8.) Giving with one click, taking with the other: e-legal deposit, web archives and researcher access

Jane Winters

If people are familiar with web archives at all, they will most often have heard of the Internet Archive (IA), a non-profit organisation based in San Francisco, California. This is perhaps only to be expected. The IA is 'amongst the earliest systematic attempts at web archiving, operates at a global scale, and gives unrestricted access to its content via the Wayback Machine' (Webster, 2017, p. 176). It has been archiving the web since late 1996, and at the time of writing makes available more than 325 billion historical web pages for browsing and limited searching (Internet Archive, n.d.).[1] Much less well known is that archives and libraries around the world, from Iceland to Australia, are also busy archiving the web. The nature, scale and scope of this archiving activity varies enormously, but unlike the IA these institutions are concerned either solely or primarily with national web domains (usually delimited by a 'country code Top Level Domain', or ccTLD, such as .fr or .uk) rather than with the web as a whole. This chapter will outline the different legal frameworks within which this national web archiving takes places, focusing on the impact of electronic legal deposit. It will discuss the vitally important enabling role of e-legal deposit, but also describe the challenges posed by the legislation – to access, use, re-use and publication. It concludes by suggesting why researchers should concern

themselves with sometimes complex legal issues, and how they might contribute their voices to ongoing discussions about access to our digital cultural heritage.

The situation in the UK is complicated by Crown Copyright (more of which below), but in general national web archives fall into two main categories: those which are created under a legal deposit regime; and those which are collected on a permissions or fair-use basis. The International Internet Preservation Consortium's (IIPC) list of countries or regions in which some form of domain-based archiving takes place includes 17 countries in the former category and 12 in the latter (IIPC, n.d.).[2] Permissions-based and ad hoc archiving is characterised by diversity, but so too is the web archiving that takes place within the framework of legal deposit legislation.

A quick glance at the IIPC list reveals a fractured international timeline for archiving the web, and a bewildering array of access arrangements. Iceland, for example, introduced legal deposit legislation in 1997, which was extended to incorporate digital materials in 2002, and web archiving commenced at the National and University Library of Iceland two years later in 2004. Access to the archive of the .is ccTLD is open online to researchers anywhere in the world (Icelandic Web Archive, n.d.). Legal deposit in Denmark, by contrast, dates back to 1697. It was revised in 1997 to include audiovisual material (CDs and video cassettes etc.) and 'static web documents', and again in 2004 to take in radio, television and 'the dynamic internet', giving Denmark 'one of the most advanced and comprehensive legal deposit laws in the world' (Schostag & Fønss-Jørgensen, 2012, p. 110). Despite this advanced legislation,

however, off-site access to the Danish Web Archive is restricted to researchers with a specific project, who are required to submit an application to the Royal Danish Library (Royal Danish Library, n.d.(a)). In France, legal deposit dates back to the Ordonnance de Montpellier introduced by Francis I. Legislation was extended to include 'multimedia, software and databases' in 1992, and websites falling within the .fr domain in 2006. Access is even more restricted than in Denmark, however, with the archive only available on-site at the Bibliothèque nationale de France (BNF) and in 25 French regional libraries (Stirling, 2012, pp. 2-7).[3] Finally, in Sweden, legislation relating to the legal deposit of digital materials was enacted in 2012, but there is as yet no researcher access, either off- or on-site to web archives at the National Library of Sweden (Hjerpe, 2014; IIPC n.d.). In the majority of institutions where legal-deposit web archiving takes place, there is also a degree of selective, permissions-based archiving, often focused on particular events or anniversaries deemed to be significant for national cultural heritage. These open collections serve as a useful taster for the larger domain archives, but currently no more than that.

Turning to the UK, if anything the landscape becomes even more complicated.[4] There are two main organisations which have responsibility for archiving the web: The National Archives (TNA), which harvests and preserves the online presence of UK central government; and the British Library, which is charged with archiving the entire .uk ccTLD. The legislation which governs their respective web archiving activities is very different. The National Archives began to archive government websites in 2003, under the terms of the Public Records Act. The definition of what constituted 'public records' was sufficiently broad to require no changes to

the Act in order to accommodate born-digital data; and archived websites, as Crown Copyright material, may be freely reused under the terms of the Open Government Licence.[5] The result is the open UK Government Web Archive, which is accessible on-site at The National Archives in Kew, but can also be searched online, either as a standalone service or through TNA's main Discovery catalogue.

The British Library began archiving UK websites in 2004, but selectively and on a permissions basis. At the time of writing, 59 of these open special collections are available through the public UK Web Archive. In order to move beyond this time-consuming and limiting approach to archiving the web and harvest data at the domain level, it was necessary, as elsewhere, for legal deposit to be extended to include digital publications, broadly defined. The Library has been undertaking an annual crawl of the .uk domain since April 2013. This first crawl took 70 days to complete and resulted in the collection of 1.9 billion web pages and other assets, amounting to 31TB of data (Webster, 2013), and the scale of the task increases every year. In just five years, the annual domain harvest has generated an extraordinarily rich and diverse primary source for historical research, combining many different types of media, from the records of government to personal blogs to online newspapers. It has the potential to provide unique insights into life in the UK in the early 21st century, to preserve official voices alongside the words of ordinary people – bloggers, citizen journalists, teenagers – who are so often absent from the historical record. And this potential is mirrored in the domain collections being developed by other national memory institutions. As Webster (2017) notes:

> The user of web archives has reason to be thankful for the existence of a network of national libraries with a mission to preserve published heritage at a large scale. Without this network, with its long-established channels of communication and co-operation, users would be even more reliant on the Internet Archive than they already are (p. 179).

