Web Archiving

Alex Ball
Digital Curation Centre, UKOLN, University of Bath

DCC STATE OF THE ART REPORT
Deliverable B 7.5.8

Version: 1.1
Status: Final
Date: 1st March 2010

Copyright

Catalogue Entry

Title	Web Archiving
Creator	Alex Ball (author)
Subject	World Wide Web; Web harvesting; temporal consistency; significant properties; archiving strategy; software tools for Web archiving; malware; Web spam; OAI-ORE; blog archiving
Description	Web archiving is important not only for future research but also for organisations' records management processes. There are technical, organisational, legal and social issues that Web archivists need to address, some general and some specific to types of content or archiving operations of a given scope. Many of these issues are being addressed in current research and development projects, as are questions concerning how archived Web material may integrate with the live Web.
Publisher	University of Edinburgh; UKOLN, University of Bath; HATII, University of Glasgow; Science and Technology Facilities Council
Date	8th January 2010 (creation)
Type	Text
Format	Portable Document Format version 1.4
Language	English
Rights	© 2010 Digital Curation Centre, UKOLN, University of Bath

Citation Guidelines

Alex Ball. (2010). *Web Archiving* (version 1.1). Edinburgh, UK: Digital Curation Centre.

Contents

1 Introduction

Since its invention in 1989 and subsequent release in 1991, the World Wide Web has grown in both size and popularity with such vigour that it is has eclipsed most of the other applications that run on the Internet. Its importance as an information resource is undisputed – the number of reference works scaling down or abandoning their print runs in favour of online editions is testament to that (Cohen, 2008) – but it is also increasingly important as an expression of contemporary culture. The advent of blog service providers and social networking sites has lowered the barriers for people wishing to express themselves on the Web, meaning even those with few technical skills and limited Internet access can publish their thoughts, ideas and opinions.

The value of preserving snapshots of the Web for future reference and study was quickly recognised, with the Internet Archive and the National Library of Sweden both starting their large-scale harvests of Web sites in 1996 (Gray, 2001; Masanès, 2009). Since that time, Web archiving – the selection, collection, storage, retrieval, and maintenance of the integrity of Web resources – has become more widespread, assisted by ever more advanced tools, but perfecting the process is something of a moving target, both in terms of the quantities involved and the sophistication and complexity of the subject material. This is without factoring in the growing demands of the research questions for which the archived material might be expected to act as evidence.

This report provides a snapshot of the state of the art of Web archiving, noting areas of current research and development. It should be of interest to individuals and organisations concerned about the longevity of the Web resources to which they contribute or refer, and who wish to consider the issues and options in a broad context. The report begins by reviewing in more detail the motivations that lie behind Web archiving, both from an organisational and a research perspective. The most common challenges faced by Web archivists are discussed in section 3. The following two sections examine Web archiving at extremes of scale, with section 4 dealing with full-domain harvesting and the building of large-scale collections, and section 5 dealing with the ad hoc archiving of individual resources and small-scale collections. The challenges associated with particular types of difficult content are summarised in section 6, while methods for integrating archived material with the live Web are reviewed in section 7. Finally, some conclusions are drawn in section 8.

2 Motivations for Web archiving

The original motivation for the Web was to provide a constantly evolving information resource (Berners-Lee, 1989). It was designed to fulfil a need that could not be served either by asking lots of people directly or by consulting books: a need for quick and simple access to up-to-date information collected from many contributors. For a medium with such an emphasis on currency and constant revision, it may seem odd at first that people should want to keep out-of-date versions of it. The reality is that, in common with other forms of ephemera, Web resources have a secondary use as evidence about the time they were created or modified. This evidence has an application in everything from marketing to legal proceedings to historical research.

The Preservation of Web Resources (PoWR) Handbook (Pinsent et al., 2008) provides several business cases for a higher or further education institution engaging with Web archiving. They apply equally well to other types of organisation.

- Web sites provide evidence of an organisation's activity. While they may not seem as vital to business continuity as, say, financial records, they may contain valuable evidence for auditing and investigation purposes, and indeed may be vital for compliance with Freedom of Information legislation among other obligations. Such evidence may also be valuable for promotional purposes, for example when providing a historical context to anniversary celebrations.
- Web sites provide evidence of the information published by an organisation. If a site provides advice or guidance, the precise wording and presentation used may be important evidence if that guidance is later called into question.
- While Web sites contain much that is ephemeral, they also contain much that could be reused in future provided it is not lost through deletion in the meantime. They may also contain scholarly, cultural and scientific resources that are or could be cited in traditional publications as well as by other Web sites.

Research commissioned by TNA and carried out by Inforesight, intended for publication in a report entitled *Delivering coordinated UK Web archives to user communities*, found that the main users of Web archives at the time were journalists, litigants, detectives, civil servants, web designers and researchers (Smith, 2009). This provides some insight into the variety of uses to which society in general may put archival Web material.

Even within the sphere of academic research a number of different avenues may be identified. As one example, the World Wide Web of Humanities project collected and analysed over 6 million Web pages from the Internet Archive, to determine how they were interlinked and how they had changed over time (Meyer, 2009).[1] A further

1. World Wide Web of Humanities project Web page, URL: `http://www.oii.ox.ac.uk/research/project.cfm?id=48`

example is the New Media programme of the University of Amsterdam Media Studies Department, which is examining ways in which various Web metrics – patterns of links between contemporaneous sites, search engine results for a particular query – can be used to infer characteristics of society in general (Weltevrede, 2009).[2]

Ashley (2009) gives a number of research questions to which archived Web resources could one day provide the answer, given a sufficient standard of archiving.

- How has the Web changed society (visualisation of Web traffic and how it changed)?
- How did a particular site evolve over time?
- How did the language of a site change over time (did it become more or less formal, when was neologism X first used, how did the balance between concepts X and Y change)?[3]
- Which formats were used, and how did this change over time?
- What would a search for X at time T have returned?
- How were pages linked, and how much traffic flowed along those links?
- What would a mashup of services X and Y have looked like?

In order to answer some of these questions, more than just archived Web pages would be required; some lines of research would also require access to underlying databases, service software/APIs, server logs, and perhaps even DNS lookup tables.

2. See, for example, the Digital Methods Initiative Web site, URL: `http://wiki.digitalmethods.net/`
3. For an example of first steps in this area, see Tahmasebi, Ramesh and Risse (2009).

3 Challenges for Web archiving

In order to create Web archives of sufficient quality to support the activities described in section 2, there are both technical and organisational challenges that need to be addressed by those harvesting and preserving the content.

3.1 Technical challenges

Many of the technical challenges associated with Web archiving involve particular types of Web content, or archiving with a particular scope. Such issues are examined in the relevant sections of this report. In contrast, any Web archiving effort involving resources of some complexity – depending on multiple pages or binary files, each of which may be updated separately – will be affected to a greater or lesser degree by *temporal consistency*.

3.1.1 Temporal consistency

Temporal consistency, otherwise known as temporal coherence or temporal cohesion, is a property of a set of archival Web pages, indicating that there was a point in time at which all the archived pages were live simultaneously on the Web. This is relatively easy to accomplish for a small set of infrequently updated pages, but becomes increasingly difficult to achieve as the number of pages in the set increases, and as the frequency at which pages are updated increases.

Whether a set of archived pages has temporal consistency may be calculated as follows. Consider a set P of pages p_1 to p_N, where each page is harvested once during a crawl. Let $t_{uk}(p_i)$ be the time at which page p_i was updated for the kth time since $t = 0$. Let $t_h(p_i)$ be the time at which page p_i was harvested, adopting the conventions that:

$$t_h(p_1) = 0$$
$$t_h(p_i) \leq t_h(p_{i+1}) \quad \text{for } 1 \leq i \leq N - 1$$

The set of pages P has temporal consistency if and only if there does *not* exist a k and m such that:

$$t_h(p_i) < t_{uk}(p_i) < t_{um}(p_j) < t_h(p_j)$$

for any i and j where $i < j$.

