Users, technologies, organisations: Towards a cultural history of world web archiving

Peter Webster

If 2015 marked the elapse of 25 years since the birth of the web, 2016 marked the 20th anniversary of web archiving: of systematic attempts to preserve web content and make it accessible to scholars and the public. As such, the time is ripe to make an initial assessment of the history of the movement, and the patterns into which it has already fallen. Although there have been short sketches of this history (Brown, 2006, pp. 8–23; Brügger, 2011, pp. 29–32), this chapter represents the first attempt to document the subject at length. In the space available, it could not be hoped to provide an exhaustive account of the activities of diverse organisations and individuals in many countries. The chapter attempts to draw the main contours of a landscape, the details of which may be filled by other more local and thematic studies. The timing is particularly significant since several of the pioneers of web archiving have reached or are approaching retirement, and so this study uses interview evidence as a supplement to written documentation.

Some notes on scope are necessary. The story of the technical evolution of web archiving is a complex one, reflecting the sheer speed of the evolution of the web itself and the technological

'arms race' in which the community has been engaged, in order to develop and maintain tools that can keep pace. The task of preserving web content has also necessitated fresh thinking about digital preservation as a discipline (Day, 2006). This chapter, however, leaves these questions aside, to concentrate on what might be termed the cultural history of the movement. It does not address the question of how web archiving has been carried out, but *why*, *by whom*, and *on whose behalf*.

Historians have for long known that, in order to interpret archival materials properly, it is first necessary to understand how that archive came into being. Why is a particular object to be found, and not another? What does the archive seek to document, and whose interests does it serve? The last very few years has seen a very welcome growth in interest in the archived web among scholars (see, for example, Brügger & Schroeder 2017). However, that interest is not yet accompanied by the necessary familiarity with how the archived web came into being, and to be thus familiar is arguably even more important in this context than for traditional paper-based archives. Older distinctions with which historians are familiar—between published document, 'grey literature' and institutional records—have become blurred, as have those between personal and institutional publication. As a result, it has become less clear where the responsibility for preserving which types of content lies among the established institutions in the library and archives field. In addition, the archived web resource is unlike the live version from which it was derived in subtle and complex ways that do not apply to print publications or to manuscripts (Brügger & Finnemann, 2013, pp. 74–76). If this chapter serves to orient users as to some of the questions they should be asking of their sources, and of the institutions that provide them, it will have achieved its aim. It dwells on certain projects and organisations as illustrative of more

general trends. Proceeding in a broadly chronological order, it begins where most narrations of the story have begun, with the Internet Archive.

The Internet Archive (1996–)

Insofar as the general public are aware of web archiving at all, it is likely that the Internet Archive and its Wayback Machine is the thing they know. This is hardly surprising, since the Archive 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. By contrast, the majority of other web archives restrict their collections either by geography or by subject matter, and (in the case of many of the national libraries) are required to impose restrictions on access, due to the legal frameworks under which they operate.

The story of the Internet Archive is relatively well-known (see Kimpton & Ubois, 2006; Livingston, 2007, pp. 274–278). The Archive's founder, Brewster Kahle, had already developed the Wide Area Information Server, acquired by AOL for a multi-million dollar sum. In 1996 he founded two organisations: the Internet Archive, as a not-for-profit organisation (with Bruce Gilliat), and Alexa Internet, the business model of which was based on the analysis of data describing usage patterns online. (Alexa was also later sold, this time to Amazon.) The early holdings of the Archive were composed of the content first collected by Alexa, although over time the Archive began to capture content in its own right. In 2001 the Archive launched the Wayback Machine, the first browser-based access mechanism to archived content. The software on which the Machine was built, also known as Wayback, remains the most widely used means of enabling access to archived web pages. Similarly dominant have been the successive versions of Heritrix, the web crawler application built by the Archive to enable the capture of content. By

2006 the Archive had already collected some 50 billion individual web pages and was serving 70,000 visitors per day; at the time of writing it held some 472 billion archived objects (Kimpton & Ubois, 2006, p. 203).

The achievement of the Internet Archive is an extraordinary one. From the very beginning Kahle was aware of the many technological and legal obstacles in the path of successful web archiving: obstacles which even now still preoccupy the web archiving community. Despite this, the Archive pressed ahead with archiving, motivated by both the fragility of web content and the rate at which it which disappeared, and by the possibilities offered to users in the future (Kahle, 1997). This was in line with a realisation in the mid-1990s of a need to avoid the period becoming known as "a digital Dark Ages" exacerbated by the euphoria and cultural amnesia of the newly emerging internet industry, an "epoch of forgetting" (Kuny, 1997, p. 1, citing Umberto Eco). The Internet Archive remains the only web archive for a substantial majority of national domains.