The work of the Internet Archive has been both ground-breaking and foundational, but the IA does not have a statutory responsibility to preserve the outputs of a nation's culture; nor is it able to draw on, in some cases, centuries of experience in adapting to changes in the technologies of publishing and reading. In so far as it is possible to guarantee the long-term preservation and sustainability of digital cultural heritage, national libraries and archives are exceptionally well placed at least to try.

The extension of legal deposit to include the archiving of various national web domains has, however, proven to be something of 'a double-edged sword' for researchers, and indeed for libraries and archives (Milligan, 2015). There can be no doubt that it underpins the collection, preservation, (partial) curation and republication of the archived web in a stable form.[6] But legal deposit also builds in barriers to the use and reuse of web archives by researchers, and even more by the wider public. Some of these barriers are there by design; others fall under the heading of unintended consequences.

The first difficulty posed by the archiving of the web on the basis of legal deposit arises from the national nature of the relevant legislative traditions and frameworks. The live web is (relatively) open, networked and international, but national web archives are primarily defined

by a ccTLD.[7] They are explicitly, and necessarily, national in focus. This is not to say that some material from one national domain is not 'accidentally' harvested during a web crawl for another, but that is not the purpose of the exercise. Artificial geographical boundaries are imposed on web archives which are simply not present on the live web. This is problematic for most kinds of research. Even if the object of study is the web presence of a single UK institution, it is likely that the researcher will come across links to websites and web pages which lie within a different national domain and therefore have not been archived by the British Library. They will hit a virtual brick wall. In this example, the fact that a handful of web pages are not accessible is more of an annoyance than an insurmountable barrier, but what of the researcher who is interested in tracing an international phenomenon? The UK Web Archive holds a large amount of material for a historian interested in the growth of Euroscepticism in Britain (Deswarte, 2015), but reactions to that Euroscepticism in other European countries will not be present; nor will it be possible to gain much insight into parallel movements and trends in other countries.[8]

The scattering of primary sources across different archives and countries is far from a new challenge for researchers, but it has not generally been deliberate.[9] It may even be useful for the historian. The final archival destination of a manuscript, document or book can shed fascinating light on the movement of people, knowledge, money and influence over time. Web archives, however, are artificially segmented by the requirements of legal deposit, and thus become further removed from the live web of which they will be our only record. This is in striking contrast to the ways in which libraries and archives have used digital technologies to

remove geographical barriers to access for their analogue holdings, by digitising some of their collections and publishing them on the web. There are numerous examples of projects and initiatives which have gone a step further and virtually reunited physically dispersed corpora.[10] Digitisation and the web have transformed access to much archival material, but counterintuitively, access to born-digital and readily accessible primary sources is closed off by archiving processes.

The obvious solution to this problem would be to aggregate the national domain collections in some way, perhaps even to develop, as a first step, a service which would allow cross-searching of European web archives.[11] Options for regional rather than national web archiving initiatives have also been explored, for example among the Nordic countries (Hallgrímsson and Bang, 2003). But as noted above, many of these national web archives are not accessible online even to the citizens of the countries whose history they record. And once again, this is a function of legal deposit. Some national domain archives which rely on legal deposit are, in fact, open by default, for example the Croatian Web Archive at the National and University Library in Zagreb; but they are the exception rather than the rule. In the UK, access to the domain crawl archive is restricted to reading rooms in the UK's six legal deposit libraries.[12] An open, selective archive is publicly accessible off-site, but it contains only a fraction of the content collected under legal deposit.[13] Even on-site, the user experience is far from seamless, as recounted by Milligan (2015). A new beta search interface has been launched since Milligan reported, but the key points remain the same: it is not possible (or rather, not legal) to take photographs of information on the computer screen (whether or not that might be

useful in practice); intrusive copyright warning messages interfere with browsing; all dynamic content is disabled; and it is not possible to cut and paste text or images. As Milligan notes, 'My research process would be slower using this archive than any other traditional archive that allowed digital photography'. In essence, an archive of born-digital data has been rendered less easy to use than a collection of medieval manuscripts or early modern printed books held in the same Library. Functionality that is taken for granted on the live web has been stripped away in order to comply with legal deposit restrictions.

It is the disjuncture between browsing and searching the live web and the web archive that is so disconcerting to a researcher. Some of this results from the archiving process itself: 'the fundamental characteristic of the archived website is that it is a reconstructed and unique version of what was once online, rather than a copy of it'; it is 're-born digital' and does not have the same functionality as a live website (Brügger, 2012a, p. 758). Some of it, however, is imposed by the legislation. One of the most obviously peculiar effects of the legal framework which allows the archiving of the web in the UK is the de facto treatment of a web page as akin to a printed book. From that starting point, it follows that no two people in the same legal deposit library can simultaneously view the same instance of a captured web page.[14] In order to make sure that this does not happen, the web archiving team at the British Library have had to devise and implement a technical solution that automatically refers the user to a different instance of the desired web page, archived as close in time as possible to the originally selected date. If this remains the case, it will be exceptionally difficult to use the UK Web Archive in a teaching context. Students would be unable to explore and critique the same archived website

as a group, for example, unless they crowded round a single monitor – behaviour which is not exactly encouraged in a library reading room. Similar problems will face anyone wishing to undertake collaborative research.

