The International Journal of Digital Curation
Issue 2, Volume 4 | 2009

Missing Links: The Enduring Web

Marieke Guy, Alex Ball, Michael Day
UKOLN, University of Bath

Summary

The Web runs at risk. Our generation has witnessed a revolution in human communications on a trajectory similar to that of the origins of the written word and language itself. Early Web pages have an historical importance comparable with prehistoric cave paintings or proto-historic pressed clay ciphers. They are just as fragile. The ease of creation, editing and revising gives content a flexible immediacy: ensuring that sources are up to date and, with appropriate concern for interoperability, content can be folded seamlessly into any number of presentation layers. How can we carve a legacy from such complexity and volatility?

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. ISSN: 1746-8256 The IJDC is published by UKOLN at the University of Bath and is a publication of the Digital Curation Centre.

Introduction

The opening lines on the Digital Preservation Coalition Web page[1] for the Missing Links workshop pull no punches. Web resource preservation is an important area from the institutional level upwards and the one-day event sponsored by the Digital Preservation Coalition (DPC) and the Joint Information Systems Committee (JISC) along with the six partners of the UK Web Archiving Consortium (the British Library, the National Library of Wales, JISC, the Wellcome Library, The National Archives and the National Library of Scotland) was an attempt to validate the work in hand. Its key aim was to develop and strengthen the links between content creators, tool developers, preservation services and users in order to "secure an enduring Web".

Many of the UK institutions most committed to, and interested in, Web preservation were represented at the event, held at the British Library in London. There were over 100 delegates, representing academia, galleries, museums, libraries, archives, government, research groups, professional societies, and so forth. While back in the early 1970s no one had thought to keep the very first e-mail message, the wide range of participants at this workshop demonstrated that there is a great deal of interest in exploring what needs to be done to ensure that Web content endures.

Session 1: Setting the Scene

Keynote: Adrian Brown, Assistant Clerk of the Records, the Parliamentary Archives

Adrian Brown (formerly of The National Archives) kicked off the workshop with a keynote address. In it he noted that there has always been an interest in Web archiving, but that recent years have seen a flurry of activity as the consequences of link rot (and similar problems) become ever more apparent. (He also bravely suggested that sites hosted by the endangered Geocities service were cultural icons that should be preserved, if only to torture future generations - or words to that effect.) Adrian adeptly set the scene and described the main challenges ahead. His talk centred on the key jigsaw pieces of Web archiving: selection, capture, storage, access and preservation. He pointed out that the Web used to be primarily a medium for fixed content and it is the new types of interactive content (e.g., discussions on wikis, Twitter, blogs, etc.) that now offer the biggest challenge to preservation. Parliament itself is using many new forms of communication, and although gaining permission to archive official forms of content is not always a problem, getting permission to archive content presented on third-party sites often can be. You may own the content that you push to an external site, but another party owns the presentation of that content. Adrian also emphasised the need for us to coordinate selection and be as explicit about what we are not going to capture as what we are (the unpublishing button on Facebook being a classic example of how difficult it can be to go back!). Another major challenge is that of temporal cohesion – the rate of capture of content is often far slower than the rate of change of content. He ended by appealing for more work to make the archived Web a seamless part of the current Web.

Following this keynote address, the workshop proceeded at a brisk pace with most speakers squeezing their messages into what seemed a scant twenty minutes.

[1] Missing links: the enduring Web http://www.dpconline.org/graphics/events/090721MissingLinks.html

Web Archive and Citation Repository in One: DACHS: Hanno Lecher, Librarian, Sinological Library, Leiden, Netherlands

After an introduction to DACHS (Digital Archive for Chinese Studies)[2] , Hanno Lecher of Leiden University promoted the idea of a citation repository. Its approach is that, instead of populating a bibliography with links to live resources – links that could easily fail after a short time – one would instead provide a single link to a page in the citation repository. This page would list all the Web-accessible resources in the bibliography and link to archived copies stored within the repository. This is already implemented in DACHS. Hanno pointed out that there are other, less optimal, ways of dealing with the same problem: from banning Web links in bibliographies altogether; to making personal copies of Web resources with tools like Snagit or Zotero; or making public copies of linked Web texts using services like WebCite[3] .

The future of researching the past of the Internet: Eric T. Meyer, Research Fellow, Oxford Internet Institute, Oxford University.