Recent years have seen a significant growth in mainstream press coverage of web archiving, and of the Internet Archive in particular. As a result, Kahle has had something of the status of a hero thrust upon him, as shown by the 2015 campaign to promote him as the new Librarian of Congress. The Archive is headquartered in San Francisco, and in one sense its story is a classic Californian story: of an entrepreneur with a disruptive idea, creating an organisation the history of which is characterised by (in the words of a well-informed observer) "the dual themes of visionary experimentation and whimsy" (Scott, 2015). This story of the Archive has tended to obscure other streams of web archiving activity, carried out by different kinds of organisations acting in response to different drivers. It is to these other streams that we now turn.

National libraries

At the same time that the Internet Archive was founded, national libraries on three continents were also taking their first steps towards systematic archiving of the web. In Canada, the issue was first discussed in 1994 by the Executive Committee of the National Library of Canada (now part of Library and Archives Canada), leading to the Electronic Publications Pilot Project which reported in 1995. The Library's historic remit included the duty "to collect, preserve and promote access to Canada's published heritage", now understood to include publications in whichever format, whether print, physical storage media such as disks, or delivered via the internet. (National Library of Canada, 1996).

The National Library of Australia, under the National Library Act of 1960, had a similar remit to maintain a comprehensive collection of materials relating to Australia and the Australian people. As in Canada, it was seen as a natural extension of that remit to take in material made available via the internet, and the PANDORA project was established in 1996, with harvesting of content beginning the following year. Faced with the need to obtain permission from the owners of websites to harvest their material, and a simple lack of resources, the NLA took a pragmatic decision to take a selective approach from the beginning (Koerbin, 2004, pp. 1–2, 2016).

This selective mode has been one of two patterns into which national library archiving has subsequently fallen, often although not always on a permissions basis. As such, many collections of web material exist, created by decisions by subject experts as to scope and importance, and structured variously by content type (such as blogs, or news media), by theme (such as climate change), or by events, such as elections. In fact, several web archiving programs have begun with election collections, since consensus about their importance is relatively easy to achieve. In 1996 the Internet Archive collected the sites of candidates for the presidency of the United

States, in partnership with the Smithsonian Institution, and in 2000 collected sites related to the election on behalf of the Library of Congress (Kimpton & Ubois, 2006, pp. 202–203). In Denmark, a test case was provided by the 2001 municipal elections (Brügger, 2016).

In Sweden, the Royal Library had been responsible for collecting, preserving and providing access to Swedish printed publications since 1661. As in Canada and Australia, the archiving of the web as a distribution mechanism closely analogous to publication was viewed as a natural extension of that remit. As a result, the Kulturarw3 project was begun by the Royal Library in 1996. In contrast to the Australian case, the Swedish project took a comprehensive approach, for several reasons: it was more cost-effective than a selective approach, since the latter involved the deployment of human effort on a very large scale, and also because "[o]ne doesn't know what information future generations will consider important" (Arvidson, Persson, & Mannerheim, 2000). This agnosticism about the relative potential value of different kinds of content has been a common theme in subsequent comprehensive web archiving.

At this point, the history of web archiving becomes enmeshed with the larger history of systems of legal deposit. Several states have centuries-old systems of legal deposit that entitle organisations such as national libraries to receive copies of everything published within that jurisdiction. In nations where print legal deposit was already in force there have been moves to extend that legal framework to cover non-print content. One of the first nations to implement a new law was Denmark, in 1997, although in 2004 it was to be substantially revised and its scope widened. The relevant act for New Zealand was the National Library Act of 2003, which coincided with the Legal Deposit Libraries Act in the United Kingdom (Elliott, 2011; Field, 2004; Larsen, 2005). Several other nations have followed suit, including France in 2006 (Aubry, 2010).

To be sure, the implementation of these schemes varies between nations. The types of content that are covered have varied, with exclusions applied to audio-visual content in the UK, for instance. The national web sphere has been defined in various ways: by country-code Top Level Domains, by domain registration, by the physical location of the hosting server, by the intended audience, and by language, or by some combination of those criteria. However, from the point of view of the present cultural history, there were certain key similarities between the contexts in which these frameworks have been formed.

