Archivo de la Internet española: Webs y archivos personales
Madrid, 12 December 2005

Iniciativas de preservación de la Web: una visión actual

Michael Day

Digital Curation Centre
UKOLN, University of Bath, Bath BA2 7AY, United Kingdom
m.day@ukoln.ac.uk
http://www.ukoln.ac.uk/

1. Introduction

Since its origins as a tool to support the management of scientific information in the early 1990s, the World Wide Web has rapidly become a ubiquitous part of the world we live in today. The perceived cultural and scientific importance of the Web means that it has become a highly visible exemplar of a more general concern with the long-term preservation of digital objects. This presentation will introduce some initiatives dealing with collecting and preserving the Web, focusing on the techniques used to collect Web content and highlighting some important issues.

2. The World Wide Web

Reflecting its origins in a information management project at CERN (Gillies & Cailliau, 2000), the Web continues to play a major role in supporting scientific research and other scholarly activities. Initial use focused on its role as a publication medium, for example for the dissemination of information about institutions and projects, but also as an easy means of sharing various types of research output. In many subject disciplines, the Web - as part of what is sometimes known as 'cyberinfrastructure' - has acted as a means of promoting the concept of open access to the outputs of research. The development and deployment of (and political support for) subject and institution-based repositories of research publications is but one example of this, and the same open access principles are increasingly being applied to other kinds of research output. For example, in January 2004, government ministers from all OECD member states (and some others) endorsed a declaration based on the principle that publicly funded research data should be made openly available to the maximum extent possible (Arzberger, *et al*., 2004). James Hendler (2003) has said that scientists have become "increasingly reliant" on the Web for supporting their research activities, noting that the "Web is used for finding preprints and papers in online repositories, for participating in online discussions at sites such as Science Online, for accessing databases through specialized Web interfaces, and even for ordering scientific supplies." It also plays an increasingly important role in supporting e-learning activities in higher education institutions (e.g., Laurillard, 2005).

Of particular concern for scholarship is the longevity of Internet references in the published literature. A study of Internet citations in three major scientific and medical journals (*Science*, *New England Journal of Medicine*, *JAMA*) revealed that the proportion of inactive links rose to 13 per cent 27 months after publication (Dellavalle, *et al*., 2003).

Similar trends have been noted in other biomedical journals (e.g., Hester, *et al*., 2004; Crichlow, Davies & Winbush, 2004; Wren, 2004), in computer science (Spinellis, 2003) and the informetrics sub-discipline of information science (Bar-Ilan & Peritz, 2004). These results support Wallace Koehler's longitudinal studies of the persistence of Web pages, which found that just 33.8 per cent of a sample of pages selected in December 1996 persisted at their original URLs by May 2003 (Koehler, 2004).

The Web is also widely used outside higher education. It has developed very rapidly as a major facilitator of personal communication, commerce, publishing, marketing, and much else across all sectors of society. For example, since its inception, the Web has supported the development of many new types of online commerce (e.g., companies like eBay or Amazon) as well as seen a major move by existing organisations (e.g., the news media, television companies, retailers, etc.) to develop a significant presence online. In addition, private individuals are making increasing use of Web technologies to share information about their personal interests and hobbies. For example, a Pew Internet & American Life Project survey undertaken in 2003 concluded that 44 per cent of US Internet users had contributed some kind of content to the Internet (Lenhart, Horrigan & Fallows, 2004). To summarise, the Web's current importance can be gauged by the comment by Peter Lyman (2002) that it has become "the information source of first resort for millions of readers." It is for this reason that some cultural heritage organisations have begun to consider their role with regard to the collection and preservation of Web content.

2. Why preserve the Web?

There are a variety of reasons why preserving Web content has begun to be addressed by cultural heritage and other organisations.