The restrictions on use, however, are even more serious than this. As Milligan (2015) discovered, the locked-down nature of the legal deposit collection means that researchers are prevented from accessing the source code for a web page. They cannot see or analyse the underlying html. Critical code studies is an emerging field of research, and is already playing an important role in web historiography:

> Focusing on the archived source code of websites does not only enable analysing web technologies used to construct them, which can tell us something about the web's underlying infrastructure, providing insights into how the web is built and how websites are connected, but can also serve as a means to investigate the web's economic underpinnings, to understand the business models of websites and third parties and trace the economically valuable data flowing between them (Helmond, p. 139).

This is an enormously promising mode of investigation, but one to which the UK Web Archive is not, and is unlikely to be, susceptible – at least until researcher access is transformed. This is a clear example of unintended consequences. At no point in the drafting of the legal deposit legislation was there a formal decision to allow researchers only to read and analyse the visible text of a web page, but that is the practical effect of the current wording of the law.[15] Policy-makers and legislators cannot be expected to anticipate the future directions of research in the humanities and social sciences, but care should be taken not to close off opportunities; to limit

the kinds of research questions that can be asked; even to distort whole disciplines and fields of enquiry. This will only happen if researchers collectively feed in to conversations about the development and evolution of legal deposit in relation to born digital archives.

The freedom of researchers to develop their own methodologies, to choose what and how they research, is crucial for the national research base and the strength of the university sector as a whole. The risk of undermining this in relation to web historiography is clear. Nowhere is this more apparent than when it comes to analysing web archives at scale. There are currently two clear trends in research drawing on web archives: qualitative, case-study approaches to the history and development of particular websites or web spheres[16] (see, for example, Raffal, 2018; Chakraborty and Nanni, 2017; and Dougherty, 2017); and quantitative analyses either of text corpora or of derived data, for example link analysis (see, for example, Ben-David and Amram, 2018; Hale et al., 2014; and Milligan, 2017). Researchers wishing to work with the UK Web Archive, or legal deposit collections in other countries, are forced into adopting a qualitative approach. They can, laboriously, work through the archived pages of an organisation or search for pages related to an event such as the London 2012 Olympics. The British Library's curated special collections similarly encourage this kind of research. What researchers cannot do is create their own web corpora, download all of the images associated with a particular event, analyse links between websites in a .uk subdomain, for example .ac.uk or police.uk. The archived web is a source of enormous potential value for linguistic study, for example in relation to the emergence and take-up of neologisms (see, for example, Winters, 2018), but at present the archive can neither be brought to the appropriate analytical tools, nor

the appropriate analytical tools brought to the archive. This creates difficulties for researchers, but it also poses present and future challenges for archiving institutions. If their web archives are to be used at all, they will have no option but to develop at least some of the necessary tools themselves. In some cases they will be forced to reinvent the wheel because of legal restrictions that may well not persist. In straitened times, this is an enormous waste of resources – human, financial and technical.

This does not mean that there will be no quantitative research using web archives – there is a burgeoning literature to the contrary[17] – but it may well mean that those interested in quantitative analysis will simply bypass legal deposit collections. Those responsible for web archiving in national memory institutions are all too aware of this risk. Paul Koerbin, the Assistant Director for Web Archiving and Government Publications at the National Library of Australia, argues that national libraries are in a good position to archive the web because to do so effectively 'requires a strategic purpose with a long term objective' – this is something that national libraries know how to do. But

> There is much that is required in terms of technologies, systems, policies, procedures and resources to make archiving more than merely harvesting and storing. Not the least is the *sine qua non* purpose of archiving and preservation: sustainable long-term access. Without access preservation is little more than a costly and meaningless storage burden (Koerbin, 2017, p. 195).

Legal deposit legislation does not only hamper those whose research calls for quantitative, big data approaches; it also prevents a range of non-academic users from engaging with these

collections. The British Library is, of course, open to anyone over the age of 18 'who has a need to see specific items in our collection' (British Library, n.d.(a)), but travelling to the Library or to one of the other legal deposit libraries in the UK in order to register for a reader pass is a significant hurdle, and one that the majority of people will never overcome (or indeed even consider overcoming). If they are interested in reading a particular book or journal article though, they might just consider ordering a copy using British Library On Demand (formerly the BL Document Supply Service), which at the time of writing provides access to approximately 87.5 million items, with publisher agreement (British Library, n.d.(b)). There is as yet no option to request a print-out of an archived web page. There is a hierarchy of access in place, and web archives are firmly at the bottom of the pyramid.

The effects of these restrictive access arrangements are becoming increasingly obvious, particularly in the media. As I have noted elsewhere, 2016 was something of a breakthrough year for the visibility of web archives in the mainstream media (Winters, 2017b, pp. 175-176). Journalists began to turn to the archived web as a source of evidence when information disappeared from the live web or was thought to have been altered in some way.[18] In the overwhelming majority of cases, it is content archived by the IA to which these newspaper and magazine articles refer; information which is readily available to journalists at their desks and to which readers will also have seamless access. You would not expect a reporter from a US newspaper to consult, or even know about, the archived collections of the British Library or the Bibliothèque Nationale de France, but these vast national legal deposit collections are also being overlooked by journalists in the UK, France and elsewhere. This has potentially very

damaging implications for memory institutions with responsibility for archiving the web, who are at a significant disadvantage when it comes to promoting and encouraging awareness about and use of their digital holdings. This was demonstrated very clearly when the BBC announced in May 2016 that it would be taking down its much-loved BBC Food website. There was a great deal of news coverage in the UK media, a lot of it pointing to the Internet Archive as the place where people would still be able to find their favourite recipes. The team at the UK Web Archive responded with a blog post announcing that they had 'instigated a … crawl of the BBC website with the specific aim of ensuring that we save the recipes from the food pages', but with no content to link to the message was easy to overlook (Winters, 2017b, pp. 175-176; Webber, 2016). The relative inaccessibility of the legal deposit web archive stands in stark contrast to the readily available live web, and to archived content on that live web. The (more or less) public has become (more or less) private.

