Preserving the Fabric of Our Lives: A Survey of Web Preservation Initiatives

Michael Day

UKOLN, University of Bath, Bath BA2 7AY, United Kingdom
m.day@ukoln.ac.uk

Abstract. This paper argues that the growing importance of the World Wide Web means that Web sites are key candidates for digital preservation. After an brief outline of some of the main reasons why the preservation of Web sites can be problematic, a review of selected Web archiving initiatives shows that most current initiatives are based on combinations of three main approaches: automatic harvesting, selection and deposit. The paper ends with a discussion of issues relating to collection and access policies, software, costs and preservation.

1 Introduction

In a relatively short period of time, the Internet has become a pervasive communication medium. For example, Castells opens his book on *The Internet Galaxy* by saying, "the Internet is the fabric of our lives" [1]. Of the many tools that make-up the Internet, perhaps the most widely used is the World Wide Web.

The Web now plays a major role in research, e.g., being used as a medium for the dissemination of information about institutions and research projects, and as a means of distributing data, publications, learning resources, etc. The Web is also used to provide user-friendly interfaces to a wide range of important databases, e.g. of bibliographic or sequence data, many of which predate the Web itself. Hendler has accurately written that scientists have become "increasingly reliant" on the Web for supporting their research. For example, he notes 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" [2].

The Web is also now widely used in non-research contexts. It has developed very rapidly as a major facilitator of personal communication, electronic commerce, publishing, marketing, and much else. Since its inception, the Web has seen the development of new types of online commerce (e.g., companies like eBay or Amazon.com) as well as a major move by existing organisations (e.g., the news media, television companies, retailers, etc.) to develop a significant presence on the Web. On a smaller scale, many individuals have begun to use services like GeoCities (http://geocities.yahoo.com/) to create Web pages that focus on their personal interests and hobbies, e.g. for genealogy. In summary, the Web's importance can be

T. Koch and I.T. Sølvberg (Eds.): ECDL 2003, LNCS 2769, pp. 461–472, 2003.

gauged by Lyman's recent comment that it has become "the information source of first resort for millions of readers" [3]. For this reason, the preservation of Web sites has begun to be addressed by a variety of different initiatives.

UKOLN undertook a survey of existing Web archiving initiatives as part of a feasibility study carried out for the Joint Information Systems Committee (JISC) of the UK further and higher education funding councils and the Library of the Wellcome Trust [4]. After a brief description of some of the main problems with collecting and preserving the Web, this paper outlines the key findings of this survey.

2 Problems with Web Archiving

There are a number of reasons why the Web or Web sites can be difficult to collect and preserve. Some of these are technical, e.g. related to the size and nature of the Web itself, while others are related to legal or organisational issues.

2.1 Technical Challenges

One general problem is that the Web is huge and still growing. This means that no single organisation can realistically hope to collect the entire Web for preservation. Until now, the Internet Archive has attempted to do this, but in the longer term Web preservation will be best seen as a collaborative activity. Estimates of Web size and growth rates vary, but all agree that the Web has until now demonstrated a consistent year on year growth. Rapid growth rates are attested by studies undertaken in the late 1990s at the NEC Research Institute [5, 6], by a survey undertaken in 2000 by Cyveillance [7] and by the annual statistics on Web server numbers collected by the Web Characterization Project of OCLC Research (http://wcp.oclc.org/). In 2000, Lyman and Varian collated these (and other) figures and concluded that the total amount of information on the 'surface Web' was somewhere between 25 and 50 terabytes [8]. It is likely to be far larger by now.

It is worth pointing out that these figures hide a large proportion of the Web. In 2001, Bar-Ilan pointed out that size estimates of the Web only tended to count "static pages, freely accessible to search engines and Web users" [9]. A large number of other pages were not so accessible; chiefly those created dynamically from databases, or with other accessibility barriers (e.g., with password protection) or format problems. A much-cited paper produced by the search company BrightPlanet estimated that this subset of the Web - sometimes known as the 'invisible,' 'hidden' or 'deep Web' - could be up to 400 to 500 times bigger than the surface Web [10].