National libraries have tended to focus on the cultural importance of the Web and its role as a publishing medium. For example, the development of the National Library of Australia's pioneering PANDORA Archive reflected the recognition that the NLA, as a national library, had "a responsibility to develop collections of library materials, regardless of format, which document the history and culture of Australia and the Australian people" (http://pandora.nla.gov.au/overview.html). Following the same principles, many other national libraries have launched operational or pilot Web archiving initiatives. In Europe, these include the national libraries in Austria, the Czech Republic, Denmark, Finland, France, Iceland, Norway, Portugal, Slovenia and the United Kingdom, outside Europe, the national libraries of China, Japan, New Zealand and the United States. An e-mail survey by Hallgrímsson (2003) of European national libraries in 2002-03 showed that 15 out of the 25 libraries that responded had some kind of Web-archiving initiative underway. Some countries, most notably France, have dealt with intellectual property rights problems by including Web archiving amongst the national library's legal deposit responsibilities. Other national library initiatives either seek explicit permission from Web site owners before adding them to the library's collections or severely restrict end-user access.

Outside the national library sector, the main Web archiving initiative is the Internet Archive, a non-profit organisation based in San Franscisco that has been collecting snapshots of the Web since 1996. Historically, the Internet Archive has acquired much Web content from the search company Alexa Internet, which it makes available through its Wayback Machine interface (http://www.archive.org/), but increasingly it co-operates with cultural heritage organisations, providing their technical knowledge and tools to

support the creation of special collections. These include collaborations with the Smithsonian Institution and the Library of Congress (e.g. on collections relating to political elections, 9/11, and Hurricanes Katrina and Rita), and more recently the UK National Archives. A mirror of the Wayback Machine is also available from the Web site of the Bibliotheca Alexandrina in Egypt.

Some national archives have also begun to get involved in the collection and preservation of Web sites, especially where Web sites are understood to have some kind of evidential value. Sometimes this interest manifests itself in the form of guidance for Web managers. For example, the National Archives of Australia (2001a; 2001b) and The National Archives (TNA) in the UK (Public Record Office, 2001) have both issued detailed electronic records management (ERM) guidelines for government Web site managers. TNA has also developed an operational selection policy for Government Web sites (The National Archives, 2003) and initiated, in co-operation with the Internet Archive, a UK Government Web Archive (http://www.nationalarchives.gov.uk/preservation/webarchive/) that has been collecting periodic snapshots of selected Web sites since late 2003. Other national archives have begun to move in the same direction. For example, the US National Archives and Records Administration (NARA) arranged for all federal agencies to take a 'snapshot' of their public Web sites at the end of the Clinton Administration for deposit with their Electronic and Special Media Records Services Division (Bellardo, 2001).

Some universities and scholarly societies have supported smaller Web archiving initiatives. These tend to focus on particular subject domains or events, e.g. political elections. Examples include the Archipol project (http://www.archipol.nl/), dedicated to the collection of Dutch political Web sites, and the Occasio archive of Internet newsgroups gathered by the Dutch International Institute of Social History (http://www.iisg.nl/occasio/). The Digital Archive for Chinese Studies (DACHS), run jointly by the universities of Heidelberg and Leiden (http://www.sino.uni-heidelberg.de/dachs/), collects Web content related to social and political discourse in the Peoples Republic of China, aiming thereby to place this beyond the control of state or party institutions (Wagner & Gross, 2004).

The UK is fairly unique in that a group of cultural heritage organisations have joined together to form a co-operative approach to collecting and preserving Web sites, the UK Web Archiving Consortium (http://www.webarchive.org.uk/). The British Library leads the consortium, partners including The National Archives, the national libraries of Wales and Scotland, the Joint Information Systems Committee (an organisation concerned primarily with IT in further and higher education), and the library of the Wellcome Trust (a medical charity). The members hope that working together will help share costs, risks and experiences, and reduce the burden on any single organisation. Each partner focuses on selecting sites relevant to their main interests and negotiating the rights to preserve them with their owners.