The severity of the restrictions imposed by legal deposit legislation on the use and reuse of web archives is, moreover, increasingly at odds with current trends in academic research and publishing worldwide. In the UK, all journal articles and conference proceedings published after 1 April 2016 have been required to comply with an open-access mandate in order to be considered in national research evaluation processes (specifically the Research Excellence Framework or REF) (HEFCE, 2016).[19] This mandate will be extended, in one form or another, to monographs after the 2021 REF (Hill, 2018). Researchers are also encouraged, wherever possible, not only to publish articles and similar works on an open-access basis, but to share the data that underpins their research findings in a similarly open fashion. The European

Commission, for example, launched an Open Research Data pilot which was extended to cover 'all thematic areas of Horizon 2020 from the start of the 2017 work programme', instituting a regime in which 'open access becomes the default setting for research data' (European Commission, 2017, pp. 8-9). It is permissible for researchers to opt out of publishing their data openly, either wholly or partially, and this will necessarily be the default position for researchers who work with web archives. In many instances, the open publication of research data places new or previously hidden information into the public sphere; users of the archived web, by contrast, are forced to restrict access to data which was previously open. Uncertainty about what is and is not permissible with regard to the republication of data from legal deposit web archives places a significant burden on individuals to engage with a complex legal framework, and potentially encourages an unnecessarily risk-averse approach. Best practice guidance and publication norms will undoubtedly emerge as users begin to test what is acceptable, but for the moment it remains easier to include a textual description of a web archive screenshot in a research article than to publish even a thumbnail image.[20]

Faced with this litany of problems and challenges, it would be all too easy to throw up one's hands in exasperation and decide to study only publicly accessible web archives, or even choose to work with alternative primary source materials. This is not, however, either a reasonable or a desirable position for researchers to adopt. Web archives developed under a legal deposit regime are inherently more stable, comprehensive and sustainable than those generated by individuals, collectives, businesses or philanthropic organisations, whatever their motivation, commitment or degree of financial investment. We may rail against the failure of

legislation to keep up with developments in both technology and research, to adapt sufficiently swiftly to new ways of living and working, but it is that same legislation which both makes possible the harvest of national web domains and assigns responsibility for the task. The long-term preservation of our digital cultural heritage is not left to chance, made vulnerable to privatisation or commercialisation, to creeping changes in 'Terms and conditions'. Like other national libraries and archives, 'The British Library's mission is to make [the UK's] intellectual heritage accessible to everyone for research, inspiration and enjoyment'. Its duties and functions are defined by law, which may be slow to change but is also slow to overturn (British Library n.d.(c)). Legal deposit brings undoubted problems, but it also brings unique benefits.

Web archives have all the characteristics of a black box; indeed 'understanding a Web archive implies opening several black boxes, the first being that of its collection' (Schafer et al., 2016, p. 3). But the archives collected by national memory institutions, in line with electronic legal deposit, are far more likely than other kinds of web archive to have been documented to the level of detail required by researchers if their analyses are to be sufficiently robust. Historians in particular have always needed to consider the archival context for their primary sources – why did a manuscript end up in one collection rather than another? – but this is particularly true for web archives. History and provenance are vital, but also exceptionally hard to establish. Failure to take into account the different sources of data in a superficially coherent web archive, or to acknowledge the impact of changes in the technologies of web crawling over time, can distort or even fatally undermine research findings (see, for example, Winters, 2017a). An example of good practice in this area is the publicly available

documentation for the UK Government Web Archive (UKGWA), which clearly sets out scope and selection policy (The National Archives of the UK, 2014).[21] The UKGWA is not a legal deposit collection, but it similarly has a basis in statute (The National Archives of the UK, n.d.(c)). The team behind the UK Web Archive at the British Library have documented their processes on a public blog, which has itself been archived (Webber, 2016), and more detailed information still will become available to researchers as it falls within the remit of the BL Corporate Archive Policy (British Library, 2012). Initiatives like the IA and the Common Crawl have established themselves as important parts of digital research infrastructure in the early 21st century, but they do not face the same imperative to explain or account for the ways in which they have worked in the past, or how they will work in the future.[22]

To date, web archiving has been characterised by an extraordinary degree of collaboration between archiving institutions and researchers to explore the full potential of legal deposit archives, to open them up to a wide contemporary audience, and to try to account for the kinds of access that future historians will need and expect. These collaborations demonstrate what is already possible within the constraints of present legislation and point the way to future innovation. In France, 'Web90: patrimoine, mémoires et histoire du Web dans les années 1990' has combined the study of web archives at the BNF with an exploration of the material culture of the web and internet; in the UK, the BUDDAH project has demonstrated the diversity of the research that can be conducted using web archives, from the investigation of disability action groups online (Millward, 2015) to an ethnosemiotic study of the French in London (Huc-Hepher, 2016). A long-term collaboration between researchers at Aarhus

University and the Royal Danish Library has produced ground-breaking research illuminating the shape of a national web domain and provides a model for research in other countries (Laursen & Møldrup-Dalum, 2017). Others, notably Ian Milligan, have shown what can be achieved when historians and computer scientists work together to explore unrestricted web archives (see, for example, Milligan, 2016a and 2017). It is to be hoped that the tools and methods Milligan and his colleagues at Waterloo University are developing will be applicable to legal deposit materials at some point in the future. And we know that there is a good chance that those collections will remain available to us in decades to come because of, not despite, electronic legal deposit. We simply need the legislation to develop, to empower national memory institutions to share the digital materials that they are collecting with the widest possible audience, and to allow us easily to access our own digital stories and histories. Electronic legal deposit has laid the groundwork, but more still needs to be done if the full potential of this new type of primary source is to be realised.