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. At the highest level, there was international collaboration from the first, in the shape of a working group on non-print legal deposit set up by the Conference of Directors of National Libraries, that worked between 1994 and 1996 (Field, 2004, p. 90). The International Internet Preservation Consortium, formed in 2003 by the Internet Archive and a nucleus of national libraries, has been of vital importance (Illien, 2011). However, the location of this effort within institutions so steeped in print culture has tended to shape that effort in particular and not always helpful ways.

Denmark first revised its legal framework to allow the Royal Library to collect non-print content in 1997. However, in relation to online content, the revised law applied only to materials that had the character of print publications, and thus excluded the bulk of the web. The inadequacy of this approach soon became apparent to the libraries concerned (Henriksen, 2016; Larsen, 2005, p. 81). The same point for scholarly users was brought home forcibly in 1999 to one media studies specialist, Niels Ole Finnemann of Aarhus University, when the website about

which a graduate student was about to submit a thesis was suddenly and radically changed (Finnemann, 2015). This event was in part responsible for a press release by Finnemann and his colleague Niels Brügger, announcing their intention to work towards the establishment of a Danish web archive. This catalysed the formation of a partnership with representatives of the Royal Library in Copenhagen and the State and University Library in Aarhus which led in turn to the establishment of netarkivet.dk, the Danish web archive (Brügger, 2016).

At this stage (2002), there was an institutional basis for archiving of the Danish web, but not yet the legal backing. In the process that then led to the revised legislation in 2004, the Danish case is highly unusual in that the interests of researchers were represented, by the presence of Niels Ole Finnemann on the committee that helped draft the legislation. The law when passed also stipulated that there be a standing editorial committee, including researchers, to guide and inform the development of netarkivet.dk (Larsen, 2005).

A common feature of most web archiving backed by legal deposit legislation is some sort of restrictions on the access afforded to the end user of the archive. In cases where archiving is limited to a single copy of a work in a particular institution, it is possible to see the ghost of the print legal deposit paradigm: a curious paradigm to apply to the web. It is also in the development of these restrictions that one can see most clearly the interplay of the interests of the three key stakeholders: the libraries, the owners of the content (and the established media companies in particular) and the end user. In different contexts greater or lesser emphasis has been placed on the different reasons for restricting access: copyright and the rights of content owners to exploit their intellectual property; the risk to the libraries of republishing libellous material or other content that is in breach of the law; and the treatment of sensitive personal data relating to individuals. Naturally much of the process leading to new legislation was not

documented publicly, but from those accounts that have emerged it would seem that in at least some cases the influence of the larger commercial publishers has weighed disproportionately heavily.

One such account is that of Andrew Green, former Librarian of the National Library of Wales and participant in the highly protracted process that led from the initial discussions over non-print legal deposit in the UK in 1997 to the final implementation in 2013. Green noted a "mutual suspicion—sometimes bordering on hostility" between librarians and publishers, particularly the news media companies. The latter were part of an industry on the defensive against commercial pressure, "and defensiveness often breeds aggression, and it is no surprise that newspaper owners, who are under most market pressure, proved the least tractable interlocutors" (Green, 2012, p. 105). In Green's account, even after the 2003 Act restricted access to library premises, thus removing any significant threat to prevailing business models, the publishers pressed for further restrictions. As a result, at the time of writing, users of the Legal Deposit Web Archive in the UK is permitted to print only a small proportion of an archived page, may not make digital copies of any sort, and may not consult an archived resource simultaneously with any other user at the same library: this last restriction being the single-copy model of print legal deposit combined with commercial pressure to produce a manifest absurdity.[1]

The full history of the development of non-print legal deposit must of course wait until minutes of private meetings become publicly available. When that story is told, it will require an articulation with the histories of other movements in media and publishing, including the Open Access movement for scholarly literature, and the radical disruption in traditional markets for

_____

1. As engagement manager for the UK Web Archive at the time the 2013 regulations came into force, when making public presentations I was often met with little short of incredulity from users when outlining these restrictions.

news, both print and broadcast (for which see, for example, Burns & Brügger, 2012; Ji & Waterman, 2014). Indeed, the story may be one of a clash of cultures, between owners of valuable intellectual capital and advocates of freer dissemination of the products of human effort, in which librarians have found themselves in a perhaps somewhat surprising alliance with some of the rhetoric surrounding Silicon Valley and the argument that "information wants to be free". For now it is reasonable to note, with Andrew Green, that delays in the process leading to the implementation of non-print legal deposit have led to the loss of very significant bodies of content from the most formative years of the live web, for which users must rely almost entirely on the Internet Archive (Green, 2012). In addition, the fact that the Danish case is so exceptional in having a strong representation of academic users from the very beginning shows the degree to which the needs of the end user have been relatively neglected in the midst of often confrontational negotiations between libraries and publishers.

