host link graph – deal with the period 1996-2010, while the fourth – an index of crawled URLs – provides coverage from 1996 up to the date of the first annual domain crawl in 2013. These are dauntingly large files, accompanied by warnings about 'extreme sizes' and the need to use 'a dedicated zip archive application',[19] but they are an enormously rich source of information and repay time spent in exploration.

While always bearing in mind that the available data is derived from an archived collection of web pages rather than perfectly delineating the UK web as it existed at a particular point in time, certain broad trends can begin to be identified. Taking the first three years for which data is available, 1996-1998, these trends are indeed best understood as reflecting the nature of the crawl process rather than the growth of the web. For example, 878,614 unique URLs are recorded for 1996; for 1997, there are 7,830,448; and for 1998, this figure falls back to 5,893,056. The Internet Archive was only launched in April 1996 (Kahle, 1997) and the crawling process did not get underway until September of that year (Rosenzweig, 2003, p. 749), so the dataset for 1996 is bound to be much smaller than for subsequent years. Similarly, it is also much more likely that there was a step-change in archiving activity in 1997 and a relative reduction the following year, rather than that the UK web contracted by almost 25% in 1998. However, even with these caveats, glimmers of insight begin to emerge. The numbers may be small, but the presence of CSS files clearly reflects a change – there are only two such files in the 1996 dataset, 32 in 1997 and 221 in 1998. This coincides with the emergence of CSS in late 1995 and the W3C's publication of the first standard on 17 December 1996 (Bos, 2016). The growth in 1998 is particularly striking given the overall reduction in the number of unique URLs compared to the previous year. My own investigation of these datasets is still at an early stage, but Peter Webster, for example, has already made good use of the host link graph data to explore online

connections among churches on either side of the border between Northern Ireland and the Republic.

Historians are used to working with multiple sources to piece together the stories that they want to tell, and this holds as true for the history of the UK web as for any other. The archived web at least partially documents its own development, as the records of bodies like Nominet, the official registry for .uk domain names, are preserved within it. Details of domain name registrations are available for each month from August 1996 to 2008, broken down first into six categories (.co.uk, .org.uk, .ltd.uk, .plc.uk, .net.uk and .sch.uk) and then from 2002, seven (with the addition of .me.uk).[20] Comparable data was published in subsequent years, but unfortunately there are significant gaps in the Internet Archive from 2009 onwards. There will be other sources, both on- and offline, which can be used further to round out the picture, but certain aspects of the story can already be outlined. For example, the Nominet figures reveal the very limited appeal of .plc.uk and .net.uk throughout the period, as well distinct peaks in applications for .sch.uk domains in August 2000 to January 2001 (an average of 1,938 a month) and in March 2004 (2,263) (see Figure 1). Figures without context, of course, can only get the researcher so far. Why, for example, were no .sch.uk domains registered between January and July 2000, and does this explain the relative surge in applications in the following six months?

It is this wider context that it is harder to retrieve, the information that may well not be contained in any web archive. Weber (2017) notes the pressing need to preserve 'enough historical materials about earlier [web and internet] systems to be able meaningfully to examine them', and this is as true for administrative as it is technological systems. Negotiating the archives of UK web space requires understanding how the archives themselves have been constructed, but also how those archives relate to the once live web of which they are a pseudo-facsimile. And this relationship is neither simple nor linear. Oral histories may be one means to uncover and preserve missing contexts, so too may be the application of ethnographical approaches.[21] The creators and users of

websites, both new and old, are still available to be interviewed; as are the librarians, archivists and technologists whose knowledge, expertise and, ultimately, decisions have shaped and will continue to shape how we will be able to access the historical web now and in decades to come.