Intuitively, there are several techniques that can be used to improve temporal consistency. One is to harvest pages in order according to the rate at which they are updated. In

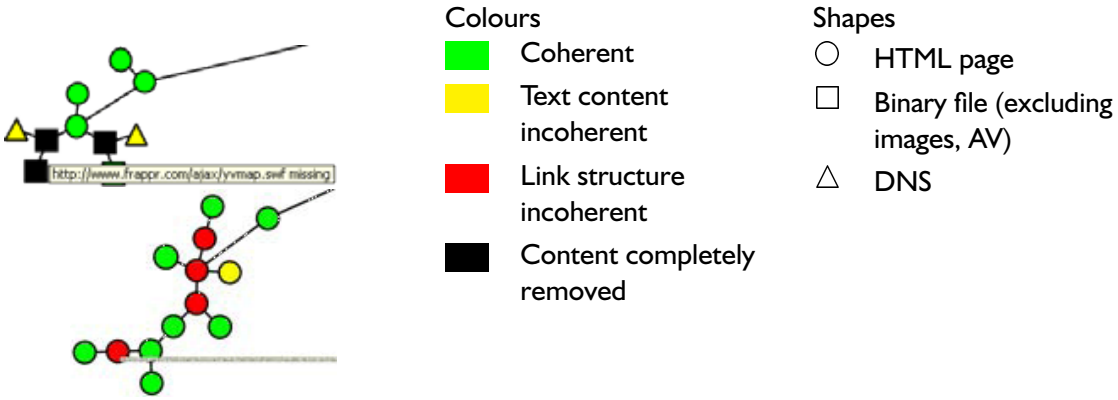

Figure 1: Extract from a coherence defect visualisation (Spaniol et al., 2009, p. 32).

this way, from the perspective of the page with the shortest lifespan, the gap between it being harvested and another page being harvested is shortest for the page with the next shortest lifespan, and longest for the page with the longest lifespan. Another technique is to take advantage of patterns in the way pages are updated. For example, if several pages with approximately the same lifespan are always updated in a certain order (within a shorter period of time than that lifespan), then harvesting them in that same order should ensure a temporally consistent set is harvested. Other techniques include running several harvesters in parallel to reduce the time period in which a snapshot is taken – although bombarding a server with many requests over a short period of time is generally considered bad practice – and (in continuous rather than snapshot harvesting) taking snapshots more frequently of those pages that are updated more frequently.

The LiWA Project has developed a tool for visualizing the temporal consistency or otherwise of a set of harvested Web pages resulting from at least two crawls (Spaniol, Mazeika, Denev & Weikum, 2009). The tool infers the lifetime of Web pages (i.e. the period during which they were live on the Web) in an archived collection from an estimate of when each page was first published to the Web. This is inferred from the HTTP Last-Modified header, or failing that from timestamps present in the page content, or failing that from content comparison with other archived versions of the page. While it is possible to load data into the tool from ARC and WARC files, the tool works best when integrated directly into Heritrix; this is because it makes gathering certain data less computationally expensive, and it allows checks for changes (recrawls) to be made directly after a crawl has taken place: again, this is less computationally expensive than performing another full crawl.

The data from the crawl is used to generate a spanning tree, in which each node in the tree (i.e. Web resource) is flagged according to whether the second copy is the unchanged from the first, differing in textual content only, differing in link structure as well, or missing entirely. The tree is then simplified so that every fully coherent (unchanged) subtree is replaced by a single, proportionately larger node. The results are then converted to GraphML format (Brandes, Eiglsperger & Lerner, n.d.), which can then be converted to a visual representation using the visone tool.[4] For example,

4. Visone software Web site, URL: `http://visone.info/`

Figure 1 shows an extract from a coherence defect visualisation generated from a crawl and subsequent recrawl of the `mpi-inf.mpg.de` site.

3.2 Management challenges

Quite apart from the technical challenges of performing Web archiving, there are perhaps more fundamental issues to address from a management perspective: namely, who should perform the Web archiving, and according to what policies and strategies should the Web archiving take place (Jordison, 2009; Pinsent et al., 2008).

3.2.1 Web archiving responsibilities

The question of who should perform Web archiving does not admit of a straightforward answer; as with other forms of archiving, there are clear benefits both to centralising the activity on a small group of experts, and to distributing the task among many generalists. The question is somewhat easier to answer if narrowed to a particular part of the Web, or a particular set of motivations.

The archiving of entire top-level domains (such as `.uk`, `.eu` or `.com`) is a large-scale operation that suits a well-resourced and stable organisation with a natural national or international scope. As such, this task has largely fallen to national libraries. Such large-scale Web archiving is discussed in more detail in section 4.

Given the size of a typical top-level domain, it is not practical for the staff performing the archiving to ensure that all sites are fully harvested and that all important modifications to those sites are captured. There is therefore a place for more targeted forms of Web archiving, where special attention can be given to ensuring harvests are successful and reflect the relative importance and inertia of individual sites. Special collections relating to specific subjects would naturally fall to libraries, archives, museums and galleries that specialise in those subjects. It is debatable whether each such institution should aim to keep its own archive, or to build a collection through the resources and infrastructure of a larger, more generalist institution; the UK Web Archive, for example, is operated principally by the British Library, but has a collection on women's issues built in collaboration with the Women's Library at London Metropolitan University, and a Quakers collection built in collaboration with the Society of Friends Library.

At a finer level of detail, individual organisations have a responsibility to make sure their own Web presence is archived. This can either be achieved through co-operation with an external Web archiving effort, or through some internal archiving process.

Finally, if individuals rely on particular Web resources, perhaps as citations in documents or as a record of their own intellectual output, they ought to consider ensuring those resources are archived somewhere, whether that is privately within the individual's own digital collections, in an institutional Web archive, or a general Web archive. Such small-scale Web archiving is discussed in more detail in section 5.

3.2.2 Choice of preservation strategies

When choosing a preservation strategy for Web-based content, one of the most important factors to consider is the priority that should be assigned to the different properties of the page. In other words, what would the users of the archive consider the most *significant properties* of the Web content?

In considering this question, it is useful to bear in mind the performance model of the National Archives of Australia (Heslop, Davis & Wilson, 2002). In this model, a researcher does not experience a digital record directly, but instead experiences a performance arising as a result of a process (software, hardware) running on some source data. In the Web context, a researcher looking at a Web page is in fact looking at a rendering of a collection of source files performed by a browser.

The significant properties of the Web page are those aspects of the performance that should remain the same over time. These aspects could include some or all of the following:

- *Text.* For most purposes it is likely that the textual content of a Web page will be important. Counter-examples include pages used as a wrapper for embedded content, and pages with dummy text used to showcase a design.

  For the most part the text will be present and marked up in the source of an (X)HTML page, but may also have been provided in embedded content (images, Flash animations) or generated on the client side by JavaScript. While it is possible to present plain text in (X)HTML pages using the '`pre`' element, it is normally formatted either with presentationally precise markup (e.g. bold, italic, underline) or with more semantic markup that implies but does not demand a particular form of presentation (e.g. heading, paragraph, emphasis). (X)HTML also provides mechanisms for various types of annotations, such as title text, alternative text and links – taken on a textual level, a link may be viewed as annotating a portion of text with the Web address of a related piece of information. A less obvious form of annotation comes from giving portions of web content meaningful `class` attributes drawn from a *microformat* specification; automated tools are usually alerted to this usage by the presence of a link to the microformat's profile (Berriman et al., n.d.; Halpin & Davis, 2007). It is therefore possible to consider text at various levels of richness.
- *Appearance.* While (X)HTML provides some scope for specifying how a Web page is presented beyond textual formatting – for example, background colours, the widths of certain elements – the most powerful method for doing so in version 4 and later is using Cascading Style Sheets (CSS). CSS rules may be applied to (X)HTML directly in `style` attributes, or indirectly through selectors that operate on element names, `class` attributes and `id` attributes. The CSS language itself is in an odd position at the time of writing, as the only officially adopted version is version 1, which is deprecated in favour of the Candidate Recommendation, version 2.1; version 3 is also in development. There is no mechanism in either (X)HTML or CSS to declare the version of CSS in use.

  Support for CSS varies considerably between different browsers and different browser versions. In the worst case scenario, this allows a Web designer to

write CSS code that only renders as intended in one version of one browser. It is possible, though, for Web designers to write CSS code aimed at fully compliant browsers and to exploit bugs or extensions in the (X)HTML/CSS handling of less compliant browsers to feed them patched CSS code (Gallant, 2009). With small-scale archiving, but not with large-scale archiving, it may be possible to distinguish these two cases and treat them accordingly.

- *Interactivity.* Within (X)HTML itself interactivity is largely restricted to links and forms, both of which provide a user-friendly way of requesting more content from a Web server. A limited number of effects may be achieved with CSS, while more complicated interactivity is possible through the use of JavaScript or by embedding interactive content encoded in another format (e.g. Flash, SVG).