Another potential problem is the Web's dynamic nature, meaning that many pages, sites and domains are continually changing or disappearing. In 2001, Lawrence, *et al.* cited an Alexa Internet (http://www.alexa.com/) estimate that Web pages disappear on average after 75 days [11]. This rate of decay means that, without some form of collection and preservation, there is a danger that invaluable scholarly, cultural and scientific resources will be unavailable to future generations. The process of change

often leaves no trace. Casey has commented that she got the impression that "a significant percentage of Web sites have the life span of a housefly and about as much chance as meeting an untimely end" [12]. A major concern has been the Web sites of major events, e.g. political elections or sporting events. Colin Webb of the National Library of Australia (NLA) noted that much of the Web presence associated with the Sydney Olympic Games in 2000 disappeared almost faster than the athletes themselves [13].

A further set of problems relates to the ongoing evolution of Web-based technologies. While some 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. All of this provides technical challenges for those wishing to collect and preserve Web sites.

It is perhaps also worth emphasising that the Web is also a 'moving-target' for preservation purposes. In the near future, there are likely to be changes as the Web evolves to take account of the W3C's vision of a 'Semantic Web,' whereby information is given well-defined meanings, so that machines can begin to understand it, and process it accordingly (http://www.w3c.org/2001/sw/). Other drivers of change will be the development of Web services technology for business to business activity and the continued adoption of computational grid technologies by scientists.

2.2 Legal Challenges

Some of the most significant challenges to Web archiving initiatives are legal ones, chiefly related to copyright or liability for content made available through archives. As part of the JISC/Wellcome Trust feasibility study, Charlesworth undertook a detailed survey of the legal issues related to the collection and preservation of Internet resources [14]. This noted that the legal environment in many countries is unappreciative of - or sometimes inhospitable to - the potential role of Web archives. While the most obvious legal problem relates to copyright law, there are also potential problems with defamation, content liability and data protection. The 'safest' way of overcoming these challenges would be to select resources carefully - thus excluding at source those resources that may 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.

2.3 Organisational Challenges

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 point out that "the Web database draws from many sources, each with its own organization" [15].

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.

Another general issue is quality. While the Web contains much that would definitely be considered to have continuing value, (e.g., the outputs of scholarly and scientific research, the Web sites of political parties, etc.), there is much other content that is of low-quality (or even worse). Chakrabarti, *et al*. note that each Web page might "range from a few characters to a few hundred thousand, containing truth, falsehood, wisdom, propaganda or sheer nonsense" [16]. A survey of academic opinion in 2001 showed that while there was a general satisfaction with the Web as a research tool, many had significant concerns about accuracy, reliability and value of the information available [17].

3 Web Archiving Initiatives

At the present time, there are a variety of different organisation types pursuing Web archiving initiatives. These have been initiated by archives, national libraries, historical data archives and even some Web site owners themselves (e.g., the British Broadcasting Corporation). Perhaps the most ambitious and well-known Web archiving initiative at the moment is that run by the US-based Internet Archive [18]. This privately funded organisation has been collecting Web pages since 1996 and has generated a huge database of Web pages that can be accessed via the 'Wayback Machine' as well as co-operating with the Library of Congress and the Smithsonian Institution on the creation of special collections.

National Libraries are responsible for some of the more visible and successful Web archiving initiatives. Following the early examples of the Swedish and Australian national libraries, pilot Web archiving initiatives have now been launched in many other countries, including Austria, Denmark, Finland, France, Germany, Iceland, New Zealand, the United States and the United Kingdom. A survey by Hallgrímsson of European national libraries in late 2002 and early 2003 showed that 15 out of the 25 libraries that responded had some kind of Web-archiving initiative underway [19]. In some countries (e.g., France) some of the intellectual property rights issues have

been dealt with by including Web archiving amongst the national library's legal deposit responsibilities. Other national library initiatives, following the example of the National Library of Australia in the PANDORA archive, seek permission from Web site owners before adding them to the library's collections.