Web archiving as the corporate record

Thus far, this chapter has been concerned with organisations making archival copies of other organisations' content: either as part of a national responsibility for the published record or—as in the case of the Internet Archive—in pursuit of a more generalised philanthropic goal. The second half of the period under discussion saw a further strand of web archiving activity emerge in response to quite different drivers: the archiving by organisations of their own content. Within this broad movement there have been several distinct streams.

Scholars of politics and government have noted the simultaneous shift in many countries towards the delivery of government services on a 'digital by default' basis, particularly since 2011 (Lips, 2014). In some contexts, this has necessitated a reinterpretation of the traditional

demarcation between official publications (usually considered part of the published record), and a public or government record, traditionally managed in paper form and the responsibility of a national archival administration. The dividing line became especially hard to see clearly as government activity online widened from the simple delivery of documents to include general communication and the conduct of transactions between state and citizen via web interfaces.

In by no means all countries have national archives engaged with web archiving: in some cases the task has been left in the hands of other organisations. Two examples, one from the USA and one from Europe, will illustrate where such engagement has taken place. The National Archives of the United Kingdom were among the earliest to institute a comprehensive program for archiving government sites. This was a consequence of two movements within government: a 1999 decision that all newly-created public records were to be stored and retrieved digitally by 2004, and a target set (first for 2008, then for 2005) that all services to business and to the citizen should be delivered online. In consequence, it was determined that the websites used to deliver those services should perforce be considered as public records, and not just documents delivered via those services. The UK Government Web Archive was formally founded in 2003 after a period of experimentation begun in 2001 (Brown, 2006, pp. 178–179).

In the USA, the responsibility for government web archiving has been shared between institutions, and in different combinations at different times. Some of the earliest government web archiving took place not under the auspices of the National Archives and Records Administration (NARA), but as part of the Federal Depository Library Content Partnerships Program. This was a continuation of an established tradition of distributed collection of government publications by federal deposit libraries, under the overall direction of the Government Printing Office. The priority was the websites of federal agencies that had ceased

operation, such as the Advisory Commission on Intergovernmental Relations, archived in 1996 by the Libraries of the University of North Texas (Advisory Commission on Intergovernmental Relations [ACIR], 1996; Hartman, 2000, 2016). In 2000–2001 the NARA first took a single snapshot of federal government websites for the USA in connection with the end of the presidential term of Bill Clinton, followed in 2004 by a similar collection at the end of the first term of George W. Bush. Quite separately, the NARA has also been harvesting Congressional websites since 2006. However, in 2008 the NARA issued guidance that placed responsibility for preservation of federal agency web estate back in the hands of individual agencies (National Archives and Records Administration [NARA], 2008). As a result, the 'end of term' collection in 2008–2009 and in 2012–2013 was carried out by a group of agencies in collaboration: the Library of Congress and the Government Printing Office (from within government) and the University of North Texas, the California Digital Library (part of the University of California) and the Internet Archive.[2]

Governments have not been the only kind of organisation that has wished to archive its own web content. Since the mid-2000s universities, schools, churches, commercial organisations and many other organisations besides have done so. However, few of these organisations have chosen to create a full web archiving programme within their own walls, since the costs in IT infrastructure are considerable, and the specific skills required often in short supply. As such, the growth of a small but global group of organisations providing web archiving services has made outsourcing an option. The Internet Archive for a time provided such contracted services, for instance to the National Archives of the UK from 2003. The Internet Archive was also instrumental in the foundation of the European Web Archive in Amsterdam in 2004, a non-profit organisation providing similar services in Europe (Brown 2006, pp. 18, 180–181). The European

2. The End of Term Web Archive may be accessed at http://eotarchive.cdlib.org

Archive became the Internet Memory Foundation, offering web archiving services via its Internet Memory Research subsidiary. In 2006 the Internet Archive itself also launched its Archive-It service, delivered via a web application allowing easy management of the process by its clients.