  From an archival perspective, there are at least three dimensions of interactivity to consider. The first is the where the interaction takes place. Interaction that takes place entirely on the client-side – for example, using downloaded JavaScript files or Java applets – can be handled in much the same way as more static archival content. Interaction that depends on communication with a server (or several) is rather more difficult to achieve without providing an emulation of the server(s) in question. The second is the distribution of the resources providing the interactivity. If the resources hail from different servers, they and the Web page are much less likely to be temporally consistent than if they all come from one server, and it increases the magnitude of work needed to emulate the server behaviour. A third is the level of interactivity. Conceptually the simplest case is where a script operates on user input to generate an output, as this does not involve additional data. The next is where the interaction prompts a request for more data from a server; examples include buttons that switch a page's style sheet, or a search and retrieval interface. The most complex case involves operations that alter data on the server, such as facilities for adding comments to blog posts. In normal circumstances the alteration of archival data would be undesirable, but an emulated server might be set up to use a fresh copy of the archived data each time on starting.

- *Dynamism.* A dynamic Web page is one in which the content is generated on demand. It is possible to achieve some dynamism on the client side using JavaScript, but it is more common for content to be generated on the server side, often using internal databases or data streams from other sources (e.g. RSS feeds, XML or JSON files).

  If dynamism is not considered significant, the content of such pages can be preserved in static snapshots. If dynamism is important, then it might be necessary to take a snapshot of the Web server itself rather than the generated pages, ensuring temporally consistent snapshots of the data dependencies (or the servers that provide them) are also archived.

For a more formal approach to determining the significant properties of a digital object, see the InSPECT Project's *Framework for the Definition of Significant Properties* (Knight, 2008).

The particular mix of properties given priority within a harvesting exercise, alongside the nature of the Web resources themselves, will influence the strategy to be adopted.

If all that matters is the designed appearance of a page, then saving a snapshot of the rendered page as an image might be sufficient. If all that matters is the textual content, then preserving the (X)HTML code of the served page verbatim would suffice for simple pages. Preserving the behaviour of a Web resource usually requires more complex solutions, which can range from rewriting URLs to point to temporally consistent archived Web resources, all the way through to maintaining browsers on emulated platforms. Indeed, there have been moves within projects such as Dioscuri, GRATE and KEEP to provide emulation solutions for Web browsing (van der Hoeven, 2009).

3.2.3 Strategies for content stored by third parties

For archives responsible for preserving the Web presence of an organisation, it is worth noting that it is increasingly common for organisations and their staff to surface content through third-party services such as SlideShare, Flickr and YouTube. Ideally, organisations should harvest such content as part of their Web archiving activity but this may be difficult both technically, especially where streaming media are involved, and from a rights perspective. In such cases, it may be preferable to set up policies and infrastructure for dealing with the master copies – for example, requiring that slides uploaded to SlideShare are also deposited in the institutional repository. More extensive advice on this matter is available in the PoWR Handbook (Pinsent et al., 2008).

3.2.4 Transfer and exit strategies

In common with other types of archive, Web archives need a strategy to follow in case they are no longer able to preserve the archived material. Typically this will involve transferring the contents of the Web archive to another Web archive, so preparations need to be made in advance to ensure that this can happen. For example, if the strategy indicates a particular archive as a suitable recipient for transferred material, that archive should be contacted to enquire about archival formats and minimum levels of metadata that it would be willing to accept, and suitable transfer procedures. The first archive can then adjust its own procedures so that these conditions can be met if necessary. In the same vein, if permissions are likely to be an issue, these need to be sought around the time of harvest, and not at the point at which the transfer would need to be made.

4 Large-scale Web archiving

There is a spectrum of scales at which Web archiving may be performed, ranging from snapshots of individual pages to archiving entire top-level domains. At the large-scale end of the spectrum, infrastructural issues become particularly important: automating the harvesting of content, providing mechanisms for access, sustaining the archive over the long term, and so on. This section introduces the organisations performing large-scale Web archiving currently, the ways in which such activity is scoped, the tools used in the course of archiving, and some of issues that arise when archiving the Web at the scale of entire domains.

4.1 Agents and scope

4.1.1 Large-scale Web archiving in the UK

In the UK, national-scale archiving is performed by the UK Web Archive.[5] When the UK Web Archive was first set up in 2004, it was operated by the UK Web Archiving Consortium (UKWAC), a collaboration between the British Library, the Joint Information Systems Committee (JISC), the National Archives, the National Library of Wales, the National Library of Scotland, and the Wellcome Library. In 2009, it was decided that UKWAC would be re-organised as a strategic group within the Digital Preservation Coalition (DPC), under the new name of the UK Web Archiving Task Force.[6] Currently the Archive is provided by the British Library in partnership with the National Library of Wales, JISC and the Wellcome Library; special collections within the Archive are built with the co-operation of key institutions such as the Live Art Development Agency, the Society of Friends Library and the Women's Library at London Metropolitan University.

4.1.2 International Internet Preservation Consortium

The International Internet Preservation Consortium (IIPC) is a body made up of national Web archiving initiatives.[7] It was founded in 2003 by twelve libraries: the national libraries of Australia, Canada, Denmark, Finland, France, Iceland, Italy, Norway, and Sweden; The British Library; The Library of Congress; and the Internet Archive. The Consortium now has 36 members from across Asia, Europe, North America and Oceania.

5. UK Web Archive Web site, URL: `http://www.webarchive.org.uk/`
6. UK Web Archiving Task Force Web page, URL: `http://www.dpconline.org/about/web-archiving-and-preservation-task-force.html`
7. IIPC Web site, URL: `http://www.netpreserve.org/`

According to its Web site, the goals of the Consortium are

- to enable the collection, preservation and long-term access of a rich body of Internet content from around the world;
- to foster the development and use of common tools, techniques and standards for the creation of international archives;
- to be a strong international advocate for initiatives and legislation that encourage the collection, preservation and access to Internet content; and
- to encourage and support libraries, archives, museums and cultural heritage institutions everywhere to address Internet content collecting and preservation.

The IIPC does not perform Web archiving corporately, but rather provides a forum for Web archiving initiatives to share good practice and harmonise their efforts.

4.1.3 Scope

Large-scale Web archiving is usually performed according to a combination of two different approaches: full-domain harvesting and selective harvesting. *Full domain harvesting* refers to attempts to collect a comprehensive snapshot of a (top-level) domain such as `.uk`. It is usual for such harvests to take place on an annual basis or thereabouts. *Selective harvesting* refers to collecting only sites that fulfil a certain set of criteria. Selection is usually used to build collections within the wider domain scope, harvested *ad hoc* or at frequent (daily, weekly) intervals. It can also be used to extend the scope of an annual harvest beyond just the top-level domain. Criteria can include

- cultural interest (e.g. Web sites of high perceived value in country X, material in language Y, material about country X, material written in country X but published on a site hosted in another country or internationally)
- relevance to a particular subject area
- relevance to a particular event (such as an election or major sporting event)
- general historical interest (e.g. news sites)
- frequent updates (usually used in conjunction with another criterion)
- permission granted – the UK Web Archive currently works on this basis.

4.2 Tools employed

There are several different tools and formats available for large-scale Web archiving. Listed below are the tools and formats in most widespread use at present; this list is based on the one compiled by Anderson et al. (2010).

4.2.1 Acquisition

Heritrix An open-source, extensible, Web-scale, archiving quality Web crawler.
Developed by the Internet Archive with the Nordic National Libraries.
Current versions: 1.14.3 (3 Mar 2009) and 3.0.0 (5 Dec 2009).
More information: `http://crawler.archive.org/`
Download: `http://sourceforge.net/projects/archive-crawler/`

DeepArc A portable graphical editor which allow users to map a relational data model to an XML Schema and export the database content into an XML document.
Developed by the National Library of France.
Current version: 1.0rc1 (18 Jan 2005).
More information: `http://deeparc.sourceforge.net/`
Download: `http://sourceforge.net/projects/deeparc/`

PageVault A commercial server-side utility for archiving every unique response body sent by the server to its clients.
Developed by Project Computing.
Current version 1.10 (2 Dec 2002).
More information: `http://www.projectcomputing.com/products/pageVault/`

HTTrack A free (GPL) and easy-to-use offline browser utility.
Developed by Xavier Roche and others.
Current version: 3.43-9 (4 Jan 2010).
More information: `http://www.httrack.com/`
Download: `http://www.httrack.com/page/2/en/`

WGet A free command-line (or scripted) tool for downloading content via HTTP, HTTPS and FTP.
Developed by the GNU Project.
Current version: 1.12 (22 Sep 2009).
More information: `http://www.gnu.org/software/wget/`
Download: `http://wget.addictivecode.org/Faq#download`