Eric Meyer (Oxford Internet Institute) reported on the World Wide Web of Humanities Project[4] that concluded in March 2009 after collecting and analysing over 6 million Web pages concerning the two world wars, drawn from the Internet Archive. This work was carried out as part of the JISC and National Endowment for the Humanities (NEH) Transatlantic Digitisation Collaborative Grants Programme. While the visualizations of how sites changed over time and how they interlinked were interesting, the most challenging aspect of the research seemed to be assembling the collections in the first place. After pointing out some future avenues of research, Eric concluded by stressing that what researchers really want is a global way to access archives, not a national one.

At this point, the workshop moved from the general to the specific.

Session 2: Creation, Capture & Collection

An overview of Web Archiving Tools by Helen Hockx-Yu, Web Archiving Programme Manager, the British Library

After some general observations on the current size and nature of the UK Web domain, Helen Hockx-Yu of the British Library gave a practical overview of Web archiving tools. This introduced the range of tools currently used to support the selection and capture of Web sites, including harvesting programs like Heritrix and HTTrack, as well as the workflow management tools like NetarchiveSuite, the Web Curator Tool and PANDAS (PANDORA Digital Archiving System) that are used in combination with harvesting programs to manage the entire capture process, including things like permissions, job scheduling, metadata capture and quality control. The presentation also briefly introduced some of the formats used to store harvested content, noting the growing importance of the WARC (Web ARCive) file format (WARC, 2009) developed by the International Internet Preservation Consortium (IIPC) – recently accepted as an international standard (ISO 28500:2009) – the adoption of which was described as "highly desirable from a long-term archival

[2] DACHS (Digital Archive for Chinese Studies) Citation Repository http://leiden.dachs-archive.org/citrep/
[3] WebCite http://www.Webcitation.org/
[4] World Wide Web of Humanities project http://www.oii.ox.ac.uk/research/project.cfm?id=48

standpoint." In addition, a number of tools that would support access to Web archives were also in development. They included an open source implementation of the Internet Archive's Wayback Machine, various WARC-based search tools being developed by Hanzo Archive, and the IIPC's NutchWAX (Nutch + Web Archive eXtensions) search tool. Helen concluded with a brief discussion on the limitations and challenges of the current generation of Web archiving tools, including problems with the "snapshot" approach to Web capture (e.g., ensuring temporal consistency), the difficulty of harvesting dynamic or interactive Web content, the challenges raised by "bad" content (e.g., spam, malware, crawler traps, illegal content), and problems with rendering archived sites. The presentation also made the generic point that community reliance on constantly evolving open source Web archiving tools made such difficulties even more burdensome.

Context and Content: Delivering Coordinated UK Web Archive to User Communities: Cathy Smith, Collections Strategy Manager, The National Archives

Cathy Smith reported on a study conducted on behalf of The National Archives (TNA) by consultants Inforesight entitled "Delivering coordinated UK Web archives to user communities," an analysis of Web archive users funded as part of the JISC's Preservation Programme. Initial findings were based on interviews with Web archive users combined with consultation with the UK Web Archiving Consortium (UKWAC). The study suggested that the main users of a Web archive would include journalists and investigative reporters, detectives, civil servants, Web designers, researchers, and so on, however, it was emphasised that the UKWAC user base, while growing, remained quite small. Findings included the need for a single user view across archives, since users need to find appropriate content without having to worry about which particular crawl led to a resource being archived, or which archive is currently holding that resource. There were also concerns that the selective harvesting approach used by UKWAC, while effective in capturing Web sites in depth, would not be able to scale up to the whole UK Web domain. For this reason, it was suggested that that the "deep" selective harvesting approach currently used by UKWAC could be supplemented by "shallow" whole domain capture approaches – although this would probably require changes in legislation. The study recommended that the UK Web Archive could provide the nucleus of a "National Collection of Web sites," inclusive of Web archiving activities undertaken by those outside the consortium. The recommendations also outlined the roles and responsibilities of individual collections vis-à-vis the "National Collection," and proposed some methods for reducing overlap or duplication from the user perspective, comparing different techniques for coordinating whole-domain and selective crawls.

Capture and Continuity: Broken links and the UK Central Government Web Presence: Amanda Spencer and Tom Storrar, The National Archives.