3. Current approaches to collecting Web sites

To date, there have been two main approaches to the collection of Web sites, one based on the automatic harvesting of Web content that match certain criteria, the other on the selection and capture of sites considered to be of value. Increasingly, these approaches are seen as being complimentary, and many current initiatives use (or propose to use) a combination of collection approaches.

3.1 Automatic harvesting

The automatic harvesting of Web content is based on the use of Web crawler programs - similar to those used by Web search services - which are periodically deployed on the Web to follow hyperlinks and download the content that matches particular user-defined criteria. The criteria used by national library based initiatives tend to focus on a particular national domain. For example, the Swedish Royal Library's Kulturarw[3] project (http://www.kb.se/kw3/) defined this as Web sites in the `.se` domain, those sites physically located in Sweden, and sites in other domains selected for their relevance to Sweden (Arvidson, Persson & Mannerheim, 2001). Hakala (2004) outlines how the harvesting approach was used to collect the Finnish Web domain.

> *A harvester is first fed a set of links (URLs) to qualifying Web documents ... These pages are fetched and analysed in order to find hyperlinks (further URLs) embedded in them. Those URLs which match the specified selection criteria are put aside. The next step is to use these URLs to retrieve a second batch of documents, which is processed in a similar manner. This process goes on, until every valid document has been retrieved. With this simple method, large portions of the Web can be covered quickly.*

The Internet Archive (http://www.archive.org/) uses broadly the same type of harvesting technique, but takes a much less selective approach. It has been collecting Web sites on a broad scale since 1996, and in that time has amassed a huge amount (>1 Petabyte) of online content, much of it currently available through the 'Wayback Machine.' The Internet Archive is unique amongst Web archiving initiatives in that it is focused on collecting and storing as much Web content as possible, meaning that it has in less than ten years become a unique record of Web history. Despite its importance, however, the harvesting techniques used mean that sites are sometimes missing significant content or functionality (e.g., Day, 2003). In addition, an evaluative study by Thelwall and Vaughan (2004) found evidence of uneven representation of different countries and potentially other biases related to site age and link structures.

With the exception of the NEDLIB harvester (Hakala, 2001), the software used for harvesting Web sites has usually been adapted from crawler programs originally developed for other purposes. However, as part of the International Internet Preservation Consortium (http://www.netpreserve.org/), the Internet Archive and some of its national library partners are developing and testing a configurable, open-source, 'archival-quality' harvester program known as Heritrix (http://crawler.archive.org/).

3.2 Selective capture or deposit

Other Web archiving initiatives have taken a more selective strategy, based on the selection of individual Web sites for capture or deposit. This was the approach pioneered by the National Library of Australia (NLA) with the development of its PANDORA archive (http://pandora.nla.gov.au/). This was initiated in 1997 with the development of a 'proof-of-concept' archive and a conceptual framework for a sustainable service. Sites are first selected according to the library's selection guidelines (http://pandora.nla.gov.au/selectionguidelines.html) and the appropriate rights negotiated with their owners. Once this has been agreed, the sites are collected using gathering or mirroring tools. If this is not possible, the national library makes arrangements with the site owner to receive the files on physical media or via ftp or e-mail. The general selection criteria for PANDORA include the resource's relevance to Australia (regardless of physical location), its 'authority' and perceived long-term research value. There are more 'inclusive' selection guidelines for particular social and topical issues and specific ones for particular types of

material. The NLA has also developed a suite of software tools known as the PANDORA Digital Archiving System (PANDAS) that can initiate the gathering process, create and manage metadata, undertake quality control and manage access to gathered resources.

The NLA currently uses Web site capture tools like HTTrack to harvest selected sites. Alternatively, site owners or administrators could be persuaded to deposit a copy (or snapshot) of Web content. This was the method used, for example, by the National Archives and Records Administration (NARA) for its collection of US federal agency Web sites at the end of the Clinton Administration (Bellardo, 2001).

3.3 Combined approaches

There has always been some discussion as to which one of the approaches is most appropriate. In practice, however, they all have some advantages and some disadvantages.