National archives have also begun to get involved in the collection and preservation of Web sites, especially where Web sites are understood to have evidential value. Sometimes this interest manifests itself in the form of guidance for Web managers. For example, the National Archives of Australia [20, 21] and the Public Record Office (now the National Archives) in the UK [22] have each issued detailed electronic records management (ERM) guidelines for Web site managers. Some other archives have already begun to capture and accession Web sites. 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 [23]. In the UK, the Public Record Office (PRO) accessioned a snapshot of the No. 10 Downing Street Web site (http://www.number-10.gov.uk/) just before the General Election of June 2001. Ryan has described some of the technical problems that the PRO had with migrating the Web site so that it could work in a different technical environment [24].

Some universities and scholarly societies have supported smaller Web archiving initiatives. These 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/).

3.1 Approaches to the Collection of Web Sites

Currently, there are three main approaches to the collection of Web sites. The first of these is based on the deployment of *automatic harvesting* or gathering tools, generally utilising Web crawler technologies. The second is based on the *selection and capture* of individual Web sites. The third approach is based on a more traditional *deposit* model.

Automatic harvesting approaches

The Internet Archive (http://www.archive.org/) and the Swedish Royal Library's Kulturarw^3 project (http://www.kb.se/kw3/) were amongst the first to adopt the automatic harvesting approach. In this, Web crawler programs - similar to those used by Web search services - are used to follow links and download content according to particular collection rules. The Kulturarw^3 crawler, for example, is set-up to only collect Web sites in the `.se` domain, those sites physically located in Sweden, and sites in other domains selected for their relevance to Sweden [25]. The Internet Archive collects the Web on a much broader scale, but their crawlers will not harvest sites (or parts of them) protected by the robot exclusion protocol. A number of other national-based initiatives have followed the automatic approach, most notably the

Helsinki University Library in its experiments with the NEDLIB harvester [26] and the Austrian On-Line Archive (AOLA) [27].

Selective capture approaches

Some initiatives have taken a much more selective strategy based on the selection of individual Web sites for inclusion in an archive. This was the general 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 NLA's selection guidelines 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 (http://pandora.nla.gov.au/selectionguidelines.html). 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 selective approach has also been experimented with by the Library of Congress in its Minerva project [28] and the British Library for the 'Britain on the Web' pilot.

Deposit approaches

Deposit approaches are based on site owners or administrators depositing a copy or snapshot of their site in a repository. This is a strategy used, for example, by NARA for its collection of US federal agency Web sites in 2001 and by Die Deutsche Bibliothek (DDB) for the collection of dissertations and some online publications (http://deposit.ddb.de/).

Combined approaches

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

The deposit approach, for example, may work in particular situations where there is agreement with depositors and where the incremental cost of deposit (to the depositors) is not too expensive. Supporters of the automatic crawler-based approach argue that it is by far the cheapest way to collect Web content. Thus Mannerheim notes that, "it is a fact that the selective projects use more staff than the comprehensive ones" [29]. However, the current generation of Web crawler technologies cannot cope with some database-driven sites and can sometimes 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 limits the range of resources that can be collected. For some of these reasons, some initiatives are increas-

ingly emphasising the need to use a combination of approaches. The pioneer of this approach has been the Bibliothèque nationale de France (BnF), which has investigated the preservation of the 'French Web' as part of its responsibilities for the legal deposit of French publications.

The French initiative has, for example, experimented with refining the automatic harvesting approach by taking account of a Web page's change frequency and by attempting to measure site importance automatically:

- Change frequency - the French Web crawling experiments collected information about when each page was retrieved and whether there had been any updating. This information was then stored in an XML-based 'site-delta' that could be used to judge the required frequency of crawl [30].
- Page importance - this can be calculated in a similar way to that used by search services like Google [31]. The BnF regards the automatic calculation of page importance based on link structures as a way of focusing attention on the part of the Web that is most well-used. An initial evaluation, comparing a sample of automated rankings with evaluations of site relevance by library staff, showed a good degree of correlation [32].