As an individual researcher, it is easy to feel rather helpless in the face of legal deposit restrictions. After all, more than 15 years elapsed between the first discussions about extending legal deposit to include non-print material and the implementation of the new legislation (Webster, 2017, p. 180). What can one person do to influence change? One person probably cannot achieve very much, but researchers who work with the archived web, or would like to do so in the future, can find ways to influence national discussions. We can produce case studies which will help libraries and archives to demonstrate the value of the data that they hold and argue for improved access conditions – and we can publish them openly so that they are

accessible to journalists and policy-makers. We can talk (and listen) to librarians and archivists about those areas where we are most likely to be able to make progress, so that energy is not wasted on the (for now) unachievable. Privileged research access, for example, might be a first step towards wider openness in the long term. Over time, we can produce an evidence base and build collaborations that will help to bring about legislative evolution. We can afford to wait a little – the data, thanks to legal deposit and the work of national memory institutions, will not be going anywhere. But we cannot afford to wait too long, or research and teaching will be held back, and opportunities missed.

References

Ben-David, A. and Amram, A. 2018. The Internet Archive and the socio-technical construction of historical facts, *Internet Histories: Digital Technology, Culture and Society*, viewed 6 April 2018. DOI: 10.1080/24701475.2018.1455412.

Bibliothèque nationale de France (BNF) n.d. *Digital legal deposit: four questions about web archiving at the BNF*, viewed 4 April 2018, http://www.bnf.fr/en/professionals/digital_legal_deposit/a.digital_legal_deposit_web_archiving.html.

Big UK Domain Data for the Arts and Humanities, viewed 5 April 2018, https://buddah.projects.history.ac.uk/.

British Library 2017. UK Web Archive (beta version), viewed 5 April 2018, https://beta.webarchive.org.uk/.

British Library 2012. *British Library Corporate Archive Policy*, viewed 14 April 2018, https://www.bl.uk/aboutus/foi/pubsch/pubscheme5/20130124%20BLCA%20Policy.pdf.

British Library 2004. UK Web Archive, viewed 5 April 2018, https://www.webarchive.org.uk/ukwa/.

British Library n.d.(a) *Get a reader pass*, viewed 13 April 2018, https://www.bl.uk/help/how-to-get-a-reader-pass.

British Library n.d.(b) *On demand*, viewed 13 April 2018, https://www.bl.uk/on-demand.

British Library n.d.(c) *British Library – Statement of public task*, viewed 13 April 2018, https://www.bl.uk/aboutus/stratpolprog/pubsect-info-regulations/faq/bl_psi_public_task_statement.pdf.

Bodleian Libraries, n.d. *Identifying UK websites and electronic publications*, viewed 5 April 2018, https://www.bodleian.ox.ac.uk/our-work/legal-deposit/information-for-publishers/identifying-uk-websites-and-electronic-publications.

Brügger, N. 2012a. Web historiography and internet studies: challenges and perspectives, *New Media and Society*, 15:5, pp. 752-764. DOI: 10.1177/1461444812462852.

Brügger, N. 2012b. Web history and the web as a historical source, *Zeithistorische Forschungen/Studies in Contemporary History*, 9:2, pp. 1-11, viewed 6 April 2018, http://www.zeithistorische-forschungen.de/2-2012/id=4426.

Chakraborty, A. and Nanni, F. 2017. The changing digital faces of science museums: a diachronic analysis of museum websites. In: Brügger, N. ed., *Web 25: Histories from the First 25 Years of the World Wide Web*. New York: Peter Lang Publishing, pp. 157-172.

Common Crawl, viewed 13 April 2018, http://commoncrawl.org/.

Cowls, J. 2017. Cultures of the UK web. In: Brügger, N. and Schroeder, R. eds. *The Web as History: Using Web Archives to Understand the Past and the Present*. London: UCL Press, pp. 220-237, viewed 5 April 2018, http://discovery.ucl.ac.uk/1542998/1/The-Web-as-History.pdf.

Cowls, J. and Bright, J. 2017. International hyperlinks in online news media. In: Brügger, N. and Schroeder, R. eds. *The Web as History: Using Web Archives to Understand the Past and the Present*. London: UCL Press, pp. 101-116, viewed 30 August 2018, http://discovery.ucl.ac.uk/1542998/1/The-Web-as-History.pdf

Cummings, M. 2014. Beinecke Library acquires 'treasure trove' of medieval manuscripts from a famed 'book breaker'. *YaleNews*, viewed 30 August 2018, http://discovery.nationalarchives.gov.uk/details/r/C433.

Deswarte, R. 2015. Revealing British Euroscepticism in the UK web domain and archive case study. *Web Archives as Big Data*, viewed 5 April 2018, http://sas-space.sas.ac.uk/6103/.

Early English Laws, viewed 5 April 2018, http://www.earlyenglishlaws.ac.uk/.