The deposit approach, for example, may work best in situations where there is close co-operation with depositors and where the incremental cost of deposit is not too high. Supporters of the automatic crawler-based approach argue that it is by far the cheapest way to collect Web content. Thus Mannerheim (2001) has noted that, "it is a fact that the selective projects use more staff than the comprehensive ones." However, most of the current generation of Web crawler programs are still unable to cope with some types of database-driven sites, and additionally can run into difficulty with items that need browser plug-ins or use scripting techniques. The selective approach allows more time to address and rectify these problems but strictly limits the range of resources that can be collected. It also provides an opportunity to address legal issues like rights clearance, where this would be required. For some of these reasons, some initiatives are increasingly emphasising the need to use a combination of approaches. It is perhaps indicative that the National Library of Australia, the pioneer of the selective approach, has recently (June-July 2005) undertaken its first whole domain harvest using the Heritrix crawler (Koerbin, 2005). The Bibliothèque nationale de France (BnF) has also supported the use of both approaches, proposing the use of harvesting for those Web sites that can be adequately reached by crawler programs, and more selective approaches for the collection of the (so-called) 'deep Web' (Masanès, 2002).

3.4 The International Internet Preservation Consortium

Perhaps the most significant development in Web archiving in recent years has been the establishment of the International Internet Preservation Consortium (IIPC) in 2003. The IIPC provides a means for national library initiatives to co-operate with the Internet Archive over the development of the technical tools required for harvesting, storing and providing access to Web sites. The BnF leads the consortium, current partners including the Internet Archive and the national libraries of Australia, Canada (Library and Archives Canada), Denmark, Finland, Iceland, Italy, Norway, Sweden, the UK (British Library), and the United States (Library of Congress). While currently primarily focused on the development of tools and technologies to support Web preservation activities, the consortium also provides a means for the Internet Archive and the libraries to co-operate on facilitating international coverage and access to Web content stored across national domains - a global distributed collection preserved by sustainable organisations (Hallgrímsson, 2004).

Consortium work to date has included the development of the following tools, part of the IIPC Web Archiving Toolset (from: http://www.netpreserve.org/software/downloads.php):

- *Heritrix* - an open-source, configurable Web crawler designed specifically for the harvesting of Web sites, developed by the Internet Archive with the Nordic National Libraries (http://crawler.archive.org/)
- *DeepArc* - an editor that facilitates the mapping of a relational data model to an XML Schema, e.g. to support the export or capture of deep Web content, developed by the BnF (http://bibnum.bnf.fr/downloads/deeparc/)
- *BAT* (BnfArcTools) - tools to support the processing of file formats, e.g. the ARC format used by the Internet Archive and generated by Heritrix, developed by the BnF (http://bibnum.bnf.fr/downloads/bat/)
- *NutchWAX* (Nutch with Web Archive eXtensions) - a tool for searching Web archives, adapted from the open source Nutch search engine, developed by the Internet Archive and the Nordic National Libraries (http://archive-access.sourceforge.net/projects/nutch/)
- *WERA* (Web aRchive Access) - a viewer for the navigation and full-text searching of Web archives, building on the Nordic Web Archive's toolset (Brygfjeld, 2002), developed by the Internet Archive and the National Library of Norway (http://archive-access.sourceforge.net/projects/wera/)
- *Xinq* (XML INQuire) - a tool for browsing XML databases, developed by the National Library of Australia (http://www.nla.gov.au/xinq/)

Other IIPC work includes the development of a Web Archiving Metadata Set defining the metadata (about harvesting parameters, Web site contexts, etc.) that can be automatically generated or captured by IIPC tools and various working groups dealing with less technical issues. More information on the IIPC can be found at the consortium's Web site (http://www.netpreserve.org/).

4. Issues

Despite almost ten years of experience with Web archiving a number of potential issues remain. This final section will outline a selection of these.