4.2.2 Curator Tools

Web Curator Tool (WCT) A tool for managing the selective Web harvesting process. It is designed for use in libraries and other collecting organisations, and supports collection by non-technical users while still allowing complete control of the Web harvesting process. The WCT is now available under the terms of the Apache Public License.
Developed by the National Library of New Zealand and the British Library and initiated by the International Internet Preservation Consortium.
Current version: 1.5 (24 Nov 2009).
More information: `http://webcurator.sourceforge.net/`
Download: `http://sourceforge.net/projects/webcurator/`

NetarchiveSuite A curator tool allowing librarians to define and control harvests of web material. The system scales from small selective harvests to harvests of entire national domains. The system is fully distributable on any number of machines

and includes a secure storage module handling multiple copies of the harvested material as well as a quality assurance tool automating the quality assurance process.
Developed by the Royal Library and the State and University Library in the virtual organisation netarchive.dk.
Current version: 3.11.0 (22 Dec 2009).
More information: http://netarchive.dk/suite
Download: http://netarchive.dk/suite/Get_NetarchiveSuite

PANDAS (PANDORA Digital Archiving System) A web-based management system to facilitate the processes involved in the archiving and preservation of online publications. It was designed for the PANDORA Archive but is available to other libraries on a cost recovery basis.
Developed by the National Library of Australia.
Current version: 3.0 (27 June 2007).
More information: http://pandora.nla.gov.au/pandas.html

4.2.3 Collection storage and maintenance

BAT (BnFArcTools) An API for processing ARC, DAT or CDX files.
Developed by the National Library of France.
Current version: 0.07 (3 Feb 2005).
More information and download: http://bibnum.bnf.fr/downloads/bat/

4.2.4 Access and finding aids

Wayback A tool that allows users to see archived versions of web pages across time.
Developed by the Internet Archive.
Current version: 1.4.2 (17 Jul 2009).
More information: http://archive-access.sourceforge.net/projects/wayback/
Download: http://sourceforge.net/projects/archive-access/

NutchWAX (Nutch with Web Archive eXtensions) A tool for indexing and searching Web archives using the Nutch search engine and extensions for searching Web archives.
Developed by the Internet Archive and the Nordic National Libraries.
Current version: 0.12.9 (13 Jan 2010).
More information: http://archive-access.sourceforge.net/projects/nutchwax/ Download: http://sourceforge.net/projects/archive-access/

WERA (WEb aRchive Access) A Web archive search and navigation application. WERA was built from the NWA Toolset, gives an Internet Archive Wayback Machine-like access to Web archives and allows full-text search.
Developed by the Internet Archive and the National Library of Norway.
Current version: 0.4.1 (17 Jan 2006).
More information: http://archive-access.sourceforge.net/projects/

wera/
Download: http://sourceforge.net/projects/archive-access/

Xinq (XML INQuire) A search and browse tool for accessing an XML database.
Developed by the National Library of Australia.
Current version: 0.5 (26 July 2005).
More information: http://www.nla.gov.au/xinq/
Download: http://sourceforge.net/projects/xinq/

4.2.5 Web archiving formats

ARC Developed by Mike Burner and Brewster Kahle, Internet Archive.
Current version: 1.0 (15 Sep 1996).
More information: http://www.archive.org/web/researcher/ArcFileFormat.php

DAT Contains meta-data about the documents stored in ARC files.
More information: http://www.archive.org/web/researcher/dat_file_format.php

CDX Individual lines of text, each of which summarises a single web document.
More information: http://www.archive.org/web/researcher/cdx_file_format.php

WARC Published as ISO 28500:2009, *Information and documentation – WARC file format*.
Developed by IIPC and ISO TC46/SC4/WG12 from ARC.
Current version: 1.0 (May 2009).
More information: http://bibnum.bnf.fr/WARC/

4.3 Legal and social issues

Web archiving raises several important issues for content creators:

- If a particular snapshot is widely cited instead of the original resource, there is a danger that the archived snapshot might rank higher in search engine results, eclipsing the live version of the resource. This in turn would reduce any page-view-related revenue for the author.
- If a content creator is obliged to edit the content of a live resource, or remove it entirely (e.g. because it contravenes another's rights; because it is libellous; because it is potentially harmful, incorrect medical information; because it is personally or professionally embarrassing) the problematic content would still remain untouched in the archive.
- Content that has been archived cannot then be subsequently withdrawn from free access and offered through toll access, or at least, such a move would be less successful than without the archived copy.

Copyright legislation is a point of concern for Web archiving initiatives, as it is in other areas of digital curation. In many countries, copyright legislation is becoming more restrictive in response to lobbying from media companies, who are themselves responding to a perceived threat to their income from file sharing. There has not been a co-ordinated response from the cultural heritage community to secure exemptions for archival purposes (and other instances of fair use/dealing) but there have been responses in individual countries. For example, in Finland lobbying from the National Library led to amendments in copyright legislation to allow the National Library's Web Archive to remove copy protection mechanisms from archived material (Hakala, 2009). In the US, the Section 108 Study Group – a joint venture of the National Digital Information Infrastructure and Preservation programme and the Copyright Office – is advising Congress on how the Copyright Act may be amended to be more sympathetic to 'memory institutions'.

In some countries, legal deposit legislation is (or will be) sufficient to void concerns about copyright with regard to archiving copies of Web materials. It does not, however, provide a sufficient mandate to overcome any concerns about how copyright legislation or ethical considerations (see the issues for content creators above) impact on the matter of providing access to the archived material. National Libraries are currently acting in a highly risk-averse manner with regard to the access they provide to their Web archives. The access policies have the following dimensions.

- *Time.* Access may be embargoed, as currently happens with census data.
- *Geographical.* Access may be limited to certain terminals. It is common for access to be restricted to national library reading rooms. In Austria, access is also available at certain research libraries. In France, the part of the archive collected by Inatheque de France (the national broadcast media archive) is available in Inatheque reading rooms as well as the BNF.
- *Personal status.* Access may be limited to academic researchers. This is the case in Denmark.

On the other hand, the Internet Archive has allowed indiscriminate Web access to its archive since the launch of the Wayback Machine in 2001 (Gray, 2001) and despite being a more fragile organisation than the national libraries, has never been sued or even seen as a threat by content providers. Neither has it been cited as a cause of lost advertising revenue. There are many possible reasons for this, but one may speculate that at root, it is because the Internet Archive does not compete with the live Web: it serves a quite separate set of needs.

One of the consequences of the stricter access policies of national Web archives is that they tend to restrict the forms of research that can be conducted on the material. The pages can only be accessed as single pages, and therefore it is not possible to perform large-scale data mining over the whole collection. It also means the archive cannot be used in the ways envisaged by the UK Government Web Continuity Project or the Memento system.

One possible solution would be to equip the archive with machine-readable statements of access rights for archived Web resources, the idea being that individual files could

be blocked from being served from the archive, possibly with different permissions for different classes of user. Hiiragi, Sakaguchi and Sugimoto (2009) take this one step further, and propose a system for redacting archived Web resources. The system they propose allow passages of text to be replaced with obscure content, optionally accompanied by metadata that explains the reason for the deletion. The advantage of this is that it means a whole resource does not have to be blocked on account of a small section of sensitive content.

4.4 Abuse of Web pages

As with any mode of communication, the Web is prone to abuse. From the perspective of Web archiving, the most problematic forms of abuse involve the spreading of malware (viruses, trojans, spyware), the propagation of illegal copies of digital resources, and certain forms of Web spam – the latter often being used to aid the first two. Web spam is a collective term for a number of techniques used to manipulate search engine rankings, either by filling a page with meaningless text designed to look relevant to a particular set of searches, or by generating a large number of incoming links to a target page. Some techniques for automatically generating pages can end up creating an infinite series of links known as a crawler trap; these can end up crashing a crawler, or else wasting a lot of resources.

While it is clear that such phenomena may be academically interesting in their own right, there are several issues that a Web archive must consider before including it – the threat posed by malware to both the archive and its visitors being particularly pertinent. While not necessarily dangerous in its own right, there is considerable cost associated with Web spam, in terms of both the time that could have been spent archiving more worthwhile content, and the amount of space it takes up in an archive. In the 2004 domain crawl for .de, for example, only 70% of HTML pages were reputable; nearly 20% of HTML pages (over 10% of sites) were spam (Risse, 2009).