European Commission, 2017. *H2020 Programme: Guidelines to the rules on open access to scientific publications and open access to research data in Horizon 2020*, viewed 13 April 2018, http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.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: Brügger, N. and Schroeder, R. eds. *The Web as History: Using Web Archives to Understand the Past and the Present*. London: UCL Press, pp. 204-219, viewed 6 April 2018, http://discovery.ucl.ac.uk/1542998/1/The-Web-as-History.pdf.

Goel, V. 2016. Beta Wayback Machine – now with site search! *Internet Archive Blogs*, viewed 4 April 2018, https://blog.archive.org/2016/10/24/beta-wayback-machine-now-with-site-search/.

Gomes, D., Nogueira, A., Miranda, J. and Costa, M. 2008. Introducing the Portuguese web archiving initiative. 8th International Web Archiving Workshop, Aarhus, Denmark, viewed 5 April 2018, http://sobre.arquivo.pt/wp-content/uploads/introducing-the-portuguese-web-archive-initiative.pdf.

Hale, S., Yasseri, T., Cowls, J., Meyer, E. T., Schroeder, R. and Margetts, H. 2014. Mapping the UK webspace: fifteen years of British universities on the web. *Proceedings of the 2014 ACM Conference on Web Science*, pp. 62-70, viewed 6 April 2018, https://dl.acm.org/citation.cfm?id=2615691.

Hallgrímsson, Þ. and Bang, S. 2003. *Nordic Web Archive*, viewed 4 April 2018, http://nwatoolset.sourceforge.net/docs/nwa@ecdl2003.pdf.

Helmond, A. 2017. Historical website ecology: analyzing past states of the web using archived source code. In: Brügger, N. ed., *Web 25: Histories from the First 25 Years of the World Wide Web*. New York: Peter Lang Publishing, pp. 139-155.

Higher Education Funding Council for England (HEFCE) 2016. *Policy for open access in the next Research Excellence Framework: updated November 2016*, viewed 13 April 2018, http://www.hefce.ac.uk/pubs/year/2016/201635/.

Hill, S. 2018. Open access monographs in the REF, *HEFCE Research Policy blog*, viewed 13 April 2018, http://blog.hefce.ac.uk/2018/02/23/open-access-monographs-ref-2027/.

Hjerpe, A. 2014. Detailed digital deposits. *Scandinavian Library Quarterly*, 47:2, viewed 4 April 2018, http://slq.nu/?article=volume-47-no-2-2014-7.

Huc-Hepher, S. 2016. Searching for home in the historic web: an ethnosemiotic study of London-French habitus as displayed in blogs, *Web archives as big data*, viewed 13 April 2018, http://sas-space.sas.ac.uk/6252/.

Icelandic Web Archive n.d. *Um íslenska vefsafnið*, viewed 4 April 2018, http://vefsafn.is/index.php?page=um-vefsafnid.

International Internet Preservation Consortium (IIPC) n.d. *Legal deposit*, viewed 4 April 2018, http://netpreserve.org/web-archiving/legal-deposit/.

Internet Archive n.d. *Wayback Machine*, viewed 4 April 2017, http://web.archive.org/.

Kamen, M. 2016. Vote Leave wipes homepage after Brexit result, *Wired*, viewed 13 April 2018, http://www.wired.co.uk/article/vote-leave-wipes-website-after-brexit.

Koerbin, P. 2017. Revisiting the World Wide Web as artefact: case 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.

Laursen, D. and Møldrup-Dalum, P. 2017. Looking back, looking forward: 10 years of development to collect, preserve, and access the Danish web. In: Brügger, N. ed., *Web 25:*

Histories from the First 25 Years of the World Wide Web. New York: Peter Lang Publishing, pp. 207–227.

Leetaru, K. 2015a. How much of the internet does the Wayback Machine really archive?, *Forbes*, viewed 13 April 2018, https://www.forbes.com/sites/kalevleetaru/2015/11/16/how-much-of-the-internet-does-the-wayback-machine-really-archive/.

Leetaru, K. 2015b. Why it's so important to understand what's in our web archives, *Forbes*, viewed 13 April 2018, https://www.forbes.com/sites/kalevleetaru/2015/11/25/why-its-so-important-to-understand-whats-in-our-web-archives/.

Library of Congress (LOC) 2009. WARC, Web ARChive file format, viewed 5 April 2018, https://www.loc.gov/preservation/digital/formats/fdd/fdd000236.shtml.

Library of Congress (LOC) 2008. *ARC_IA, Internet Archive ARC file format*, viewed 5 April 2018, https://www.loc.gov/preservation/digital/formats/fdd/fdd000235.shtml.

Milligan, I. 2017. Welcome to the web: the online community of GeoCities during the early years of the World Wide Web. In: Brügger, N. and Schroeder, R. eds. *The Web as History: Using Web Archives to Understand the Past and the Present*. London: UCL Press, pp. 137-158, viewed 6 April 2018, http://discovery.ucl.ac.uk/1542998/1/The-Web-as-History.pdf.

Milligan, I. 2016a. Lost in the infinite archive: the promise and pitfalls of web archives, *International Journal of Humanities and Arts Computing*, 10: 1-2 (2016), pp. 78-94.

Milligan, I., Ruest, N. and St. Onge, A. 2016b. The great WARC adventure: Using SIPS, AIPS, and DIPS to document SLAPPs, *Digital Studies/ Le champ numérique*. DOI: 10.16995/dscn.18.

Milligan, I. 2015. Web archive legal deposit: a double-edged sword, *Ian Milligan: Digital History, Web Archives, and Contemporary History*, viewed 6 April 2018, https://ianmilligan.ca/2015/07/14/web-archive-legal-deposit-a-double-edged-sword/.