4.1 Conceptual problems with defining the 'Web'

In order to preserve something effectively, it is important to understand exactly what is being preserved. While individual components of the Web (e.g., images, PDF documents, text marked up in HTML) are relatively easily to understand, the Web itself is a more nebulous entity. For example, many 'deep Web' sites just provide browser-friendly access to a managed database that predates the Web and will, just as likely, survive it. Also, some database-driven sites serve up Web content differently depending on the software available to the user or are otherwise customisable. In these contexts, what exactly does it mean to preserve the Web? Are crawler programs the best way to collect this content?

4.2 Copyright and content liability

Some of the most significant challenges to Web archiving initiatives are legal ones, chiefly related to copyright or liability for content made publicly available through archives. As part of the feasibility study undertaken before the setting up of the UK Web Archiving Consortium, Andrew Charlesworth (2003) undertook a detailed survey of the legal issues related to the collection and preservation of Internet resources. This noted

that the legal environment in many countries is unappreciative of - or sometimes hostile to - the potential role of Web archives and that there was also a severe lack of case law (see also: Borrull & Oppenheim, 2004). While the most obvious legal problem related to copyright law, Charlesworth also noted potential problems with liability, e.g. for defamatory or otherwise illegal content, or for breaches of data protection laws. The UK feasibility study concluded that the 'safest' way of overcoming these challenges would be adopt the selective route only - thus excluding at source those resources that could have liability problems - and to develop effective rights management policies, combined with effective processes for the removal of (or the limiting of access to) certain types of material. Other initiatives deal with this problem in different ways, e.g. by restricting end-user access to the Web content that they collect.

4.2 Scale

Another significant problem is the vast size of the Web. Because Web archives tend to collect multiple snapshots of Web content, they can grow very quickly indeed. For example, the Internet Archive (http://www.archive.org/about/faqs.php) claims to give access to over 40 billion Web pages, approximately a Petabyte of data and growing at the rate of 20 Terabytes a month. On the other hand, some national domain crawls can have relatively modest storage requirements, especially when compared with the high levels of demand (for example) associated with multimedia content. For example, Hakala (2004) reported that a crawl of the Finnish Web domain in 2002 collected a total of 500 Gigabytes. A crawl of the Portuguese Web in 2003 processed 3.8 million URLs and downloaded 78 Gigabytes of data (Gomes & Silva, 2005). By contrast, the most recent harvest of the Australian Web domain took six weeks and captured 185 million documents or 6.69 Terabytes of data (Koerbin, 2005).

4.3 Dynamism

Another potential problem is the Web's dynamic nature, meaning that many pages, sites and domains are continually changing or disappearing. Back in 2001, Lawrence, *et al*. (2001) cited an Alexa Internet (http://www.alexa.com/) estimate that Web pages disappeared on average after 75 days. This process of change often leaves no trace, except in some cases a URL cited elsewhere that retrieves a 404 or similar error. For this reason, a major concern of many Web-archiving initiatives - including the Internet Archive's special collections - have been the Web sites associated with ephemeral events, things like political elections, 9/11 or Hurricane Katrina. For example, Colin Webb (2001) of the National Library of Australia noted that much of the Web presence associated with the Sydney Olympic Games in 2000 disappeared almost faster than the athletes themselves.

A further set of problems relates to the ongoing evolution of Web-based technologies. While some of the most basic Web standards and protocols have remained relatively stable since the 1990s, there have been major changes in the way some Web sites are managed. For example, Web content is increasingly beginning to be delivered from dynamic databases. Some of these may be extremely difficult to replicate in repositories without detailed documentation about database structures and the software used. Other sites may use specific software that may not be widely available, or may adopt non-standard features that may not work in all browsers (e.g., Fitch, 2003). The user interactivity inherent in the new technologies sometimes characterised as Web 2.0 (e.g., O'Hear, 2005) also provides additional technical challenges for those wishing to collect Web sites for continued access.