The BnF have also proposed a selective strategy for the collection of the 'deep Web' - known as the 'deposit track.' This follows a similar approach to PANDORA, and is based on the evaluation of sites by library staff followed by liaison with site owners over their deposit into the library. The BnF undertook a pilot project to test this approach in 2002. The BnF notes that Web harvesting technologies may have some uses in helping to support the deposit track. For example, robots can be used to analyse the technical features of crawled material, helping to detect deep-Web sites for which selective attention may be required.

4 Discussion

The survey of Web archiving initiatives looked in particular at a number of issues of relevance to the feasibility study. These included collection and access policies, software, relative costs and sustainability.

4.1 Collection Policies and Coverage

Both automatic harvesting-based and selective approaches to Web archiving are dependent to some extent upon the development of collection policies. The automatic approaches will usually define this by national domain and server location, supplemented by a list of other sites that have been individually judged to be of interest. In some cases, sites can be automatically excluded from the collection process, e.g. by taking account of standards for robot exclusion. Selective approaches will normally develop more detailed collection guidelines, often based on a resource's relevance to the collecting institution's designated communities, their provenance or their suitability for long-term research. Sites that change frequently may have to be collected on a

regular basis. In addition, many of the Web sites that meet selection guidelines on other criteria may include errors, be incomplete or have broken links. The collecting institution will need to decide whether these 'features' are an essential part of the resource being collected and act accordingly. Once a site loses its active hyperlinks with the rest of the Web, it will be very difficult to evaluate whether these links were working at the time of collection. Whether this is a problem will depend on whether the Web site is being preserved for its informational or evidential value.

Table 4.1 is an attempt to show the relative size (in Gigabytes) of selected Web-archives as of late 2002 and early 2003. Although the figures are approximate and do not take into account things like compression or crawl frequency, it shows that those initiatives based on automatic harvesting have normally resulted in collections that are considerably larger than those using the selective approach. As may be expected, the initiative with the largest collection is the Internet Archive with over 150 terabytes of data (and growing). While the largest of the selective archives (PANDORA) has a size not dissimilar to two of the smaller harvesting-based archives (Austria and Finland), by comparison, the British Library and Library of Congress pilot archives are extremely small.

Table 1. Approximate size of selected Web-archiving initiatives, 2002

Country	Initiative/Organisation	Approach	Size (Gb.)	No. sites
International	Internet Archive	harvesting	> 150,000.00	
Sweden	Kulturarw3	harvesting	4,500.00	
France	Bibliothèque nationale de France	combined	< 1,000.00	
Austria	AOLA	harvesting	448.00	
Australia	PANDORA	selective	405.00	3,300
Finland	Helsinki University Library	harvesting	401.00	
UK	Britain on the Web	selective	0.03	100
USA	MINERVA	selective		35

Source: Day, 2003 [4]

4.2 Access Policies

More thought needs to be given to how access is provided to the large databases that can be generated by the automatic harvesting approach. The Internet Archive's Wayback Machine is a useful and interesting 'window' on the old Web, but currently users need to know the exact URLs that they are looking for before they can really begin to use it. Alternative approaches to access might involve the generation or reuse of metadata or the development of specialised Web indexes designed to search extremely large databases of Web material, possibly including multiple versions of pages harvested at different times. From 2000 to 2002, the Nordic Web Archive Access project (NWA) investigated the issue of access to collections of Web documents [33]. The result was an open-source NWA Toolset (http://nwa.nb.no/) that searches and navigates Web document collections. The current version of the NWA Toolset supports a commercial search engine provided by the Norwegian company FAST (http://www.fastsearch.com/).