Conclusion

Throughout this chapter I have used the term web archive without any kind of qualification, but perhaps part of our difficulty in dealing with this new source is that web archives are not really archives at all, at least as they have been traditionally understood. Owens (2014), for example, suggests that a web archive is 'much more in keeping with the computing usage of archive as a back-up copy of information then the disciplinary perspective of archives'. There is no common agreement that national web archiving activity should be undertaken by archival institutions rather than by libraries: in the UK, as noted above, it is divided between the two sectors; in France the responsibility is shared by the Bibliothèque national de France and the Institut national de l'audiovisuel; while in Denmark it is the work of the Royal Danish Library. It is not just in the UK that the picture is confused. Brügger (2017b, p. 187) has recently coined the term 'webrary', arguing that 'the advent of the Web is challenging and blurring the fundamental distinction between archives and libraries that has prevailed for centuries, although he acknowledges that it is probably now too late to change the 'terminological inconsistency'. Is it any wonder that we are struggling?

For the moment at least, web archives fit uneasily within our familiar information landscape; and, as set out in this volume's introduction, they have not yet fallen naturally within the purview of disciplines like digital humanities, which might have been expected to address some of the challenges described in these pages. They vary in crucial aspects both from the live web and from other archived sources; they are born digital, but access is subject to physical constraints; they duplicate and overlap with each other in ways that are difficult to determine; they cut across but also reflect institutional and national boundaries; they challenge traditional archival and research

practice. A national web sphere evades easy delineation, but as researchers we are fundamentally concerned to define our source materials, to determine boundaries and edges, to know what it is that we are reading or analysing. With a few exceptions, web archiving is conducted on a national basis by major national institutions, in keeping with well-established missions to preserve national cultural heritage. This collection of essays assumes that there is value in studying national webs in historical perspective, however difficult that might be. Taking the time and making the effort to negotiate the data that we have, wherever it may be held and whatever the restrictions on access, is the only way we can begin to describe and understand (digital) life in the UK, and in many other nations, over the past 20 years. Each national web sphere will have its own particularities and peculiarities, but there will also be commonalities and lessons to be learnt and shared, and the course will be easier to chart in the future.

Web archives are complex things, and this complexity can lead to an over-emphasis on problems. There is a temptation as a researcher to focus not on what can be done, but on what cannot; to become annoyed at the slow pace of change, both legal and technological; to concentrate on gaps and absence. Highlighting challenges is essential, and provides evidence for those individuals and organisations lobbying for the easing of access restrictions in particular, but it is even more important to demonstrate the importance of national (and international) efforts to archive the web. Researchers in the UK can add their voices to national (and international) conversations about the archiving and preservation of born-digital data, working with archivists and librarians to build on the work that I have described in this chapter.

It is to be hoped that, eventually, it will become easier to use the archives of the UK web, as we develop new tools and methods and as legal frameworks mature. In the meantime, while it may indeed be a source of frustration, perhaps the diversity of these web archives is also something to be celebrated. We might learn to be grateful for the fact that a web archive ‘is generated by an entangled and iterative system comprised of proactive human contributions, routinely operated

crawls and … external, crowd-sourced knowledge devices' (Ben-David and Amram, 2018),[22] rather than representing a single point of view or revealing the political and financial circumstances of a single institution at one point in time. The duplication within and between different web archives that has previously been viewed as problematic for most research purposes has only very recently begun to be considered as important, if not essential, in determining the value of the archived web as legal evidence (Nelson, 2018). Things change. What remains constant, however, is that national web archives embody the diversity of the nations whose stories they preserve, and this makes them both difficult and enormously rewarding to navigate.

References

Bailey, S. and Thompson, D. 2006. UKWAC: building the UK's first public web archive, *D-Lib Magazine*, 12:1, viewed 17 May 2017, http://www.dlib.org/dlib/january06/thompson/01thompson.html.

Ben-David, A. and Amram, A. 2018. The Internet Archive and the socio-technical construction of historical facts, *Internet Histories*, 2:1-2, viewed 28 May 2018, https://www.tandfonline.com/doi/citedby/10.1080/24701475.2018.1455412.

Bos, B. 2016. A brief history of CSS until 2016, *W3C*, viewed 6 June 2017, https://www.w3.org/Style/CSS20/history.html.

British Library n.d. (a). Facts and figures, viewed 5 June 2017 http://www.bl.uk/aboutus/quickinfo/facts/.

British Library n.d. (b). *UK Web Archive*, viewed 6 June 2017, https://www.webarchive.org.uk/.