These two services—Internet Memory Research and Archive-It—at the time of writing remain the two principal outsourcing services for the creation of web archives that are available freely online to end users. Both organisations have been heavily involved in the wider development of the web archiving community, with a significant degree of crossover of personnel. One of the founders of the European Archive was Julien Masanès, who had previously led the web archiving program at the Bibliothèque nationale de France from 2000. Masanès had been one of the instigators of the IIPC, and also of the series of conferences known as the International Web Archiving Workshop, which ran annually from 2001 to 2010.[3]

The same period saw the inception of attempts to provide web archiving services commercially. One early example of this was Hanzo Archives, incorporated as a limited company in the UK in 2005 by two former members of the web archiving program at the British Library, Mark Middleton and Mark Williamson, with Julien Masanès as a member of the board of directors (Hanzo Archives, 2006). Since that time, several other firms have been set up to serve the market, including amongst others Pagefreezer (Netherlands and Canada) and Aleph Archives (Switzerland, USA and Canada). It is more difficult to assess how widely these services are used, since one of the distinguishing features is that the archive is closed to everyone but the staff of the client. The value proposition is also articulated in different terms to that by Archive-It and Internet Memory Research, being in terms of enabling corporations to meet legal requirements in relation to disclosure of information, and as a defence against litigation. Already

_____

3. The proceedings of IWAW are available at http://iwaw.net

by 2005 there were cases coming to courts around the world that involved the use of archived web pages as evidence ("Keeper of expired web pages," 2005).

Research-driven archiving

The availability of outsourcing services, and in particular Archive-It, enabled a wide range of organisations to enter the web archiving arena. One particularly significant group are those scholarly organisations, mostly universities, who have begun to archive content in support of their library content development: a form of archiving in close articulation with the needs, known or inferred, of particular groups of scholars. This movement has proved particularly strong in the USA. One early example is that of Columbia University in New York, which (as well as archiving its own content) has created research collections on subjects including human rights (from 2008) and religious life in New York City (from 2010). The former is a project of the Center for Human Rights Documentation and Research which, although located within the Columbia University Libraries, engages directly in education and research activities as well as acquiring collections for research. One of the selection criteria is the relevance of the content to "current research, teaching and advocacy" (Centre for Human Rights Documentation and Research [CHRDR], 2016).

Examples of this kind of subject-based archiving are relatively few outside the USA, but one example, and possibly the earliest of all, is DACHS, the Digital Archive for Chinese Studies. DACHS was a joint venture between two specialist Sinological institutes, in the universities of Heidelberg and Leiden, although it began first in Heidelberg. Although the project was and is managed by librarians on an operational level, the initial impetus was directly from academics and first expressed in 1999; archiving began in 2001. Perhaps unsurprisingly, there was a keen

sense of the unusual fragility of the Chinese web, given the political situation in that country and the widespread use of censorship even at that time, and so the archive focussed specifically on social and political discourse. There was also a realisation that the Internet Archive and other large scale projects could not be expected to capture content for any particular subject area at the optimal depth and frequency, and so specialist organisations would have to meet that need. To aid selection, the project also drew on the the accumulated knowledge of a distributed group of collaborators—scholars and 'netizens' both within and outside China some of whom were active participants in the discourses concerned. This model of distributed participant curation is one that has rarely been emulated elsewhere, and even in this case the resources required to construct and maintain such a network have proved significant (Lecher, 2006, 2016).

Activist archiving

It may become clear after further research that the few years either side of 2010 saw a shift in the way in which the story of the web was understood by at least some of its users. According to this new narrative of web history, the individualistic spirit that had characterised the early years had given way to an increased colonisation of the web by authoritarian governments, corporate lobbyists, and technology companies with overreaching ambition (see, for instance, Jeanneney, 2007; Morozov, 2011). In place of a web with many relatively small publishers on the one hand and archivists on the other, there were now three kinds of participant: large content organisations, the individual users who entrusted their content and data to them, and the archivists charged with keeping the record.

All of the web archiving programmes examined so far have indeed been programmes: planned activity carried out by organisations in line with their wider mission and purpose. In part

because of the scale at which these programmes have operated, and the relative accessibility of the archived content, they have tended to be more prominent. There is, however, an important strand of web archiving activity that tends to be overlooked as a result: the work of individuals and small groups, responding to a particular cause. One such is the Dale Askey archive, concerning the 2012 libel suit against the academic librarian Dale Askey, then of McMaster University in Canada, which raised questions of freedom of speech and the appropriate use of the law of libel. Members of the Greater Toronto Chapter of the Progressive Librarians' Guild, seeing a fast-developing online event which would not be captured by the periodic crawls of the Internet Archive or other institutions, came together as individuals to begin capturing key discussions of the case. Using a combination of open source tools, the Dale Askey Archive was subsequently made publicly available. Even though in 2012 all the major components of the web archiving landscape were in place, there were still other ways for the librarian, acting personally but guided by "the professional ethics of libraries and archives, to choose a community to document, preserve, and support" (Milligan, Ruest, & St. Onge, 2016).