Millward, G. 2015. Digital barriers and the accessible web: disabled people, information and the internet, *Web archives as big data*, viewed 13 April 2018, http://sas-space.sas.ac.uk/6104/.

Moore, M. 2017. Museum fees are killing art history, say academics, *The Times*, viewed 13 April 2018, https://www.thetimes.co.uk/edition/news/museum-fees-are-killing-art-history-say-academics-qhfwmdws6.

The National Archives of the UK 2014. *Operational Selection Policy OSP27: UK Central Government Web Estate*, viewed 13 April 2018, http://www.nationalarchives.gov.uk/documents/information-management/osp27.pdf.

The National Archives of the UK n.d.(a). *Crown copyright*, viewed 4 April 2018, http://www.nationalarchives.gov.uk/information-management/re-using-public-sector-information/uk-government-licensing-framework/crown-copyright/.

The National Archives of the UK n.d.(b), Correspondence with the colonies, entry books and registers of correspondence, viewed 30 August 2018, http://discovery.nationalarchives.gov.uk/details/r/C433.

The National Archives of the UK n.d.(c). UK Government Web Archive, viewed 5 April 2018, http://www.nationalarchives.gov.uk/webarchive/.

Punzalan, R. L. 2014. Digital diasporas: a framework for understanding the complexities and challenges of dispersed photographic collections, *The American Archivist*, 77:2, pp. 326-349. DOI: 10.17723/aarc.77.2.729766v886w16007.

Raffal, H. 2018. Tracing the online development of the Ministry of Defence and Armed Forces through the UK web archive, *Internet Histories: Digital Technology, Culture and Society*, viewed 6 April 2018. DOI: 10.1080/24701475.2018.1456739.

RESAW: A Research Infrastructure for the Study of Archived Web Materials, viewed 5 April 2018, http://resaw.eu/.

Revesz, R. 2016. Melania Trump's website vanishes from internet as rumours swirl over her university degree, *The Independent*, viewed 13 April 2018, https://www.independent.co.uk/news/world/americas/melania-trump-vanishes-biography-website-online-disappears-rumours-plagiarised-speech-education-a7160856.html.

Royal Danish Library n.d.(a). *Netarchive*, viewed 4 April 2018, https://en.statsbiblioteket.dk/national-library-division/netarchive.

Royal Danish Library n.d.(b). *Act on Legal Deposit of Published Material (translation)*, viewed 5 April 2018, http://www.kb.dk/en/kb/service/pligtaflevering-ISSN/lov.html.

Schafer, V., Musiani, F. and Borelli, M. 2016. Negotiating the web of the past, *French Journal for Media Research, La toile négociée/Negotiating the web*, viewed 13 April 2018, http://frenchjournalformediaresearch.com/lodel/index.php?id=963.

Schostag, S. and Fønss-Jørgensen, E. 2012. Webarchiving: legal deposit of internet in Denmark. A curatorial perspective, *Microform & Digitization Review*, 41, pp. 110-120.

Stirling, P., Illien, G., Sanz, P. and Sepetjan, S. 2011. The state of e-legal deposit in France: looking back at five years of putting new legislation into practice and envisioning the future, *World Library and Information Congress: 77th IFLA General Conference and Assembly*, pp. 1-27, viewed 4 April 2018, https://www.ifla.org/past-wlic/2011/193-stirling-en.pdf.

Web90: patrimoine, mémoires et histoire du Web dans les années 1990, viewed 13 April 2018, https://web90.hypotheses.org/.

Webber, J. 2016. Saving BBC Recipes website, *UK Web Archive Blog*, viewed 30 August 2018, http://blogs.bl.uk/webarchive/2016/05/saving-bbc-recipes-website.html.

Webber, J. ed. 2016. *UK Web Archive blog*, viewed 13 April 2018, http://blogs.bl.uk/webarchive/index.html.

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. 175-190.

Webster, P. 2013. Crawling the UK web domain, *UK Web Archive blog*, viewed 4 April 2018, http://blogs.bl.uk/webarchive/2013/09/domaincrawl.html.

Winters, J. 2018. Negotiating the archives of UK web space. In: Brügger, N. and Laursen, D. eds. *The Historical Web and Digital Humanities: the Case of National Web Domains*. Abingdon: Routledge.

Winters, J. 2017a. Tackling complexity in humanities big data: from parliamentary proceedings to the archived web. In: Hiltunen, T., McVeigh, J. and Säily, T. eds. *Big and Rich Data in English Corpus Linguistics: Methods and Explorations*. Helsink: Varieng, viewed 13 April 2018, http://www.helsinki.fi/varieng/series/volumes/19/winters/.

Winters, J. 2017b. Breaking in to the mainstream: demonstrating the value of internet (and web) histories, *Internet Histories: Digital Technology, Culture and Society*, 1:1-2, pp. 173-179. DOI: 10.1080/24701475.2017.1305713.

.

[1] For most of its history, it was only possible to search the Wayback Machine by URL. This meant that researchers were only able to find web pages about which they already knew, or with a URL which could be inferred. In 2016, however, a beta search service was launched which offered full-text searching of website home pages (Goel, 2016). This functionality was subsequently rolled out to the public Wayback Machine.

[2] It should be noted that this list does not include the Portuguese Web Archive (Arquivo.pt), which unusually does not fall into either category. Legal deposit legislation in Portugal 'created a framework to support selective archiving of online publications', but Arquivo.pt does not archive the Portuguese web on this basis (Gomes et al., 2008).