The options open to an archive in the face of these considerations are (*a*) to keep all the problematic content, except perhaps to transfer malware and illegal content to offline storage and replace it with sanitised or null content, (*b*) to keep the Web spam but eliminate malware and illegal content entirely, (*c*) to discard most problematic content, keeping a small selection as a representative sample, or (*d*) to discard all problematic content. The option chosen will depend on the target user group for the archive, influences from stakeholders and the archive's technical capabilities.

The Living Web Archives (LiWA) Project has developed a tool for filtering spam from a set of archived Web pages.[8] The tool uses an approach similar to that used by e-mail clients: Bayesian filters with good training sets, updated regularly. In order to make it easier for archivists to maintain the filters, the project has prepared a Java-based user interface that simplifies the task of generating training sets (Benczúr et al., 2008).

8. LiWA Project Web site, URL: `http://www.liwa-project.eu/`

5 Small-scale Web archiving

Large-scale Web archiving is suited to gathering a comprehensive picture of the state of a Web domain at a given point in time, but due to the amount of data involved and the nature of the harvesting process it is easy for Web resources to be missed. If individuals need reliable access to particular Web resources, this requires a more targeted approach, described here as small-scale Web archiving.

One of the attractions of small-scale Web archiving is that it is distributed and economical. The costs of desktop Web archiving can easily be absorbed as part of regular research time and IT support costs. Furthermore, the archived resources are archived precisely because someone has found them interesting or useful. That a resource may be archived multiple times by different people is not necessarily a problem, as it provides a level of back-up proportional to the interest in the resource.

Another attraction lies in the particular use case of a researcher citing a Web resource in a publication. If a resource is removed from the Web, it is clear to readers that what they are seeing is not what the author saw. If they are particularly interested, they might use a large-scale archive to track down a copy of the page, which may or may not be the version read by the author. If a resource is merely modified, it is not necessarily obvious to readers that this has happened, and it is unlikely they will consider it worth the effort to track down the earlier version even if the new version gives a different impression to that imparted to the author. With small-scale Web archiving, it is possible to provide readers with the cited resources alongside the knowledge that they are seeing the version read by the author.

5.1 Forms of small-scale Web archiving

There are at least three forms of small-scale Web archiving that may be distinguished. Each has a different set of use cases, and is served by different tool set.

5.1.1 Cloud-based Web archiving

Cloud-based Web archiving is where an author stores a snapshot of a Web resource using a third-party online service provider. The provider generates a URL for the snapshot that the author can then use when citing the resource. The principal advantage of this over desktop Web archiving is that the archived copy is visible to anyone, meaning that the author's readers can see immediately what the author saw. Of course, this archiving method only works as long as the online service provider is willing and able to provide the service.

WebCite is the prime example of cloud-based Web archiving, as it is explicitly an archive-on-demand service rather than a domain crawler.[9] Other sites offering similar capabilities include BackupURL and the social bookmarking service Spurl.net.[10]

5.1.2 Citation repositories

A citation repository is a repository that collects the digital materials cited in publications written by its producer users.[11] Instead of providing access information for electronic resources cited by a document in the bibliography, the author instead recreates the bibliography as a page within the citation repository, and provides a pointer to this page within the document. As part of the process of creating the bibliography page, the electronic resources cited within it are ingested into the repository, so that anyone consulting the referenced resources will see the versions the author used while preparing the document in question.

Citation repositories do not wholly solve the problem, of course, as the URL of the bibliography page within the repository becomes a single point of failure for all the references within a document. Arguably it does make the problem more manageable, though, as it involves the repository manager(s) ensuring that the repository works, and keeping the bibliography page URLs persistent across software changes. This is clearly a more scalable solution than, say, having authors provide PURLs for each Web resource they cite, and monitor each one of them to ensure they remain up-to-date (Lecher, 2009).

5.1.3 Desktop Web archiving

Desktop Web archiving is where individuals save local copies of Web resources that are important or interesting to them. This may be achieved in several different ways. At the most lossy level, screen capture tools such as SnagIt allow the user to take a screenshot of an entire Web page. This has the advantage that the page is preserved exactly as the user saw it, but the disadvantage that the functionality of the page is lost (copy and paste, hyperlinks, etc.).[12] All browsers have some degree of functionality for saving copies of Web pages locally, with some able to follow links and save a snapshot of a site (up to a given 'depth' from the top level page). Typically the saved version is an HTML file with an accompanying folder containing the dependencies (images, stylesheets and so on), but some browsers support single file 'archives'.[13] Finally there are tools such as Zotero for saving snapshots of pages along with a certain amount of metadata.[14]

9. WebCite Web site, URL: `http://www.webcitation.org/`
10. BackupURL Web site, URL: `http://www.backupurl.com/`; Spurl.net Web site, URL: `http://spurl.net/`
11. For an example of a citation repository, see the DACHS (Digital Archive for Chinese Studies) citation repository, URL: `http://leiden.dachs-archive.org/citrep`
12. SnagIt Web page, URL: `http://www.techsmith.com/screen-capture.asp`
13. Examples include MIME HTML, which re-uses the syntax for bundling several resources in a single e-mail message (Palme, Hopmann & Shelness, 1999), KDE WAR, a tarball containing both the HTML file and its dependencies, Apple WebArchive format, and an HTML file using data URIs.
14. Zotero Web site, URL: `http://www.zotero.org/`

Where a desktop Web archiving effort has come too late to harvest live Web content, it may still be possible to reconstruct the content from material picked up from large-scale Web archiving operations. Warrick is a utility designed for such an eventuality; it searches the Internet Archive and the caches of Google, Bing and Yahoo for missing content, although support for other archives is planned.[15] Old Dominion University have an installation of it accessible from the Web, although it has limited capacity and therefore a backlog of requests. It is also freely available for download and local installation (McCown, Smith, Nelson & Bollen, 2006).

5.2 Archiving complex resources

Value may be added to material in small-scale archives and in special collections within large-scale archives by explicitly documenting the relationships between the archived resources. Such relationships may include several pages making up a single document, multiple representations of the same content, multiple documents making up a collection (e.g. a journal issue), or diverse resources with precisely the same subject. There are several ways in which such relationships could be documented, but one of the most promising is to use an extension of the Open Archives Initiative's Object Reuse and Exchange (OAI-ORE) specification.

OAI-ORE (Lagoze et al., 2008) provides a way of specifying aggregations of resources, especially Web resources; while it does not have a built-in mechanism for specifying the relationships between these resources, this functionality may be added through standard XML extension techniques. Aggregations are described in a Resource Map (ReM) with its own URI.

The ReMember Framework takes the concept behind Web site recovery tools like Warrick a step further.[16] It uses wiki technology to host (version-controlled) ReMs. When an aggregation is selected by a user, the tool looks up the resources in the ReM and returns screenshots (if possible) of all the ones it can find. If there are any it cannot find, it alerts the user and presents them with a suggested search query for finding an alternative copy. If the user finds a new copy, they submit the link and ReMember updates the ReM. At the same time, if the resource does not already exist in the Internet Archive or WebCite, ReMember adds a copy to WebCite in case the resource goes missing again.

Maintaining ReMs in this way is labour-intensive, but relatively simple. Given a suitable hook for engaging people's interest, it may be possible to crowdsource the task (McCown, 2009).

15. Warrick Web site, URL: http://warrick.cs.odu.edu/
16. ReMember Framework Web site, URL: http://african.lanl.gov/preserve/

5.3 Management issues

Insofar as small-scale Web archiving involves taking snapshots of Web resources without the content provider's permission, it faces the same legal and social issues as large-scale Web archiving. In practice, the risks are reduced as desktop archives of Web resources are invisible to the live Web, and live Web archives are typically much smaller than full-domain archives while still being difficult or impossible to search through.

The archives produced by small-scale Web archiving exist on a spectrum of distribution. Those at the highly distributed end of the spectrum (individual hard drives of Web users) have the advantage that the failure of one archive has little effect on the totality of archived material remaining. The main problem is that these archives act as silos: with no way of co-ordinating them, there is no way for their contents to be aggregated, and sharing can only be initiated through widely broadcast speculative requests. Sustaining these archives is a matter of general good practice: file naming conventions, regular backups, at least minimal metadata, and so on.

At the more centralised end of the spectrum, the greater accessibility of the archived material comes at the price of reliance on a single service provider. Such service providers need to ensure they have a sustainability model in place, and some form of exit strategy in place in case continued operation becomes uneconomic.