4.3 Software

Specific software tools have been developed or adapted to support both collecting approaches. The Swedish Royal Library's Kulturarw[3] initiative adapted the Combine crawler [34], while other countries have used or evaluated the NEDLIB harvester developed by the Finnish CSC [26]. The experiments at the BnF tested the Xyleme crawler for Web collection. The Internet Archive uses the Alexa crawler, and this software is completely rewritten every other year.

The selective approach has seen the use of a variety of site mirroring and harvesting tools. PANDORA started using Harvest, but currently has adopted a twin approach, using HTTrack and Teleport Pro/Exec. The British Library, the Library of Congress and the BnF have also used HTTrack in their pilot projects. The NLA have themselves developed an archive management system called PANDAS to help facilitate the collection process, to deal with metadata and quality control, and to manage access. This has had a significant impact by increasing automation and tools for these processes and consequently reducing staff time and costs incurred.

4.4 Costs

Costs between the initiatives vary widely. Arms, *et al*., have estimated that the selective approach - as carried out in the Library of Congress's Minerva pilot - is "at least 100 times as expensive as bulk collection" on the Internet Archive model [28]. In addition it should be recognised that a significant element of the additional cost of the selective approach can be incurred in undertaking rights clearances. However, although this approach has additional costs, it does allow many materials gathered in this way (for example in PANDORA), to be made publicly accessible from the archive via the Web. This generates substantially higher use and gives wider accessibility than other methods.

4.5 Long-Term Preservation

Many current Web archiving initiatives have been, until now, focused on the collection of resources rather than on their long-term preservation. In the short to medium-term, there is nothing wrong with this, but there remains a need to consider how those Web sites being collected at the moment can be preserved over time, and what this may mean. This may include assessments of various proposed preservation strategies (migration, emulation, etc.) and the implementation of repositories based, for example, on the standard *Reference Model for an Open Archival Information System (OAIS)* [35]. One key issue for repositories will be how to ensure the authenticity of digital objects, i.e. to verify that they are exactly what they (or their metadata) claim to be [36]. This may be dependent on cryptographic techniques applied by the repository or by the encapsulation of objects in descriptive metadata. What is clear, however, is that in many cases the nature of the repository itself will serve as a surrogate for an object's authenticity. So, for example, Hirtle has said, "the fact that digital information is found within a trusted repository may become the base upon which all further assessment of action builds" [37].

Ways of defining trusted repositories have recently been investigated by a working group established by the Research Libraries Group (RLG) and OCLC. In 2002, this group published a report outlining a framework of attributes and responsibilities of trusted digital repositories. Trusted repositories are defined as "one whose mission is to provide reliable, long-term access to managed digital resources to its designated community, now and in the future" [38]. The report defines the key attributes of such repositories (e.g. organisational viability and financial sustainability) and outlines their main responsibilities. The working group further recommended that a framework should be developed in order to support the certification of digital repositories. The RLG together with NARA is currently setting up a task force to undertake this (http://www.rlg.org/longterm/certification.html).

Web archiving initiatives need to be aware of the requirements for becoming trusted digital repositories. 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 have been funded from the host organisations' own budgets.

5 Conclusions

It is hoped that this short review of existing Web archiving initiatives has demonstrated that collecting and preserving Web sites is an interesting area of research and development that has now begun to move into a more practical implementation phase. To date, there have been three main approaches to collection, characterised in this report as 'automatic harvesting,' 'selection' and 'deposit.' Which one of these has been implemented has normally depended upon the exact purpose of the archive and the resources available. Naturally, there are some overlaps between these approaches but the current consensus is that a combination of them will enable their relative strengths to be utilised. The longer-term preservation issues of Web archiving have been explored in less detail.

Acknowledgements. This paper is based on work undertaken for a feasibility study into Web-archiving undertaken for the Joint Information Systems Committee and the Library of the Wellcome Trust. UKOLN is funded by the JISC and Resource: The Council for Museums, Archives & Libraries, as well as by project funding from the JISC and the European Union. UKOLN also receives support from the University of Bath, where it is based.