[3] Responsibility for archiving the French web is divided between the BNF and the Institut national de l'audiovisuel, with the latter dealing with online audiovisual materials (BNF, n.d.)

---

[4] For a more detailed description of the history and development of web archiving in the UK, see Winters, 2018.

[5] '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.(a)). 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.

[6] The Internet Archive developed the ARC file format for its Heritrix web crawler (LOC, 2008), but the majority of national libraries and archives now use the improved WARC file format (LOC, 2009).

[7] In most countries, the legislation does allow national memory institutions limited freedom to look beyond the appropriate ccTLD. In the UK for example, a web page falls within scope if it is 'made available to the public by a person and any of that person's activities relating to the creation or the publication of the work take place within the United Kingdom' (Bodleian Libraries, n.d.). The Danish Act on Legal Deposit of Published Material mandates that 'Material published in electronic communication networks are [*sic*] considered to be Danish when … it is published from other Internet domains etc. and is directed at a public in Denmark' (Royal Danish Library, n.d.(a)).

[8] This case study is one of 10 produced for the Big UK Domain Data for the Arts and Humanities (BUDDAH) project, funded by the UK Arts and Humanities Research Council (AHRC) as part of its Digital Transformations in the Arts and Humanities theme (grant reference AH/L009854/1). A synthesis of all 10 case studies is available in Cowls, 2017.

[9] Archives are sometimes subject to structural separation, as is immediately apparent to anyone researching the history of the British empire at The National Archives of the UK. TNA's Discovery catalogue, for example, notes in relation to 'Correspondence with the colonies, entry books and registers of correspondence' that 'Domestic records of colonial governments do not normally form part of the Colonial Office records, being kept by those governments in their own archives' (The National Archives of the UK, n.d.(b)). The profit motive has also come into play on occasion, as archives and individual manuscripts have deliberately been broken up for sale (see, for example, Cummings, 2015; I am grateful to Melissa Terras for suggesting this comparison). Some of these issues fall within the useful concept of 'archival diaspora' (see Punzalan, 2014).

[10] See, for example, Early English Laws (grant reference AH/F019394/1), which digitised and made available online 81 manuscripts from 25 libraries and archives in three different countries.

[11] Exploring the possibilities for just this kind of research infrastructure inspired the European network RESAW: A Research Infrastructure for the Study of Archived Web Materials.

[12] The British Library undertakes the crawls of the .uk domain on behalf of the wider consortium, which includes the National Library of Scotland, the National Library of Wales, the Bodleian Libraries, Oxford, the University Library, Cambridge and the Library of Trinity College, Dublin.

[13] This selective archive remains accessible as a standalone service (still referred to as the UK Web Archive at the time of writing), but in late 2017 the British Library launched an expanded beta version of the UK Web Archive, which for the first time allows users to search across both open and restricted content. This successfully removes one barrier to access – the requirement for researchers to use different interfaces to explore web archives depending on the terms by which they were collected – but it makes more prominent the problem of on-site access. The vast majority of search results are flagged with a notice in red that the archived website is 'Viewable only on Library premises'.

[14] 'A deposit library 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, no. 777, part 4, regulation 23); and 'a deposit library must ensure that only one reader uses an accessible copy of the same relevant material made under this regulation at any one time (Non-Print Works) Regulations 2013, no. 777, part 4, regulation 26). I am grateful to Andy Jackson at the British Library for clarifying that the simultaneous access restriction applies on a library-by-library basis.

[15] See Non-Print Works) Regulations 2013, no. 777, part 4, 'Permitted activities'.

[16] Brügger (2012b) defines a web sphere as 'web activity related to an event, a theme or the like (for instance political elections, catastrophes, etc.)' (p. 2).

---

[17] The journal *Internet Histories: Digital Technology, Culture and Society* is a good place to start, and particularly the bumper launch issue (1-2, 2017). See also Meyer et al., 2017; Hale et al., 2014; Cowls and Bright, 2017; Helmond, 2017; and Milligan, 2016b.

[18] Notable examples from that year include the removal of information from the website of the EU Referendum Leave campaign in the UK shortly after the vote on 23 June; and coverage in the US of the taking down of Melania Trump's website amid disputed claims about her education (Kamen, 2016; Revesz, 2016; and see Winters, 2017b, p. 176).

[19] In a striking example of the value of web archives generally, but particularly of open and accessible web archives, at the time of writing the web page which links to the relevant HEFCE policy document contains a note that, as a result of changes to the structure of UK higher education, 'The HEFCE domain – www.hefce.ac.uk – will continue to function until September 2018. At this point we will close the site entirely and all its information will only be available from the National Web Archive [*sic*]' (HEFCE, 2016). The misnamed 'National Web Archive' is, in fact, the open UKGWA.

[20] Researchers working with web archives are, of course, not the only ones to face difficulties when it comes to the publication of images in particular. The challenges for art historians, for example, are well rehearsed (see, e.g., Moore, 2017). Web archives, however, stand apart from other materials in legal deposit collections in being uniquely restricted from reproduction. It might be necessary to pay a fee to reproduce a page from an early modern incunable, but at least it may be reproduced, and the permission processes are both known and tested.

[21] Rather less good practice is evident in the disappearance from the live TNA website of a useful narrative history of the construction of the UKGWA. It does, however, remain publicly available in archived form at http://webarchive.nationalarchives.gov.uk/20170608213215/https://www.nationalarchives.gov.uk/webarchive/information.htm, viewed 13 April 2018.

[22] The problems arising from a relative lack of transparency are raised by Leetaru, 2015a & 2015b, albeit unduly critically.