6 Archiving difficult content

6.1 Hidden links

Web crawlers work by following the links present in Web documents. When these links are presented in a form other than HTML (e.g. in JavaScript that manipulates the HTML page, or in Flash presentations), they become invisible to most Web crawlers, causing the latter to miss content. In order to overcome this issue, some additional intelligence on the part of the crawler is needed. The simplest method for extracting links from JavaScript is to scan the code for fragments that resemble links (using regular expression pattern matching). This not only has the potential to generate many false positives, but also will fail to pick up truly dynamic links, that is, those generated using the values of variables. The Living Web Archives (LiWA) Project has developed an alternative method that involves running the code through a JavaScript engine running on top of a minimal WebKit-based browser, and extracting the links from the DOM tree generated as a result (Risse, 2009).[17]

6.2 Blogs

In 1997, a team from the University of North Carolina, Chapel Hill, conducted a Web-based survey of 223 bloggers on their perspectives on blog preservation (Hank, 2009a, 2009b; Sheble, Choemprayong & Hank, 2007). The survey asked about bloggers' editing behaviours, how they would cope with losing their blog data, how much they would invest to preserve their blog, and more general questions on how blogs should be selected for preservation, and who should preserve them.

The results of the survey show an enthusiasm for blog preservation, but mixed opinions on what should be preserved, and by whom. Around 71% of those surveyed thought their own blog should be preserved, but only 36% thought all blogs should be preserved. On the question of who should preserve blogs, 76% thought the author should have primary responsibility, while 20% thought the author should only have a secondary responsibility; correspondingly, 26% thought libraries and archives should have primary responsibility, and 45% thought they should have secondary responsibility.

The survey also revealed aspects of blogging behaviour that could prove challenging for preservation. Most of the respondents (96%) edit entries after posting, and 39% have deleted entries; 23% have deleted their entire blog. Some 19% use a password protection mechanism to restrict access to some of their entries, while 2% restrict access to an entire blog in this way. Most of those surveyed (86%) have blogs hosted by

17. LiWA Project Web site, URL: `http://www.liwa-project.eu/`

a service provider, but 16% host their own blog. Most divisive was the issue of paying for preservation: 54% of the sample indicated they would not pay for their own blog(s) to be preserved.

The impression given by these results is that there would be real benefits from authors taking responsibility for archiving their own blogs, namely that the author has full access to all entries, and would have control over which version of a post is archived. In order for this to happen, though, there would need to be author-level tools available: tools that work with existing blog service providers, are inexpensive (preferably free), simple and quick to use.

The ArchivePress Project is pursuing this idea of a simple archiving tool for blogs.[18] It received some inspiration from a comment of Rusbridge (Baker, 2009, comment dated 31 March 2009 at 12:34), who suggested that for blogs, 'content is primary and design secondary,' on the basis that many people read blogs through feed readers rather than Web browsers, and therefore never see the styling of the blog site. The project is thus developing the idea of using blog software to harvest and store the content of other blogs for preservation purposes, using their newsfeeds.

In order to prove this concept, the Project is attempting to archive a set of blogs from the DCC, the University of Lincoln and UKOLN using a customised WordPress installation. Two main issues are being explored. The first is how much of the original information can be saved in this fashion, information such as the author, the time at which the entry was first posted, the times at which an entry was modified, the comments attached to the entry and their authors and timestamps, and so on. The second is the extent to which it is possible to enrich an archived entry with, say, an extensive Dublin Core metadata record (Davis, 2009; Pennock & Davis, 2009).

It is possible that other sites and services that surface their content through RSS feeds (e.g. Twitter) may also be archived using the same technique.

6.3 Institutional Web resources

Higher and further education institutions produce a wealth of Web resources in the course of both corporate activity and the work of individual faculties, departments, teams and researchers. The JISC PoWR (Preservation of Web Resources) Project was given the task of producing guidance on digital preservation for these institutions, covering both technical and organisational issues.[19] The major output from the project was the PoWR Handbook (Pinsent et al., 2008), which deals with the following issues:

- Scoping the Web archiving task: determining the institution's Web resources, understanding possible risks and challenges, appraising which resources to preserve, deciding which significant properties to preserve (textual content, look and feel, change history, etc.).
- Capturing Web resources: approaches, techniques, tools.

18. ArchivePress Project Web site and blog, URL: http://archivepress.ulcc.ac.uk/
19. JISC PoWR Project Web site and blog, URL: http://jiscpowr.jiscinvolve.org/

- Managing captured resources: writing retention policies, performing case-by-case retention evaluations
- Preserving special content: resources in Content Management Systems, resources in third party services, collaborative and modular content.
- Organisational matters: the business case for Web archiving, assigning responsibilities, strategic direction.
- Creating preservation-friendly resources.
- Third-party Web preservation services.

On the matter of Content Management Systems, the Handbook recommends preserving the content as generated Web pages rather than as the underlying databases, on the basis that it is technically easier to view the generated pages later on than to emulate the Content Management System, especially with licences that forbid the use of the original software following an expiry date. Having said that, tools such as DeepArc and Xinq are specifically designed to enable these underlying databases to be archived in an XML format and queried without reference to the original software, though the point from the Handbook still stands: it would be time consuming to attempt to recreate the behaviour of the original CMS exactly.

The Handbook also contains appendices that act as introductions to relevant legal matters (copyright, Freedom of Information, Data Protection) and records management.

6.4 Virtual Worlds

The preservation of online virtual worlds, such as World of Warcraft and Second Life, presents some unique challenges, borne from the combination of computer gaming with Web technologies. There are several projects working in this area, notably the NDIIPP-funded Preserving Virtual Worlds Project and the Documenting Virtual Worlds Project.[20] As one might expect, the techniques being developed stem more from the gaming side of the preservation problem than from the Web side. For example, Antonescu, Guttenbrunner and Rauber (2009) describe one of a number of methods of preserving interactions with a virtual world in the form of a video recording. Similarly, Lowood (2009) describes methods in which log files of interactions may be used to record gaming sessions, and how maps and objects may be migrated from one gaming platform to another.

20. Preserving Virtual Worlds Web site, URL: http://pvw.illinois.edu/pvw/; Documenting Virtual Worlds Web page, URL: http://www.ifs.tuwien.ac.at/dp/second_life/

7 Web archive interfaces

Web archives are typically accessed through a search interface, although some permit browsing. The power of this search facility varies tremendously from archive to archive. The Internet Archive's Wayback Machine, for example, only permits searching by URL, optionally filtered by date. The UK Web Archive permits searching by full text, Web site title and Web site URL, optionally filtered by subject or special collection. The Government of Canada Web archive permits searching by full text, optionally filtered by date, format and URL. It is not currently possible to cross-search the publicly available Web archives, although there is evidence of support for such a service among researchers (Ashley, 2009; Meyer, 2009; Smith, 2009).

Described below are two projects that have developed alternative ways of interfacing with archived Web content. Both of them seek to more closely integrate the archived content with resources live on the Web.

7.1 UK Government Web Continuity Project

The Web Continuity Project is attempting to solve the problem of Central Government Web resources being deleted, causing any links pointing to that information to break.[21] This was brought to Parliamentary attention when Jack Straw wrote a letter to the Cabinet Office Minister about the difficulties of gaining access to government information online. While Government Web sites have been thought of as ephemeral sources of information, they have been cited in Parliament; a review of URLs quoted in Hansard between 1997 and 2006 found that 60% of them were broken, meaning a significant gap in the documentation of government in that period.

The Project – carried out by the (UK) National Archives with assistance from the European Archive – has started creating a comprehensive archive of the Web presence of Central Government: around 1200 sites, to be captured three times per year. In addition, the Project has developed an open source plugin for Apache and MS IIS servers. This plugin changes the behaviour of the server when it cannot find the requested resource. Instead of immediately serving a 404 Not Found error page, the server instead searches the archive for any resources that were harvested from the URL in the request. If it finds one (or several), it redirects the request to the most recent archived snapshot. The archived material is marked with a banner to indicate that it is not a live resource (Spencer & Storrar, 2009).

21. Web Continuity Project Web page, URL: `http://www.nationalarchives.gov.uk/webcontinuity/`

7.2 Memento

Memento is a system developed by LANL and Old Dominion University for providing different snapshots of an online resource through HTTP content negotiation (Van de Sompel et al., 2009).[22] Content negotiation is a mature part of the HTTP standard that allows a single resource (with a single URI) to have several different representations. For example, the same Web page may be available in English, French and German, and in HTML or DocBook XML.