British Library n.d. (c). London terrorist attack 7th July 2005, *UK Web Archive*, viewed 6 June 2017, https://www.webarchive.org.uk/ukwa/collection/100757/page/1.

British Library n.d. (d). *Shine: UK Web Archive*, viewed 6 June 2017, https://www.webarchive.org.uk/shine.

British Library n.d. (e). data.bl.uk – UK Web Archive, viewed 6 June 2017, https://data.bl.uk/UKWA/.

British Library n.d. (f). *Frequently Asked Questions for Webmasters*, viewed 28 May 2018, http://www.bl.uk/aboutus/legaldeposit/websites/websites/faqswebmaster/.

Brügger, N. 2017a. Probing a nation's web domain: a new approach to web history and a new kind of historical source. In: Goggin, G. and McLelland, L. ed. *The Routledge Companion to Global Internet Histories*. New York/Abingdon: Routledge, pp. 61-73.

Brügger, N. 2017b. Webraries and web archives – the web between public and private. In: Baker, D. and Evans, W. ed. *The End of Wisdom? The Future of Libraries in a Digital Age*. Cambridge, Mass.: Chandos Publishing, pp. 185-90.

Brügger, N. and Schroeder, R. ed. 2017. *The Web as History: Using Web Archives to Understand the Past and the Present*. London: UCL Press, viewed 6 June 2017, http://discovery.ucl.ac.uk/1542998/1/The-Web-as-History.pdf.

Brügger, N. 2011. Digital history and a register of websites: an old practice with new implications. In: Park, D., Jankowski, N. and Jones, S. eds. *The Long History of New Media: Technology, Historiography and Contextualizing Newness*. New York: Peter Lang Publishing, pp. 283-98.

Common Crawl Foundation n.d. *Common Crawl*, viewed 28 May 2018, http://commoncrawl.org/.

Goel, V. 2016. Beta Wayback Machine – now with site search!, *Internet Archive blogs*, viewed 17 May 2017, https://blog.archive.org/2016/10/24/beta-wayback-machine-now-with-site-search/.

Hartmann, S. 2015. 2015 UK domain crawl has started, *UK Web Archive blog*, viewed 17 May 2017, http://blogs.bl.uk/webarchive/2015/09/2015-uk-domain-crawl-has-started.html.

Hines, C. 2000. *Virtual Ethnography*. London: Sage Publications Ltd.

Jackson, A. 2017a. Revitalising the UK Web Archive, *UK Web Archive blog*, viewed 12 June 2017, http://blogs.bl.uk/webarchive/2017/06/revitalising-the-uk-web-archive.html.

Jackson, A. 2017b. The Web Archive and the Catalogue, *andrew.n.jackson blog*, viewed 30 June 2017, http://anjackson.net/2017/06/28/waw-the-shelves-and-the-mine/.

Kahle, B. 1997. Preserving the Internet, *Scientific American*, viewed 6 June 2017, https://www.scientificamerican.com/article/preserving-the-internet/.

Koerbin, P. 2017. Revisiting the World Wide Web as artefact: cases studies in archiving small data for the National Library of Australia's PANDORA Archive. In: Brügger, N. ed. *Web 25: Histories from the First 25 Years of the World Wide Web*. New York: Peter Lang Publishing, pp. 191-206.

The Legal Deposit Libraries (Non-Print Works) Regulations 2013, viewed 5 June 2017, http://www.legislation.gov.uk/uksi/2013/777/introduction/made.

Milligan, I. 2015. Web Archive legal deposit: a double-edged sword, *Ian Milligan's blog*, viewed 5 June 2017, https://ianmilligan.ca/2015/07/14/web-archive-legal-deposit-a-double-edged-sword/.

The National Archives of the UK 2016. *The UK Government Web Archive: guidance for digital and records management teams*. London: The National Archives, viewed 17 May 2017, http://nationalarchives.gov.uk/webarchive/web-archiving-technical-guidance.pdf.

The National Archives of the UK n.d. (a). *Information on web archiving*, viewed 17 May 2017, http://www.nationalarchives.gov.uk/webarchive/information.htm.

The National Archives of the UK n.d. (b). *20-year rule*, viewed 5 June 2017, http://www.nationalarchives.gov.uk/about/our-role/plans-policies-performance-and-projects/our-projects/20-year-rule/.