Amanda Spencer and Tom Storrar of The National Archives were asked to sort out the problem of broken links on UK Government Web sites. Their presentation commenced with a brief overview of recent activities, including the UK Government Web Archive, which has been harvesting selected government Web sites since 2003, and the Transformational Government strategy, a long-term plan for streamlining the government's Web presence in which TNA takes responsibility for its longer-term archiving. The core of the presentation introduced the Web Continuity project[5], which originated in a letter sent in April 2007 to the Cabinet Office Minister from the then

[5] Web Continuity Project http://www.nationalarchives.gov.uk/Webcontinuity/

Leader of the House of Commons (the Rt. Hon. Jack Straw MP) about problems in gaining access to government information online. A subsequent review revealed that 60% of the URLs quoted in Hansard between 1997-2006 were broken, which at best would make site navigation difficult, at worst leading to a potential loss of public trust in the authority of government Web sites. Working with the European Archive, the Web Continuity project has begun to create a comprehensive archive of the entire central government Web space, capturing around 1200 sites three times per year. The project has also developed an open source plugin for installation on government Web servers. Its effect would be that any request which would normally result in a '404 Not Found' error would be instead redirected to the last version of that page stored in the archive, thus providing seamless access to the appropriate resource.

In the following discussion session, the panel was asked what advice it would give to Web managers about the effective archiving of their sites. The suggestions covered both content (following accessibility guidelines, W3C standards, validating code, adding in hard links to files next to streaming content, having transcripts of audio files) and communication (supporting a dialogue between Web archivists and Web creation people). It was noted that there can be a fine line between encouraging good practice and stifling innovation and that, at times, communication with content creators was not always possible. Responsible Web harvesting also meant explaining why one was doing it to those involved.

The discussion then moved to other possible ways of capturing content, including using Google Cache, using browser plugins or working directly with ISPs.

Session 3: Issues and Approaches to Long-term Preservation of Web Archives

After lunch the programme moved on from dealing with current problems to identifying the challenges of the future.

Diamonds in the Rough: Capturing and Preserving Online Content from Blogs: Richard Davis, Project Manager, University of London Computing Centre (ULCC)

Richard Davis of ULCC started his presentation with an assertion that blogs were important, emphasising their potential value as primary sources as well as highlighting their use in supporting research and learning. He introduced the JISC-funded ArchivePress[6] Project, which was concerned with developing tools to support the preservation of blogs. Richard noted that traditional Web archiving tools could be difficult to configure and use, and with blogs might prove to be the proverbial "hammer to crack a nut". A comment made by Chris Rusbridge, Director of the DCC, to the effect that "blogs represent an area where the content is primary and design secondary"[7] suggested that blog feeds might be the key to facilitating their capture and preservation. The ArchivePress Project will experiment with capturing feeds via a WordPress database; test beds will include blogs published by the Digital Curation Centre, the
University of Lincoln and UKOLN. Looking to the future, Richard thought that

[6] ArchivePress http://archivepress.ulcc.ac.uk/
[7] Chris Rusbridge's comment on: "Preservation for scholarly blogs," Gavin Baker's blog, March 30, 2009, from http://www.gavinbaker.com/2009/03/30/preservation-for-scholarly-blogs/ is available at http://www.gavinbaker.com/?p=227#comments

ArchivePress might be a potential means of harvesting Twitter content.

Beyond Harvest: Long Term Preservation of the UK Web Archive: Maureen Pennock, Web Archive Preservation Project Manager, The British Library

Maureen Pennock began her talk by providing some general background to the UK Web Archive[8], including basic information on its size (ca. 5,000 titles, 4Tb) and the tools used for Web site capture. Now that the archive has been established, Maureen commented that the next task would be ensuring that the content could be preserved, which is seen as very much an ongoing activity involving people, processes and systems. In developing a preservation approach, the main focus so far has been on: documenting system dependencies, the consideration of containers and metadata standards, the development of preservation workflows and defining preservation strategies. On container standards and metadata, a review has led to WARC being the preferred file format for preservation, but there is also a proposal to use selected additional metadata features from METS (Metadata Encoding and Transmission Standard) and the PREMIS (Preservation Metadata: Implementation Strategies) Data Dictionary. In addition, Maureen explained that "technology watch" needed to be an embedded activity within a preservation archive, and noted that the UK Web Archive had a Technology Watch blog[9]. The presentation ended with a look to the future, which would involve the consideration of many potentially new areas, for example, whether there was a need to preserve computer viruses.