4.4 Access

The Web developed in a decentralised way. There is, therefore, no single organisation (or set of organisations) that can be held responsible for the Web. It has no governing body that can mandate the adoption of standards or Web site preservation policies. Instead, most decisions about Web content and delivery are devolved down to Web site owners themselves. Bollacker, Lawrence and Giles (2000) point out, "the Web database draws from many sources, each with its own organization."

With the exception of the Internet Archive, Web preservation initiatives tend to focus on defined subsets of the Web, e.g. by national domain, subject or organisation type. Those cultural heritage organisations interested in the preservation of the Web tend to approach it from their own professional perspective. Archives will be interested in the recordkeeping aspects of Web sites, art galleries in conserving artworks that use Web technologies, historical data archives in those sites considered to have long-term social or political importance, etc. Some national libraries have provided a slightly wider perspective, for example, viewing a whole national Web domain (however defined) as suitable for collection and preservation. In practice, this decentralised approach to Web archiving may prove useful, although it will need significant co-operation to avoid duplication and to help facilitate user access to what could become a confusing mess of different initiatives and repositories.

4.5 Digital preservation and curation

Many current Web archiving initiatives have been, until now, primarily focused on the collection of resources rather than on their long-term preservation or curation. In the short to medium-term, there is nothing wrong with this. After all, content cannot be preserved for the long term if it no longer exists. However, there remains a need to consider how those Web sites being collected at the moment can continue to be accessible over time and how Web archiving initiatives should fit into the wider landscape of digital preservation and curation.

For example, those that are now essentially project-type activities will need to be become firmly embedded into the core activities of their host institutions and have sustainable business models. In this regard, it is encouraging to note how many of the current initiatives are funded from the host organisations' own budgets.

5. Conclusions

This paper has attempted to review the current state-of-the-art in collecting and preserving Web sites. It has attempted to review some of the reasons why Web archiving initiatives have been developed and reviewed the different approaches to content collection based on domain harvesting or selective methods, and the ongoing work of the International Internet Preservation Consortium. A final section outlined some remaining issues relating to defining the Web, the size and scale of Web archives, the dynamic nature of the Web and the technologies that underlie it, and the partly unresolved issue of long-term preservation.

References

Arvidson, A., Persson, K., Mannerheim, J. (2001). "The Royal Swedish Web Archive: a 'complete' collection of Web pages." *International Preservation News*, 26, 10-12. Retrieved December 5, 2005, from http://www.ifla.org/VI/4/news/ipnn26.pdf

Arzberger, P., Schroeder, P., Beaulieu, A., Bowker, G., Casey, K., Laaksonen, L., Moorman, D., Uhlir, P., Wouters, P. (2004). "An international framework to promote access to data." *Science*, 303, 1777-1778

Bar-Ilan, J., Peritz, B.C. (2004). "Evolution, continuity, and disappearance of documents on a specific topic on the web: A longitudinal study of 'informetrics'." *Journal of the American Society for Information Science and Technology*, 55(11), 980-990

Bellardo, L. J. (2001). "Memorandum to Chief Information Officers: snapshot of public Web sites." Washington, D.C.: National Archives & Records Administration, 12 January

Bollacker, K. D., Lawrence, S., Giles, C. L. (2000). "Discovering relevant scientific literature on the Web." *IEEE Intelligent Systems*, 15, 42-47

Borrull, A. L., Oppenheim, C. (2004). "Legal aspects of the Web." *Annual Review of Information Science and Technology*, 38, 483-548

Brygfjeld, S. A. (2002). "Access to Web archives: the Nordic Web Archive Access Project." *Zeitschrift für Bibliothekswesen und Bibliographie*, 49, 227-231

Charlesworth, A. (2003). *A study of legal issues related to the preservation of Internet resources in the UK, EU, USA and Australia*. Retrieved December 5, 2005, from http://www.jisc.ac.uk/index.cfm?name=project_webarchiving

Crichlow, R., Davies, S., Wimbush, N. (2004). "Accessibility and accuracy of Web page references in 5 major medical journals." *JAMA: the Journal of the American Medical Association*, 292(22), 2723-2724