The National Archives of the UK n.d. (c). *UK Government Web Archive*, viewed 6 June 2017, http://www.nationalarchives.gov.uk/webarchive/.

The National Archives of the UK n.d. (d). *Crown copyright*, viewed 30 June 2017, http://www.nationalarchives.gov.uk/information-management/re-using-public-sector-information/uk-government-licensing-framework/crown-copyright/.

Nelson, M. L. 2018. The Internet Archive can't preserve the web's history by itself, *Motherboard*, viewed 28 May 2018, https://motherboard.vice.com/en_us/article/7xdn8y/joy-reid-and-the-weaponization-of-internet-archives.

Nyhane, J. and Flinn, A. 2016. *Computation and the Humanities: Towards and Oral History of Digital Humanities*. Springer International Publishing AG.

Owens, T. 2014. What do you mean by archive? Genres of usage for digital preservers, *The Signal*, viewed 12 June 2017, https://blogs.loc.gov/thesignal/2014/02/what-do-you-mean-by-archive-genres-of-usage-for-digital-preservers/.

Parliamentary Archives n.d. Web Archive FAQs, *Parliament's Web Archive*, viewed 17 May 2017, http://www.parliament.uk/business/publications/parliamentary-archives/search-finding-aids-to-online-collections-/web-archive/web-archive-faqs/.

Pennock, M. 2013. *Web-archiving: DPC technology watch report 13-01 March 2013*. Glasgow: Digital Preservation Coalition, viewed 17 May 2017, http://www.dpconline.org/docman/technology-watch-reports/865-dpctw13-01-pdf/file.

Rosenzweig, R. 2003. Scarcity or abundance? Preserving the past in a digital era, *American Historical Review*, 18:3, pp. 735-62.

Taylor, N. 2011. The average lifespan of a webpage, *The Signal*, viewed 5 June 2017, https://blogs.loc.gov/thesignal/2011/11/the-average-lifespan-of-a-webpage/.

Van de Sompel., H., et al. 2009. Memento: time travel for the web. arXiv:0911.1112v2 [cs.IR].

Weber, M. 2017. A common language, *Internet Histories*, 1:1-2, viewed 12 June 2017, http://www.tandfonline.com/doi/abs/10.1080/24701475.2017.1317118.

Webster, P. 2017. Users, technologies, organisations: towards a cultural history of world web archiving. In: Brügger, N. ed. *Web 25: Histories from the First 25 Years of the World Wide Web*. New York: Peter Lang Publishing, pp. 175-90.

Webster, P. 2013. Crawling the UK web domain, *UK Web Archive blog*, viewed 17 May 2017, http://blogs.bl.uk/webarchive/2013/09/domaincrawl.html.

Winters, J. Electronic Legal Deposit, web archives and researchers. In: Gooding, P. and Terras, M. eds. *Electronic Legal Deposit: Shaping the Library Collections of the Future*. London: Facet Publishing, 2018.

Acknowledgements

I am very grateful to Nicola Bingham, Helena Byrne, Andrew Jackson and Jason Webber at the British Library, for their insights into the workings of the UK Web Archive. Thanks are also due to the anonymous reviewer of this chapter for their comments and suggestions.

[1] A beta version of the Wayback Machine was launched in October 2016, supplementing the longstanding URL search facility with keyword searching on website home pages only (Goel, 2016).

[2] The Shine interface was developed as part of the Big UK Domain Data for the Arts and Humanities project, funded by the UK's Arts and Humanities Research Council (Grant reference AH/L009854/1). The project was a collaboration between the School of Advanced Study, University of London, the British Library, the Oxford Internet Institute and Aarhus University.

[3] A range of different terms are used for the various processes by which material from the live web is captured for more detailed analysis – harvesting, scraping, etc. – but web archives are created through crawling. The British Library defines web crawling as 'an automated process used to collect content and metadata that is available without access restriction on the open web' (British Library, n.d. (f)). Merriam-Webster dates the first use of 'web crawler' to 1994.