From Web Page to Living Web Archive: Thomas Risse, Senior researcher, L3S Research Center

Thomas Risse (L3S Research Center, Hannover) provided a short overview of Living Web Archives (LiWA)[10], a research project funded by the European Commission as part of the European Union's Seventh Framework Programme (FP7). LiWA was specifically concerned with developing tools and approaches that would be able to deal with current generations of Web technologies. The presentation first explained what the LiWA consortium was doing with regard to dynamic or streamed content. Chipping away at this problem, LiWA have developed tools that run harvested pages through a JavaScript engine in order to extract auto-generated links that may be hidden in code (they are hoping to do something similar with Flash). Risse also introduced LiWA work on identifying and filtering "spam" content, on supporting temporal coherence (using snapshots from rapidly changing content), and on dealing with changes in semantics over long periods of time.

Emulating access to the Web 1.0: Keep the browser: Jeffrey van der Hoeven, Koninklijke Bibliotheek, The National Library of the Netherlands

Jeffrey van der Hoeven of the Koninklijke Bibliotheek began his presentation with a whirlwind history of the Web browser, from the first World Wide Web browser in 1991 (more like a document viewer), through Mosaic and Netscape, to the current generation of browsers like Firefox and Google Chrome. He explained that today's browsers do many tasks other than rendering Web sites, such as bookmarking, caching, authentication and rendering RSS feeds. Over time, browsers have changed

[8] UK Web Archive http://www.Webarchive.org.uk/
[9] UK Web Archive Technology Watch blog
http://britishlibrary.typepad.co.uk/ukWebarchive_techwatch/
[10] LiWA (Living Web Archives) project http://www.liwa-project.eu/

from hypertext viewers to all round workplaces – the Google Wave[11] collaboration tool being an extreme and recent example of this and van der Hoeven argued that as the browser is increasingly the central application of the computer, maintaining them and continuing to run them through emulator programs would be a sustainable solution to the long-term rendering of archived Web sites. The Koninklijke Bibliotheek has been working on developing emulators in the Dioscuri Project[12] (van der Hoeven, Lohman & Verdegem, 2007) with the Planets Project[13] on the Global Remote Access to Emulation services (GRATE) tool, and the FP7 Keep Emulation Environments Portable (KEEP) project[14].

Session 4: Discussion and Next Steps

What we want with Web-archives; will we win? Kevin Ashley, Head of Digital Archives, University of London Computer Centre (ULCC)

The closing keynote address was given by Kevin Ashley of ULCC who let his imagination run free in considering what future researchers might be interested in. He began his talk with a joke about the forthcoming Web 8.0, noting that when the next major shift in the Web happens, we may need a better metaphor than Web 3.0. He suggested that future users might want Web archives to address things like: how the Web changed society; how particular sites evolved over time; how language was used within a site or across sites; how formats were used (and how they rose and fell); how pages were connected, and how much traffic followed those connections. He also suggested that people may want to perform historical searches – for example, "What results would have come up if I had searched for 'Web archiving' in 1998?" – or conduct historical mashups. He argued that we may need to archive more than just Web sites themselves, suggesting that we may also need server logs, the browser plugins, as well as what more traditional media had to say about the Web.

The workshop ended with a panel session, where delegates discussed in various degrees of depth issues of advocacy, the relationship between subject specialists and Web archivists, permissions to archive and legal deposit, whether we needed to archive DNS registries, and selection policies.

Summing up

The day left most people enthused and positive about the future of Web archiving. However the biggest challenge remains that of convincing the masses that it is something that makes sense. Most are too busy thinking about the now to consider the future. This is something of which the JISC PoWR (Preservation of Web Resources) Project[15], which looked at Web resource preservation for UK HE/FE Institutions, became increasingly aware (Eddie, 2008; Emmott, 2008).