Content negotiation may be performed in one of three different ways. *Server-driven negotiation* is where the user agent (browser) sends the server additional information along with the URI, expressing preferences for certain types of content; the server then uses this information to determine which representation to send back. RFC 2616 (Fielding et al., 1999, section 12) indicates that the following request headers may be used to express a preference among available representations:

- `Accept` – media type (HTML vs XML vs PDF...)
- `Accept-Charset` – character set (ISO-8859-1 vs UTF-8 vs CP 1252...)
- `Accept-Encoding` – compression (none vs gzip vs LZW vs zlib/deflate)
- `Accept-Language` – natural language
- `User-Agent` – allows tailored content for different browsers (e.g. narrow layout for mobile phones/PDAs)

When using this method, it is up to the server to decide whether to serve the representation at the URI requested, or redirect the request to a unique URI. *Agent-driven negotiation* is where the server sends the list of possible representations (each with their own URI) to the user agent; the user agent then decides which URI to request, either automatically or after user interaction. *Transparent negotiation* (Holtman & Mutz, 1998) is where features of server- and agent-driven negotiation are combined. On the server-side, when a request comes in, the server supplies a list of possible representations as in agent-driven negotiations, but may also perform some limited server-driven negotiation in response to a limited set of headers, so that a second request is not always needed. Furthermore, transparent negotiation encompasses cases where intermediate Web caches can act as full proxies for the server by not only returning cached copies of the list of the alternative representations, but also performing server-side negotiation.

Memento is set up to allow transparent negotiation, using a new request header, '`X-Accept-Datetime`'. It works slightly differently, depending on whether the server has archival capabilities (i.e. can provide dated copies of previous versions of pages) or not. In the former case, the server detects the '`X-Accept-Datetime`' header and returns a list of URIs for snapshots of the page in question from a period centred on the time specified. This allows the user agent to select the most appropriate version. In the latter case, the server detects the '`X-Accept-Datetime`' header and redirects the user agent to an archive holding the snapshots, which acts as previously described.

Given that the time dimension is continuous rather than discrete, it has a rather different character to the established content negotiation dimensions, even when discretised

22. Memento project Web site, URL: `http://www.mementoweb.org/`

into one-second intervals. One implication is that it is highly unlikely that a snapshot will exist for exactly the time specified, but highly likely that one from a nearby time would be acceptable. Memento therefore specifies new response headers to help with automated agent-driven negotiation: '`X-Archive-Interval`' provides the dates of the first and last available snapshot of the page, while '`X-Datetime-Validity`' indicates (if known) the time period in which the served page was live on the Web. Another implication is the unlimited number of alternative representations each resource could have. It is because of the scalability issue of returning a full list of alternatives with each request that Memento only recommends providing alternatives from a period centred on the request date. It does, however, provide a mechanism for looking up the full list of alternatives, presented as an OAI-ORE Resource Map.

One further issue is that the essential message of a resource is more likely to change across the time dimension than it is across any of the other dimensions. This has lead to some controversy over whether this is stretching the notion of content negotiation too far (Johnston, 2009; Nelson, Van de Sompel & Sanderson, 2009).

8 Conclusions

In the fourteen years since large-scale Web archiving began, the process has matured considerably. There now exists a suite of tools to aid in the creation of a Web archive, from harvesting all the way through to access. Twenty-five countries from across Asia, Europe, North America and Oceania now have Web archiving programmes. Interest has also started to trickle down to organisations and individuals. When the GeoCities Web hosting service shut down in October 2009, not only did the Internet Archive perform special deep collection crawls of the hosted sites, guided in part by public suggestions, but also no fewer than three volunteer projects were set up to try and save as much of the material as possible ('End of an Era for Early Websites', 2009).[23]

There is, however, plenty of scope for further developments in this area. At the harvest stage, techniques have yet to be developed to ensure the greatest possible level of temporal consistency between Web resources. While there are tools available to help organisations archive their Web presence, they are by no means ubiquitous, and in any case are not yet sophisticated enough to provide accurate emulations of earlier server states. At present there is no widespread solution to the challenge of making restricted content available for dark archives to harvest while preventing desktop Web archiving. At the storage stage, there is arguably too much diversity in the way desktop tools archive Web pages; as standardising on a single archival format is unlikely, there is at least a need for more comprehensive and reliable migration support between the formats in use. At the access stage, there is plenty of work to do to make it easier to find earlier copies of Web resources; there are several promising technical solutions to integrating archived material with the live Web, but the greatest barriers to making national Web archives cross-searchable are social and legal.

While the state of the art of Web archiving is not sufficiently advanced to support all the use cases we can currently imagine for it, it is good enough for some important use cases. Already a vast amount of content has been saved that would otherwise have been lost. It is important that the momentum that has already gathered behind Web archiving is not lost, so that the record of our lives online may be even richer for future generations to discover and study.

23. Internet Archive GeoCities Special Collection 2009, URL: `http://www.archive.org/web/geocities.php`; Archive Team wiki, URL: `http://www.archiveteam.org/`; Geocities Rescue Project Web page, URL: `http://transformativeworks.org/projects/geocities-rescue`; ReoCities Web site, URL: `http://reocities.com/`

Bibliography

Anderson, M., Grotke, A., Jones, G., Berko, M., Boulderstone, R., Hallgrimsson, T. et al. (2010). *Downloads*. Retrieved January 21, 2010, from the International Internet Preservation Consortium Web site: http://netpreserve.org/software/downloads.php

Antonescu, M.-D., Guttenbrunner, M. & Rauber, A. (2009). Documenting a virtual world: A case study in preserving scences from Second Life. In J. Masanès, A. Rauber & M. Spaniol (Eds.), *Proceedings of the 9th International Web Archiving Workshop (IWAW 2009)* (pp. 5–9). Retrieved January 15, 2010, from http://www.iwaw.net/09/IWAW2009.pdf#page=5

Ashley, K. (2009, July 21). *What we want with Web archives: Will we win?* Paper given at the workshop *Missing Links: the Enduring Web*, London. Retrieved January 21, 2010, from http://www.dpconline.org/technology-watch-reports/download-document/391-0907ashleymissinglinks.html

Baker, G. (2009, March 30). *Preservation for scholarly blogs*. Retrieved January 21, 2010, from the *Gavin Baker: a Journal of Insignificant Inquiry* Web site: http://www.gavinbaker.com/2009/03/30/preservation-for-scholarly-blogs/

Benczúr, A. A., Siklósi, D., Szabó, J., Bíró, I., Fekete, Z., Kurucz, M. et al. (2008). *Web spam: A survey with vision for the archivist*. Paper presented at the 8th International Web Archiving Workshop (IWAW 2008). Retrieved May 2, 2010, from http://www.iwaw.net/08/IWAW2008-Benczur.pdf

Berners-Lee, T. (1989, March). *Information management: A proposal*. CERN. Retrieved February 5, 2010, from http://www.w3.org/History/1989/proposal.html

Berriman, F., Cederholm, D., Çelik, T., Khare, R., King, R., Marks, K. et al. (n.d.). *About microformats*. Retrieved January 20, 2010, from http://microformats.org/about

Brandes, U., Eiglsperger, M. & Lerner, J. (Eds.). (n.d.). *GraphML primer*. Retrieved January 15, 2010, from the GraphML Working Group Web site: http://graphml.graphdrawing.org/primer/graphml-primer.html

Cohen, N. (2008, March 16). Start writing the eulogies for print encyclopedias. *New York Times*. Retrieved February 8, 2010, from http://www.nytimes.com/2008/03/16/weekinreview/16ncohen.html

Davis, R. (2009, December 3). *Archiving scientific blogs with ArchivePress*. Poster presented at the 5th International Digital Curation Conference. Retrieved January 21, 2010, from http://www.dcc.ac.uk/events/dcc-2009/programme/posters/ArchivePress_IDCC_Poster.pdf

End of an era for early Websites. (2009, October 26). Retrieved February 8, 2010, from the BBC News Web site: http://news.bbc.co.uk/1/hi/technology/8325749.stm

Fielding, R., Gettys, J., Mogul, J., Frystyk, H., Masinter, L., Leach, P. et al. (1999, June). *Hypertext transfer protocol – http/1.1*. Request for Comments. Retrieved January 21,

2010, from the Internet Engineering Task Force Web site: http://tools.ietf.org/html/rfc2616

Gallant, J. (2009, September 26). *Position is everything*. Retrieved January 20, 2010, from http://www.positioniseverything.net/

Gray, D. F. (2001, October 25). Archiving the net all the 'Wayback' to 1996. *InfoWorld Daily News*.