[4] At the time of writing, access to these different datasets at the British Library is via three completely separate interfaces, but work is underway to make them available through a unified search, thereby reducing some confusion for the user (Jackson, 2017a).

[5] The Common Crawl Foundation, based in California like the Internet Archive, makes available 'an open repository of web crawl data that can be accessed and analysed by anyone' (Common Crawl Foundation, n.d.).

[6] The Memento protocol, for example, supports the reconstruction of a web page from 'mementos' held in multiple distributed archives (Van de Sompel et al., 2009). The 'best' version of a particular page on a particular date might be drawn from three or four different national archives.

[7] This is true for most national web archiving programmes. Two exceptions are the PANDORA Archive at the National Library of Australia (Koerbin, 2017, p. 191) and the Kulturarw3 project at the Royal Library in Sweden (Webster, 2017, p. 178), both of which began in 1996. I owe these references to Niels Brügger.

[8] This patchwork quality is not unique to UK web archives. The data held by the Internet Archive itself is the result of 'collaboration between organisations, individuals, experts, users, and crawlers and external web based knowledge devices such as Alexa Internet and Wikipedia, each with their own epistemic logic and rules in all their richness and complexity' (Ben-David and Amram, 2018).

[9] The UK Web Archiving Consortium was set up in 2005, and involved six partner organisations: The National Archives, the British Library, the then Joint Information Systems Committee (now Jisc), the national libraries of

Scotland and Wales, and the Wellcome Library. The aim of UKWAC, which was active until 2008, was to establish a common selective web archive (Pennock, 2013, p. 6; Bailey and Thompson, 2006).

[10] This information about the history of the UKGWA was available on The National Archives' website until very recently, but has now been removed from the live web. It has, however, been preserved in the UKGWA itself http://webarchive.nationalarchives.gov.uk/20170608213215/https://www.nationalarchives.gov.uk/webarchive/information.htm, viewed 13 April 2018.

[11] In the UK, web archiving is enabled by Non-Print Legal Deposit legislation (at the British Library) and the Public Records Act (at The National Archives). These legislative frameworks have played a vital role in shaping the web archives we already have, and will continue to determine both what can be archived and how it can be accessed in years to come. For a detailed treatment of web archiving and the law in the UK, see Winters, 2018.

[12] The UK's six legal deposit libraries are the Bodleian Libraries of the University of Oxford, the British Library, Cambridge University Library, the National Library of Scotland, the Library of Trinity College Dublin and the National Library of Wales.

[13] Historians in the UK have only recently benefited from the reduction of a thirty-year closure period for government records to a mere two decades for non-sensitive material (The National Archives, n.d. (b)).

[14] This is only a slight exaggeration. While the figure for the average lifespan of a web page is much disputed, it is generally believed to be between around 44 and 100 days (Taylor, 2011).

[15] A search for 'terror*' produces 77,588 results, and the figure for archived websites increases to 116. This is a percentage increase from 0.07 to 0.15, but is still less than a third of the number of PhD theses listed (364). All searches were conducted on 6 June 2017.

[16] Brügger, 2011 explores some of the challenges of identifying the 'title' of a web page.

[17] Search conducted on 6 June 2017.

[18] 'Crown copyright is defined under section 163 of the Copyrights, Designs and Patents Act 1988 as works made by officers or servants of the Crown in the course of their duties' (The National Archives, n.d. (d)). The default licence for Crown copyright material is the Open Government Licence, which, with a handful of exemptions, allows both commercial and non-commercial copying, publication, distribution, transmission, adaptation and exploitation of data.

[19] To take one example, the index of URLs crawled for the single year 2012 is 68.3 GB.

[20] Academic websites, with the suffix .ac.uk, were particularly significant at the start of this period, but the registration of domain names was handled by a separate organisation, the then Joint Information Systems Committee or JISC.

[21] Nyhane and Flinn (2016) have used oral history to illuminate the development of digital humanities; while Hines (2000) makes a persuasive case that 'Ethnography can be used to develop an enriched sense of the meanings of the technology [the web and internet] and the cultures which enable it and are enabled by it' (p. 8).

[22] This quotation refers to the Internet Archive, but it is equally true of many of the archives of the UK web.