References

1. Castells, M.: The Internet galaxy: reflections on the Internet, business, and society. Oxford University Press, Oxford (2001)
2. Hendler, J.: Science and the Semantic Web. Science **299** (2003) 520–521

3. Lyman, P.: Archiving the World Wide Web. In: Building a national strategy for digital preservation. Council on Library and Information Resources and Library of Congress, Washington, D.C. (2002) 38–51 http://www.clir.org/pubs/abstract/pub106abst.html
4. Day, M.: Collecting and preserving the World Wide Web: a feasibility study undertaken for the JISC and Wellcome Trust (February 2003) http://library.wellcome.ac.uk/projects/archiving_feasibility.pdf
5. Lawrence, S., Giles, C.L.: Searching the World Wide Web. Science **280** (1998) 98–100
6. Lawrence, S., Giles, C.L.: Accessibility of information on the Web. Nature **400** (1999) 107–109
7. Murray, B., Moore, A.: Sizing the Internet. Cyveillance White Paper. Cyveillance, Inc. (10 July 2000) http://www.cyveillance.com/web/downloads/Sizing_the_Internet.pdf
8. Lyman, P., Varian, H.R.: How much information? University of California at Berkeley, School of Information Management and Systems, Berkeley, Calif. (2000) http://www.sims.berkeley.edu/research/projects/how-much-info/internet.html
9. Bar-Ilan, J.: Data collection methods on the Web for informetric purposes: a review and analysis. Scientometrics **50** (2001) 7–32
10. Bergman, M.K.: The deep Web: surfacing hidden value. Journal of Electronic Publishing **7** (August 2001). Available at: http://www.press.umich.edu/jep/07-01/bergman.html
11. Lawrence, S., Pennock, D.M., Flake, G.W., Krovetz, R., Coetzee, F.M., Glover, E., Nielsen, F.Å, Kruger, A., Giles, C.L.: Persistence of Web references in scientific research. Computer **34** (February 2001) 26–31
12. Casey, C.: The cyberarchive: a look at the storage and preservation of Web sites. College & Research Libraries **59** (1998) 304–310
13. Webb, C.: Who will save the Olympics? OCLC/Preservation Resources Symposium, Digital Past, Digital Future: an Introduction to Digital Preservation, OCLC, Dublin, Ohio (15 June 2001) http://www.oclc.org/events/presentations/symposium/preisswebb.shtm
14. Charlesworth, A.: A study of legal issues related to the preservation of Internet resources in the UK, EU, USA and Australia (February 2003) http://library.wellcome.ac.uk/projects/archiving_legal.pdf
15. Bollacker, K.D., Lawrence, S., Giles, C.L.: Discovering relevant scientific literature on the Web. IEEE Intelligent Systems **15** (2000) 42–47
16. Chakrabarti, S., Dom, B.E., Kumar, S.R., Raghavan, P., Rajagopalan, S., Tomkins, A., Kleinberg, J., Gibson, D.: Hypersearching the Web. Scientific American **280** (June 1999) 44–52
17. Herring, S.D.: Using the World Wide Web for research: are faculty satisfied? Journal of Academic Librarianship **27** (2001) 213–219
18. Kahle, B.: Way back when ... New Scientist **176** (23 November 2002) 46–49
19. Hallgrímsson, T.: Survey of Web archiving in Europe. E-mail sent to list web-archive@cru.fr (3 February 2003)
20. National Archives of Australia: Archiving Web resources: a policy for keeping records of Web-based activity in the Commonwealth Government (January 2001) http://www.naa.gov.au/recordkeeping/er/web_records/archweb _policy.pdf
21. National Archives of Australia: Archiving Web resources: guidelines for keeping records of Web-based activity in the Commonwealth Government (March 2001) http://www.naa.gov.au/recordkeeping/er/web_records/archweb _guide.pdf
22. Public Record Office: Managing Web resources: management of electronic records on Websites and Intranets: an ERM toolkit, v. 1.0 (December 2001)