Day, M. (2003). *Collecting and preserving the World Wide Web: a feasibility study undertaken for the JISC and Wellcome Trust*. Retrieved December 5, 2005, from http://www.jisc.ac.uk/index.cfm?name=project_webarchiving

Dellavalle, R. P., Hester, E. J., & Heilig, L. F., Drake, A. L., Kuntzman, J. W., Graber, M., Schilling, L. M. (2003). "Going, going, gone: lost Internet references." *Science*, 302, 787-788

Fitch, K. (2003). "Web site archiving: an approach to recording every materially different response produced by a Website." 9th Australasian World Wide Web Conference, Sanctuary Cove, Queensland, Australia, July 5-9, 2003. Retrieved December 5, 2005, from http://ausweb.scu.edu.au/aw03/papers/fitch/

Gillies, J., Cailliau, R. (2000). *How the Web was born: the story of the World Wide Web*. Oxford: Oxford University Press.

Gomes, D., Silva, M. J. (2005). "Characterising a national community Web." *ACM Transactions on Internet Technology*, 5(3), 508-531

Hakala, J. (2001). "The NEDLIB Harvester." *Zeitschrift für Bibliothekswesen und Bibliographie*, 48, 211-216

Hakala, J. (2004). "Archiving the Web: European experiences." *Program*, 38(3), 176-183

Hallgrímsson, Th. (2003). "Survey of Web archiving in Europe." E-mail sent to list web-archive@cru.fr, 3 February

Hallgrímsson, Th. (2004). "The International Internet Preservation Consortium (IIPC)." Conference of Directors of National Libraries (CDNL 2005), Oslo, Norway, 14-18 August 2005. Retrieved December 5, 2005, from http://consorcio.bn.br/cdnl/cdnl_2005.htm

Hendler, J. (2003). "Science and the Semantic Web." *Science*, 299, 520-521

Hester, E. J., Heilig, L. F., Drake, A. L., Johnson, K. R., Vu, C. T., Schilling, L. M., Dellavalle R. P. (2004). "Internet citations in oncology journals: A vanishing resource?" *Journal of the National Cancer Institute*, 96(12), 969-971

Koehler, W. (2004) "A longitudinal study of Web pages continued: a consideration of document persistence". *Information Research*, 9(2), January 2004. Retrieved December 5, 2005, from http://informationr.net/ir/9-2/paper174.html

Koerbin, P. (2005). "Report on the crawl and harvest of the whole Australian Web domain undertaken during June and July 2005." Retrieved December 5, 2005, from http://pandora.nla.gov.au/documents/domain_harvest_report_public.pdf

Laurillard, D. (2005). "E-learning in higher education." In: Ashwin, P. (ed.), *Changing higher education: the development of learning and teaching*." London: RoutledgeFalmer.

Lawrence, S., Pennock, D. M., Flake, G.W., Krovetz, R., Coetzee, F. M., Glover, E., Nielsen, F. Å, Kruger, A., Giles, C. L. (2001). "Persistence of Web references in scientific research." *Computer*, 34(2), 26-31

Lenhart, A., Horrigan, J., Fallows, D. (2004). *Content creation online*. Washington, D.C.: Pew Internet & American Life Project, 29 February. Retrieved December 5, 2005, from http://www.pewinternet.org/

Lyman, P. (2002). "Archiving the World Wide Web." In: *Building a national strategy for digital preservation*. Washington, D.C.: Council on Library and Information Resources, 38-51. Retrieved December 5, 2005, from http://www.clir.org/pubs/abstract/pub106abst.html

Mannerheim, J. (2001). "The new preservation tasks of the library community." *International Preservation News*, 26, 5-9. Retrieved December 5, 2005, from http://www.ifla.org/VI/4/news/ipnn26.pdf