The #freeDaleAskey team were clear that their work was within the remit of the librarian and archivist, broadly conceived, and not a call to the profession to become citizen journalists or community activists. There has however been a strand of web archiving which approaches such a status, the most prominent example of which has been the Archive Team. In 2008 Jason Scott noted the readiness of corporations to discontinue online services that were no longer profitable, often with the loss of user-generated content of significant value both to its creator and to later scholars. Motivated by the shutting-down of AOL Hometown in late 2008—which Scott described as an 'eviction' of people from their webspace—the volunteer-run Archive Team was created (Scott, 2008, 2011). Its most public case was follow in 2009 with the closure of Geocities

by Yahoo, at which several million individual websites disappeared in an instant, but of which the Archive Team, a "loose collective of rogue archivists, programmers, writers and loudmouths dedicated to saving our digital heritage" were able to capture a subset, numbering in the millions (Archive Team, 2016).

In one sense, both the Archive Team and the Dale Askey campaign represent a return to an approach closer to that of the Internet Archive than of the national libraries. A rapid response was required in order to save content that would not be archived by any of the existing institutional programmes. It was a pragmatic approach, characterised by a willingness to press ahead and archive content despite some risk relating to breaches of copyright law: risks which national libraries, by their nature, rarely contemplate taking. Both ventures were motivated by a sense of public duty, and a particular political and social vision of the kind of space that the web should be. They also represent a response to a new configuration of stakeholders after Web 2.0: publishers, users who create content, and archivists who set out to document the relationship and (at times) to redress the balance of power between them. This new articulation of interests was significantly different from the binary library-publisher relationship that so profoundly shaped the development of non-print legal deposit.

Web archiving in 2016 and the future

If the history of web archiving is now a story of 20 years, from 1996 to the time of writing, then by the mid-way point of 2006 the movement had taken its present institutional shape. The International Internet Preservation Consortium had been established, giving a global point of reference for the community of web archiving practitioners. The two key technologies—Heritrix for large-scale crawling, and Wayback for replay of content—were both in general use.

Comprehensive legal deposit frameworks for web harvesting had been formulated and put into force in several countries. Outsourcing services had become available for organisations to archive their own content, or (in the case of research-driven archiving) the content of others for research purposes. Significant publications attempting to survey the whole scene had also begun to appear (Brown, 2006; Brügger, 2005; Masanès, 2006).

I have attempted to show that the shape of each of these component pieces of that organisational pattern was a product of the interplay between institutions, their perception of their mission, and the interests (sometimes competing) of the various stakeholders in each context. A larger study (which the topic would certainly merit) would be able to tease out the complexities of these relationships in each national situation, and the growth and influence of the global web archiving community. Its approach might be exhaustive where the current chapter can only be selective, and would involve a very significant programme of oral history interviews.

The missing piece from this picture, in 2006, was the researcher, as the end user of the archive. Although the Association of Internet Researchers was well established, having begun to hold its annual conferences in 2000, there was yet little engagement with the archived web as an object of study.[4] There were, to be sure, scholars beginning to use the archived web (Brügger, 2005; Foot & Schneider, 2004), but in relative isolation. Possibly the first international conference to take up the theme took place in 2008 on the fringes of the Association of Internet Researchers conference in Copenhagen; several of the papers were subsequently published

4. For a periodisation of the discipline of Internet Studies, see Wellman (2011). In the case of the Association, an important milestone was a workshop on the fringes of the 2004 conference in London, at which scholars engaged with members of the IIPC. See, for instance, the paper given by Alex Halavais, at http://alex.halavais.net/blogs-and-archiving (retrieved June 16, 2016). I am grateful to the anonymous reviewer for drawing this meeting to my attention.

(Brügger, 2010). The first PhD from within the social sciences and humanities to use the archived web was that by Meghan Dougherty, a student of Kirsten Foot at the University of Washington (Dougherty, 2007).

Understandably, the attention of the web archiving community in the early years was focussed on developing the necessary tools to capture web content, the mechanisms by which that data might be preserved, and the organisational work of integrating web archiving in existing and often ancient institutions. If some of the access mechanisms have not served all the possible uses that researchers might have wanted, this was understandable under these circumstances, and given the small number of researchers with whom libraries and archives could engage.