23. Bellardo, L.J.: Memorandum to Chief Information Officers: snapshot of public Web sites. National Archives & Records Administration, Washington, D.C. (12 January 2001) http://www.archives.gov/records_management/cio_link/memo_to_cios.html
24. Ryan, D.: Preserving the No 10 Web site: the story so far. Web-archiving: managing and archiving online documents, DPC Forum, London (25 March 2002) http://www.jisc.ac.uk/dner/preservation/presentations/pdf/Ryan.pdf
25. Arvidson, A., Persson, K., Mannerheim, J.: The Royal Swedish Web Archive: a "complete" collection of Web pages. International Preservation News **26** (December 2001) 10–12 http://www.ifla.org/VI/4/news/ipnn26.pdf
26. Hakala, J.: The NEDLIB Harvester. Zeitschrift für Bibliothekswesen und Bibliographie **48** (2001) 211–216
27. Rauber, A., Aschenbrenner, A., Witvoet, O.: Austrian Online Archive processing: analyzing archives of the World Wide Web. In: Agosti, M., Thanos, C. (eds.): Research and advanced technology for digital libraries: 6th European conference, ECDL 2002, Rome, Italy, September 16–18, 2002. Lecture Notes in Computer Science, Vol. 2458. Springer-Verlag, Berlin (2002) 16–31
28. Arms, W.Y., Adkins, R., Ammen, C., Hayes, A.: Collecting and preserving the Web: the Minerva prototype. RLG DigiNews, **5** (April 2001) http://www.rlg.org/preserv/diginews/diginews5-2.html#feature1
29. Mannerheim, J.: The new preservation tasks of the library community. International Preservation News **26** (December 2001) 5–9 http://www.ifla.org/VI/4/news/ipnn26.pdf
30. Abiteboul, S., Cobéna, G., Masanès, J., Sedrati, G.: A first experience in archiving the French Web. In: Agosti, M., Thanos, C. (eds.): Research and advanced technology for digital libraries: 6th European conference, ECDL 2002, Rome, Italy, September 16–18, 2002. Lecture Notes in Computer Science, Vol. 2458. Springer-Verlag, Berlin (2002) 1–15
31. Brin, S., Page, L.: The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems **30** (1998) 107–117
32. Masanès, J.: Towards continuous Web archiving: first results and an agenda for the future. D-Lib Magazine **8** (December 2002) http://www.dlib.org/dlib/december02/masanes/ 12masanes.html
33. Brygfjeld, S.A.: Access to Web archives: the Nordic Web Archive Access Project. Zeitschrift für Bibliothekswesen und Bibliographie **49** (2002) 227–231
34. Ardö, A., Lundberg, S.: A regional distributed WWW search and indexing service - the DESIRE way. Computer Networks and ISDN Systems **30** (1998) 173–183
35. CCSDS 650.0-B-1: Reference Model for an Open Archival Information System (OAIS). Consultative Committee on Space Data Systems (2002) http://wwwclassic.ccsds.org/ documents/pdf/CCSDS-650.0-B-1.pdf
36. Lynch, C.: Authenticity and integrity in the digital environment: an exploratory analysis of the central role of trust. In: Authenticity in a digital environment. Council on Library and Information Resources, Washington, D.C. (2000) 32–50 http://www.clir.org/pubs/abstract/ pub92abst.html
37. Hirtle, P.B: Archival authenticity in a digital age. In: Authenticity in a digital environment. Council on Library and Information Resources, Washington, D.C., (2000) 8–23 http://www.clir.org/pubs/abstract/ pub92abst.html
38. RLG/OCLC Working Group on Digital Archive Attributes: Trusted digital repositories: attributes and responsibilities (2002) http://www.rlg.org/longterm/repositories.pdf