More broadly, the workshop demonstrated that significant challenges remain for practitioners undertaking Web archiving. Above all, a number of the presentations indicated the ever growing complexity and interactivity of the Web domain. This

[11] Google Wave http://wave.google.com/
[12] Dioscuri http://dioscuri.sourceforge.net/
[13] Planets, Preservation and Long-term Access through Networked Services http://www.planets-project.eu/
[14] KEEP (Keeping Emulation Environments Portable), http://www.keep-project.eu/
[15] JISC PoWR (Preservation of Web Resources) http://jiscpowr.jiscinvolve.org/

presents both technical and conceptual challenges to Web archiving. So, for example, Adrian Brown, Helen Hockx-Yu and Thomas Risse's presentations all noted technical limitations with the current generation of Web harvesting tools with regard to things like interactive or dynamically generated pages, temporal consistency, or the configuration of harvesting programs to avoid 'crawler traps.' Even more problematically, a growing amount of Web content is generated on the fly, incorporates services or data from third parties, or can be optimised to fit the preferences of visiting users or browsers. In these circumstances, it can be very difficult to know whether a 'definitive' version of some Web sites can ever exist. Chris Rusbridge's comment on blogs that "content is primary and design secondary" might, therefore, also apply to the Web. The incorporation of third-party content on Web sites creates potential dangers for both users and Web archives. Jonathan Zittrain explains (2008, p.56):

> The Web was designed to seamlessly integrate material from disparate sources: a single Web page can draw from hundreds of different sources on the fly, not only through hyperlinks that direct users to other locations on the Web, but through placeholders that incorporate data and code from elsewhere into the original page ... To visit a Web site is not only to be asked to trust the Web site operator, It is also to trust every third party – such as an ad syndicator – whose content is automatically incorporated into the Web site owner's pages, and every fourth party – such as an advertiser – who in turn provides content to that third party.

At the very least, incorporated Web site content from third and fourth parties will not always respect the traditional collection management boundaries embodied in a concept like a "National Collection of Web sites." Selection and appraisal policies will increasingly need to take account of third-party content and services incorporated within archived Web sites. Similar decisions may need to be made about the wisdom of preserving "bad" content like malware or phishing sites.

The focus of a number of workshop presentations on the use of Web archives was extremely welcome. The reminder from Amanda Spencer and Tom Storrar's presentation that Web archives sometimes need to be seamlessly integrated into the current Web environment was extremely timely. Perhaps more fundamentally, however, we may need to think more carefully about how the different organisational motives for collecting and preserving Web sites have a direct influence on how Web archives can be used. For example, Eric Meyer's discussion of the World Wide Web of Humanities project and Kevin Ashley's keynote presentation reminded workshop delegates that Web archiving initiatives probably need to think well beyond the preservation of content or the emulation of Web browsers. For example, there is much of interest that can be learned from historical Web linking behaviour or from a detailed study of the Web graph itself over time, for example from a Webometric (Thelwall, Vaughan & Björneborn, 2005) or network science perspective (Lazer at al, 2009; Newman, 2008). Preserving these aspects of the Web adequately will most likely challenge some of the assumptions that underlie concepts of "national" Web domains and unearth a new range of technical problems that will need to be addressed.

A conference report will be made available on the DPC Web site[16]. and announced on the dpc-discussion and digital-preservation JISCmail lists.

All presentation slides are now available in PDF from the DPC Web site WAC09 was used as the Twitter tag for the event[17].

References

Eddie, C. (2008, October). Embedding Web preservation strategies within your institution. *Ariadne, 57*. Retrieved September 30, 2009, from http://www.ariadne.ac.uk/issue57/jisc-powr-2008-09-rpt/

Emmott, S. (2008, July). Preservation of Web resources: Making a start. *Ariadne, 56*. Retrieved September 30, 2009, from http://www.ariadne.ac.uk/issue56/jisc-powr-rpt/

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A., Brewer, D. et al. (2009). Computational social science, *Science*, *323*(5915), 721-723.

Newman, M. (2008). The physics of networks. *Physics Today*, *61*(11), 33-38.

Thelwall, M., Vaughan, L., & Björneborn, L. (2005). Webometrics. *Annual Review of Information Science and Technology*, *39*, 81-135.

van der Hoeven, J., Lohman, B., & Verdegem, R. (2007). Emulation for digital preservation in practice: The results. *International Journal of Digital Curation, 2*,(2), pp. 123-132. Retrieved September 30, 2009, from http://www.ijdc.net/index.php/ijdc/article/view/50

WARC. (2009). *Geneva: International Organization for Standardization*. ISO 28500:2009: Information and Documentation.

Zittrain, J. (2008). *The future of the Internet.* Allen Lane: London.

[16] Missing links: the enduring Web http://www.dpconline.org/graphics/events/090721MissingLinks.html

[17] Twitter feed for the event #WAC09 http://hashtags.org/tag/wac09/messages