Happily, recent years have seen a growing interest, both amongst researchers and from institutions engaged in web archiving, in collaborating in order to inform both selection decisions and the development of access services. This was prefigured by the Danish collaboration noted above, and by webarchivist.org, a collaboration between researchers at the State University of New York, the University of Washington, the Library of Congress and the Internet Archive, which began in 2001 and continued until 2010 (Foot, Schneider, Xenos, & Dougherty, 2003). More recently, other examples include the collaborative curation project named Researchers and the UK Web Archive that ran between 2010 and 2011 (Webster, 2010), and the two projects in the UK to co-design a new search interface for British Library data (with acronyms of AADDA and BUDDAH) which between them ran between 2011 and 2015.[5] It is to be hoped that the next 20 years are characterised more and more by just this collaboration between archivists and their users.

_____

5. The project blogs may be found at http://domaindarkarchive.blogspot.co.uk/ and http://buddah.projects.history.ac.uk/.

Acknowledgements

References

Advisory Commission on Intergovernmental Relations. (1996). Homepage, now in University of North Texas Digital Library. Retrieved May 4, 2016 from http://digital.library.unt.edu/ark:/67531/metadc800/

Archive Team (2016). Homepage. Retrieved May 3, 2016 from http://www.archiveteam.org

Arvidson, A., Persson, K., & Mannerheim, J. (2000). *The Kulturarw3 Project: The Royal Swedish web archive—An example of 'complete' collection of web pages*. Paper given at 66th Council and General Conference of the International Federation of Library Associations and Institutions (IFLA), Jerusalem. Retrieved April 15, 2016 from http://archive.ifla.org/IV/ifla66/papers/154-157e.htm

Aubry, S. (2010). Introducing web archives as a new library service: The experience of the National Library of France. *LIBER Quarterly*, *20*(2), 179–199.

Brown, A. (2006). *Archiving websites: A practical guide for information management professionals*. London: Facet.

Brügger, N. (2005). *Archiving websites: General considerations and strategies*. Aarhus: Centre for Internet Studies.

Brügger, N. (Ed.). (2010). *Web history*. New York, NY: Peter Lang.

Brügger, N. (2011). Web archiving—Between past, present and future. In M. Consalvo & C. Ess (Eds.), *The handbook of internet studies* (pp. 24–42). Chichester: Wiley-Blackwell.

Brügger, N. (2016). Interview with the author, March 14, 2016.

Brügger, N., & Finnemann, N. O. (2013). The web and digital humanities: Theoretical and methodological concerns. *Journal of Broadcasting and Electronic Media*, *57*(1), 66–80.

Brügger, N., & Schroeder, R. (Eds.). (2017). *The web as History: The first two decades*. London: UCL Press.

Burns, M., & Brügger, N. (Eds.). (2012). *Histories of public service broadcasters on the web*. New York, NY: Peter Lang.

Centre for Human Rights Documentation and Research. (2016). Human Rights Web Archive. Retrieved May 5, 2016 from http://library.columbia.edu/locations/chrdr/hrwa.html

Day, M. (2006). The long-term preservation of web content. In J. Masanès (Ed.), *Web archiving* (pp. 177–199). Berlin: Springer.

Dougherty, M. (2007). *Archiving the web: Documentation, display and shifting knowledge production paradigms* (PhD thesis). University of Washington.

Elliott, A. (2011). *Electronic legal deposit: The New Zealand experience*. Paper given at conference of the International Federation of Library Associations and Institutions (IFLA), San Juan, Puerto Rico. Retrieved April 1, 2016 from http://www.ifla.org/past-wlic/2011/193-elliott-en.pdf

Field, C. D. (2004). Securing digital legal deposit in the UK: The Legal Deposit Libraries Act 2003. *Alexandria*, *16*(2), 87–111.

Finnemann, N. O. (2015). Speech at tenth anniversary of Netarkivet.dk, Aarhus, June 2015.

Foot, K., & Schneider, S. (2004). The web as an object of study. *New Media & Society*, *6*(1), 114–122.

Foot, K., Schneider, S., Xenos, M., & Dougherty, M. (2003). Opportunities for civic engagement on campaign sites. Retrieved June 22, 2016, from https://web.archive.org/web/20080201083014/http://politicalweb.info/reports/engagement.html

Green, A. (2012). Introducing electronic legal deposit in the UK: A Homeric tale. *Alexandria, 23*(3), 103–109.

Hanzo Archives (2006). Annual company return, 1 April 2006. Retrieved June 22, 2016, from https://beta.companieshouse.gov.uk/company/05410483/

Hartman, C. N. (2000). Storage of electronic files of federal agencies that have ceased operation: A partnership for permanent access. Retrieved June 14, 2016 from http://digital.library.unt.edu/ark:/67531/metadc181693/

Hartman, C. N. (2016). Interview with the author, 21 April.