Hakala, J. (2009, November 13). *Legal aspects of Web harvesting: The Finnish experience*. Paper given at the workshop *Archiving the Web: New Perspectives*, Stockholm, Sweden. Retrieved January 21, 2010, from http://www.kb.se/aktuellt/video/Archiving-the-Web/

Halpin, H. & Davis, I. (Eds.). (2007, June 28). *GRDDL primer*. W3C Working Group Note. Retrieved January 20, 2010, from the World Wide Web Consortium Web site: http://www.w3.org/TR/grddl-primer/

Hank, C. (2009, July 25). *Blogger perspectives on digital preservation: Attributes, behaviors, and preferences*. Paper given at the Future of Today's Legal Scholarship Symposium, Georgetown, Washington DC. Retrieved January 21, 2010, from http://ils.unc.edu/~hcarolyn/FTLS_hank_25July09.pdf

Hank, C. (2009, December 3). *Science and scholarship in the blogosphere: Blog characteristics, blogger behaviours and implications for digital curation*. Poster presented at the 5th International Digital Curation Conference. Retrieved January 21, 2010, from http://www.dcc.ac.uk/events/dcc-2009/programme/posters/poster016hank.ppt

Heslop, H., Davis, S. & Wilson, A. (2002, December). *An approach to the preservation of digital records*. National Archives of Australia. Retrieved January 18, 2010, from http://www.naa.gov.au/Images/An-approach-Green-Paper_tcm2-888.pdf

Hiiragi, W., Sakaguchi, T. & Sugimoto, S. (2009). A policy-based institutional Web archiving system with adjustable exposure of archived resources. In J. Masanès, A. Rauber & M. Spaniol (Eds.), *Proceedings of the 9th International Web Archiving Workshop (IWAW 2009)* (pp. 20–26). Retrieved January 15, 2010, from http://www.iwaw.net/09/IWAW2009.pdf#page=20

Holtman, K. & Mutz, A. (1998, March). *Transparent content negotiation in HTTP*. Request for Comments. Retrieved January 21, 2010, from the Internet Engineering Task Force Web site: http://tools.ietf.org/html/rfc2295

Johnston, P. (2009, November 23). *Memento and negotiating on time*. Retrieved January 21, 2010, from the eFoundations Web site: http://efoundations.typepad.com/efoundations/2009/11/memento-and-negotiating-on-time.html

Jordison, S. (2009, March 25). *Session: Who should preserve the Web?* Retrieved January 21, 2010, from the JISC Events Blog Web site: http://events.jiscinvolve.org/session-who-should-preserve-the-web/

Knight, G. (2008, February 14). *Framework for the definition of significant properties*. JISC. Retrieved January 7, 2010, from http://www.significantproperties.org.uk/documents/wp33-propertiesreport-v1.pdf

Lagoze, C., Van de Sompel, H., Johnston, P., Nelson, M., Sanderson, R. & Warner, S. (Eds.). (2008, October 17). *ORE user guide: Primer*. Version 1.0. Retrieved January 21, 2010, from http://www.openarchives.org/ore/1.0/primer

Lecher, H. (2009, July 21). *Web archive and citation repository in one: DACHS*. Paper given at the workshop *Missing Links: the Enduring Web*, London. Retrieved January 21, 2010, from http://www.dpconline.org/technology-watch-reports/download-document/395-0907lechermissinglinks.html

Lowood, H. (2009, October). *Memento mundi: Are virtual worlds history?* Paper presented at the 6th International Conference on Preservation of Digital Objects, San Francisco, CA. Retrieved February 2, 2010, from http://www.cdlib.org/iPres/presentations/Lowood.pdf

McCown, F. (2009, April 3). *Everyone is a curator: Human-assisted preservation for ORE aggregations*. Paper given at the 2nd Digital Curation Curriculum Symposium, Chapel Hill, NC. Retrieved January 21, 2010, from http://www.ils.unc.edu/digccurr2009/5b-mccown.pdf

Masanès, J. (2009, November 13). *Research challenges in Web archiving*. Paper given at the workshop *Archiving the Web: New Perspectives*, Stockholm, Sweden. Retrieved January 21, 2010, from http://www.kb.se/aktuellt/video/Archiving-the-Web/

McCown, F., Smith, J. A., Nelson, M. L. & Bollen, J. (2006). Lazy preservation. In *Proceedings of the 8th ACM International Workshop on Web Information and Data Management (WIDM 2006)* (pp. 67–74). Retrieved February 9, 2010, from http://www.cs.odu.edu/~fmccown/pubs/lazyp-widm06.pdf

Meyer, E. T. (2009, July 21). *The future of researching the past of the Internet*. Paper given at the workshop *Missing Links: the Enduring Web*, London. Retrieved January 21, 2010, from http://www.dpconline.org/technology-watch-reports/download-document/396-0907meyermissinglinks.html

Nelson, M., Van de Sompel, H. & Sanderson, R. (2009, November 24). *Memento response*. Retrieved January 21, 2010, from http://www.cs.odu.edu/~mln/memento/response-2009-11-24.html

Palme, J., Hopmann, A. & Shelness, N. (1999, March). *MIME encapsulation of aggregate documents, such as HTML (MHTML)*. Request for Comments. Retrieved February 2, 2010, from the Internet Engineering Task Force Web site: http://tools.ietf.org/html/rfc2557

Pennock, M. & Davis, R. (2009, October). *ArchivePress: A really simple solution to archiving Web content*. Paper presented at the 6th International Conference on Preservation of Digital Objects, San Francisco, CA. Retrieved January 21, 2010, from http://archivepress.ulcc.ac.uk/wp-content/uploads/2009/10/pennockm_archivepress_ipres09_1.pdf

Pinsent, E., Davis, R., Ashley, K., Kelly, B., Guy, M. & Hatcher, J. (2008). *PoWR: The Preservation of Web Resources Handbook*. Version 1.0. London: JISC. Retrieved February 10, 2010, from http://www.jisc.ac.uk/publications/programmerelated/2008/powrhandbook.aspx

Risse, T. (2009, July 21). *From Web page storage to Living Web Archives*. Paper given at the workshop *Missing Links: the Enduring Web*, London. Retrieved January 21, 2010, from http://www.dpconline.org/technology-watch-reports/download-document/403-0907rissemissinglinks.html

Sheble, L., Choemprayong, S. & Hank, C. (2007, December 13). *Surveying bloggers' perspectives on digital preservation: Methodological issues*. Paper given at the 3rd

International Digital Curation Conference, Washington, DC. Retrieved January 21, 2010, from http://www.dcc.ac.uk/events/dcc-2007/papers/P19.pdf

Smith, C. (2009, July 21). *Context and content: Delivering coordinated UK Web archives to user communities*. Paper given at the workshop *Missing Links: the Enduring Web*, London. Retrieved January 21, 2010, from http://www.dpconline.org/technology-watch-reports/download-document/398-0907smithmissinglinks.html

Spaniol, M., Mazeika, A., Denev, D. & Weikum, G. (2009). 'Catch me if you can': Visual analysis of coherence defects in Web archiving. In J. Masanès, A. Rauber & M. Spaniol (Eds.), *Proceedings of the 9th International Web Archiving Workshop (IWAW 2009)* (pp. 27–37). Retrieved January 15, 2010, from http://www.iwaw.net/09/IWAW2009.pdf#page=27

Spencer, A. & Storrar, T. (2009, July 21). *Capture and continuity: Broken links and the UK Central Government Web presence*. Paper given at the workshop *Missing Links: the Enduring Web*, London. Retrieved January 21, 2010, from http://www.dpconline.org/technology-watch-reports/download-document/399-0907spencermissinglinks.html

Tahmasebi, N., Ramesh, S. & Risse, T. (2009). First results on detecting term evolutions. In J. Masanès, A. Rauber & M. Spaniol (Eds.), *Proceedings of the 9th International Web Archiving Workshop (IWAW 2009)* (pp. 50–55). Retrieved January 15, 2010, from http://www.iwaw.net/09/IWAW2009.pdf#page=50

Van de Sompel, H., Nelson, M. L., Sanderson, R., Balakireva, L. L., Ainsworth, S. & Shankar, H. (2009, November 6). Memento: Time travel for the Web.

van der Hoeven, J. (2009, July 21). *Emulating access to the Web 1.0: Keep the browser*. Paper given at the workshop *Missing Links: the Enduring Web*, London. Retrieved January 21, 2010, from http://www.dpconline.org/technology-watch-reports/download-document/400-0907vanderhoevenmissinglinks.html

Weltevrede, E. (2009, November 13). *Web archiving: The Dutch experience*. Paper given at the workshop *Archiving the Web: New Perspectives*, Stockholm, Sweden. Retrieved January 21, 2010, from http://www.kb.se/aktuellt/video/Archiving-the-Web/