Masanès, J. (2002). "Towards continuous Web archiving: first results and an agenda for the future." *D-Lib Magazine*, 8(12), December. Retrieved December 5, 2005, from http://www.dlib.org/dlib/december02/masanes/12masanes.html

National Archives of Australia. (2001). *Archiving Web resources: a policy for keeping records of Web-based activity in the Commonwealth Government*. Retrieved December 5, 2005, from ttp://www.naa.gov.au/recordkeeping/er/web_records/intro.html

National Archives of Australia. (2001). *Archiving Web resources: guidelines for keeping records of Web-based activity in the Commonwealth Government*. Retrieved December 5, 2005, from http://www.naa.gov.au/recordkeeping/er/web_records/intro.html

O'Hear, S. (2005, 15 November). "Seconds out, round two." *The Guardian*, 15 November. Retrieved December 5, 2005, from http://education.guardian.co.uk/elearning/story/0,10577,1642281,00.html

Public Record Office. (2001). *Management of electronic records on Websites and Intranets: an ERM toolkit*, v. 1.0, December. Retrieved December 5, 2005, from http://www.nationalarchives.gov.uk/electronicrecords/advice/

Spinellis, D. (2003). "The decay and failure of Web references." *Communications of the ACM*, 46(1), 71-77.

The National Archives. (2003). *Operational Selection Policy 27: The selection of Government Websites*. Retrieved December 5, 2005, from http://www.nationalarchives.gov.uk/recordsmanagement/selection/ospintro.htm

Thelwall, M., Vaughan, L. (2004). "A fair history of the Web? Examining country balance in the Internet Archive." *Library & Information Science Research*, 26, 162-176

Wagner, R. G. (2004) "Harvesting the Web, preserving Chinese voices: the Digital Archive for Chinese Studies (DACHS) in Heidelberg." International Conference on Sinological Resources in the Digital Era, Taipei, Taiwan, Republic of China, 7-9 December 2004. Retrieved December 5, 2005, from http://www.sino.uni-heidelberg.de/dachs/publ.htm

Webb, C. (2001). "Who will save the Olympics?" OCLC/Preservation Resources Symposium, Digital Past, Digital Future: an Introduction to Digital Preservation, OCLC, Dublin, Ohio, 15 June 2001. Retrieved December 5, 2005, from http://www.oclc.org/education/conferences/presentations/2001/preservation/

Wren, J. D. (2004). "404 not found: the stability and persistence of URLs published in MEDLINE." *Bioinformatics*, 20(5), 668-672.

Acknowledgements

UKOLN is funded by the Joint Information Systems Committee (JISC) and the Museums, Libraries and Archives Council (MLA), as well as by project funding from the JISC, the UK research councils, the European Union, and other sources. UKOLN also receives support from the University of Bath, where it is based.

Author details

Michael Day is a member of UKOLN's research and development team based at the University of Bath (United Kingdom). Since joining UKOLN in 1996, he has worked on a large number of research projects, mostly focused on the topics of metadata, interoperability and digital preservation. For example, he led UKOLN's involvement in the Cedars (CURL Exemplars in Digital Archives) project, contributed to the development of its preservation metadata specification, and produced the *Cedars Guide to Preservation Metadata* (2002). He was also a member of the international working group on preservation metadata commissioned by the Research Libraries Group and OCLC Online Computer Library Center that produced *A Metadata Framework to Support the Preservation of Digital Objects* in 2002. He also project managed a feasibility study into Web-archiving in the United Kingdom, undertaken on behalf of the Joint Information Systems Committee (JISC) and the Wellcome Trust between 2002 and 2003. More recently, he worked on the first phase of the eBank UK project, which explored the use of repository tools to support the dissemination and aggregation of crystallographic datasets. Since 2004 he has mainly worked for the Digital Curation Centre, a service funded by the JISC and the e-Science Core Programme to provide a UK focus for research and development into digital curation issues and to promote good practice for the preservation of digital objects in the higher education sector. He is also a member of the digital preservation cluster of the DELOS Network of Excellence on Digital Libraries.