Henriksen, B. N. (2016). Interview with the author, 15 April.

Illien, G. (2011). Une histoire politique de l'archivage du web. *Bulletin des bibliothèques de France*, *2*. Retrieved December 1, 2013 from http://bbf.enssib.fr/consulter/bbf-2011-02-0060-012

Jeanneney, J.-N. (2007). *Google and the myth of universal knowledge: A view from Europe*. Chicago, IL: Chicago University Press.

Ji, S. W., & Waterman, D. (2014). The impact of the internet on media industries: An economic perspective. In M. Graham & W. H. Dutton (Eds.), *Society and the internet: How networks of information and communication are changing our lives* (pp. 149–163). Oxford: Oxford University Press.

Kahle, B. (1997, March 1). Preserving the internet: An archive of the internet may prove to be a vital record for historians, businesses and government. *Scientific American*. 276 (3).

Keeper of expired web pages is sued because archive was used in another suit. (2005, July 13). *New York Times,* p. C (L).

Kimpton, M., & Ubois, J. (2006). Year-by-year: From an archive of the internet to an archive on the internet. In J. Masanès (Ed.), *Web archiving* (pp. 201–212). Berlin: Springer.

Koerbin, P. (2004). *Managing web archiving in Australia: A case study*. Paper given at IWAW (International Web Archiving Workshop), Bath (UK), 2004. Retrieved May 1, 2016 from http://iwaw.net/04/

Koerbin, P. (2016). Interview with the author, May 4, 2016.

Kuny, T. (1997). A digital dark ages? Challenges in the preservation of electronic information. Paper presented at the 63rd Council and General Conference of the International Federation of Library Associations and Institutions (IFLA), Copenhagen. Retrieved May 1, 2016 from http://archive.ifla.org/IV/ifla63/63kuny1.pdf

Larsen, S. (2005). Preserving the digital heritage: New legal deposit act in Denmark. *Alexandria*, *17*(2), 81–87.

Lecher, H. (2006). Small scale academic web archiving: DACHS. In J. Masanès (Ed.), *Web archiving* (pp. 213–226). Berlin: Springer.

Lecher, H. (2016). Interview with the author, April 20, 2016.

Lips, M. (2014). Transforming government—By default? In M. Graham & W. H. Dutton (Eds.), *Society and the internet: How networks of information and communication are changing our lives* (pp. 179–194). Oxford: Oxford University Press.

Livingston, J. (2007). *Founders at work. Stories of startups' early days*. Berkeley, CA: Apress.

Masanès, J. (Ed.) (2006). *Web archiving*. Berlin: Springer.

Milligan, I., Ruest, N., & St. Onge, A. (2016). The great WARC adventure: Using SIPS, AIPS and DIPS to document SLAPPS. *Digital Studies/Le Champ Numerique*, 2016. Retrieved June 14, 2016 from https://www.digitalstudies.org/ojs/index.php/digital_studies/article/view/325/412

Morozov, E. (2011). *The net delusion: How not to liberate the world*. London: Allen Lane.

National Archives and Records Administration. (2008). Web harvest background information [15 April]. Retrieved June 14, 2016 from http://www.archives.gov/records-mgmt/memos/nwm13-2008-brief.html

National Library of Canada. (1996). Electronic Publications Pilot Project (EPPP). Summary of the final report. Retrieved April 22, 2016 from http://epe.lac-bac.gc.ca/100/200/301/nlc-bnc/eppp_summary-e/ereport.htm

Scott, J. (2008). Eviction, or the coming datapocalype. Retrieved May 1, 2016 from http://ascii.textfiles.com/archives/1617

Scott, J. (2011). Presentation at Personal Digital Archiving conference [Internet Archive]. Retrieved April 1, 2016 from https://archive.org/details/PDA2011-jasonscott

Scott, J. (2015). The case for #DraftBrewster. (n.d.). Retrieved April 11, 2016 from https://medium.com/@textfiles/the-case-for-draftbrewster-abca1fd3cf71

Webster, P. (2010). Using the UK Web Archive. Retrieved June 22, 2016 from https://peterwebster.me/2010/12/03/using-the-uk-web-archive/

Wellman, B. (2011). Studying the internet through the ages. In M. Consalvo & C. Ess (Eds.), *The handbook of internet studies* (pp. 17–23). Chichester: Wiley-Blackwell.