



---

# Collecting and preserving the World Wide Web

---

A feasibility study undertaken for the  
JISC and Wellcome Trust

Michael Day  
UKOLN, University of Bath

Version 1.0 - 25 February 2003

## UKOLN

UKOLN (<http://www.ukoln.ac.uk/>) is a national focus of expertise in digital information management. It provides policy, research and awareness services to the UK library, information and cultural heritage communities.

UKOLN aims to inform practice and influence policy in the areas of: digital libraries, metadata and resource discovery, distributed library and information systems, bibliographic management, web technologies, and public library networking. It provides network information services, including the *Ariadne* and *Cultivate Interactive* magazines and runs workshops and conferences.

UKOLN is funded by Resource: the Council for Museums, Archives and Libraries, the Joint Information Systems Committee of the Higher Education Funding Councils (JISC), 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.

## Contents

|   |    |
|---|----|
| Acknowledgements .....                                  | v  |
| 1. Management summary and recommendations .....         | 1  |
| 1.1. Management summary .....                           | 1  |
| 1.2. Recommendations.....                               | 3  |
| 2. Introduction .....                                   | 5  |
| 2.1. Report background.....                             | 5  |
| 2.2. Why preserve the World Wide Web? .....             | 5  |
| 2.3. Challenges of preserving the Web.....              | 6  |
| 2.4. Responsibility for preserving the Web.....         | 8  |
| 2.5. The role of the Wellcome Trust and the JISC.....   | 9  |
| 3. The size and nature of the World Wide Web .....      | 15 |
| 3.1. Size and growth of the Web .....                   | 15 |
| 3.2. The dynamic nature of the Web.....                 | 15 |
| 3.3. The 'deep Web'.....                                | 15 |
| 3.4. The future of the Web .....                        | 16 |
| 3.5. Defining and characterising the UK Web space ..... | 16 |
| 3.6. Conclusions.....                                   | 17 |
| 4. Review of existing Web archiving initiatives.....    | 18 |
| 4.1. The Internet Archive.....                          | 19 |
| 4.2. Selective approaches.....                          | 20 |
| 4.3. Harvesting-based approaches.....                   | 24 |
| 4.4. Combined approaches.....                           | 26 |
| 4.5. Other initiatives.....                             | 28 |
| 4.6. Evaluation .....                                   | 29 |
| 4.7. Other issues.....                                  | 30 |
| 4.8. Conclusions.....                                   | 31 |
| 5. Implementation.....                                  | 32 |
| 5.1. Cost estimates.....                                | 32 |
| 5.2. Staff skills .....                                 | 33 |
| 5.3. Collaboration .....                                | 34 |
| 5.4. Sustainability.....                                | 36 |
| 6. References.....                                      | 37 |
| Appendix A: Evaluation of selected medical sites .....  | 45 |

|  |    |
|--|----|
| Appendix B: Evaluation of selected eLib project Web sites .....  | 61 |
| Appendix C: The World Wide Web.....                              | 66 |
| Appendix D: The UK Web space.....                                | 76 |
| Appendix E: Questionnaire sent to Web archiving initiatives..... | 80 |
| Appendix F: Abbreviations used.....                              | 83 |

## Acknowledgements

In particular, I would like to thank:

- Neil Beagrie (JISC) and Robert Kiley (Wellcome Library), the project co-ordinators, for their contribution to the process and content of this report
- Andrew Charlesworth (University of Bristol) for providing an excellent overview of legal issues
- The members of the advisory board who provided useful comments on draft versions of both reports: Stephen Bury (British Library), Adrian Brown (Public Record Office), Simon Jennings (Resource Discovery Network), Julien Masanès (Bibliothèque nationale de France), Margaret Phillips (National Library of Australia), David Ryan (Public Record Office), John Tuck (British Library), Colin Webb (National Library of Australia)
- Brewster Kahle and Michele Kimpton (Internet Archive) for answering questions on their visit to the UK in August 2002
- Those who replied to the questionnaire on Web archiving initiatives or who provided other information. These include: Allan Arvidson (Royal Library, Sweden), Andreas Aschenbrenner (Austrian On-Line Archive), Stephen Bury (British Library), Gerard Clifton (National Library of Australia), Leonard DiFranza (Internet Archive), Lisa Gray (RDN BIOME), Julien Masanès (Bibliothèque nationale de France), Margaret Phillips (National Library of Australia), Andreas Rauber (TU Wien), Dave Thompson (National Library of New Zealand), Deborah Woodyard (British Library)
- Brian Kelly (UKOLN, UK Web Focus) for obtaining statistics from Netcraft, and for providing information on the size and composition of the UK Web (Appendix D)

Michael Day  
Bath, 25th February 2003



## 1. Management summary and recommendations

### 1.1. Management summary

#### 1.1.1. Why collect and preserve the Web?

- In the short time since its invention, the World Wide Web has become a vital means of facilitating global communication and an important medium for scientific communication, publishing, e-commerce, and much else. The 'fluid' nature of the Web, however, means that pages or entire sites frequently change or disappear, often without leaving any trace.
- In order to help counter this change and decay, Web archiving initiatives are required to help preserve the informational, cultural and evidential value of the World Wide Web (or particular subsets of it).

#### 1.1.2. Why should the Wellcome Library be interested in this?

- The Wellcome Library has a particular focus on the history and understanding of medicine. The Web has had a huge impact on the availability of medical information and has also facilitated new types of communication between patients and practitioners as well as between these and other types of organisations. The medical Web, therefore, has potential long-term documentary value for historians of medicine.
- To date, however, there has been no specific focus on collecting and preserving medical Web sites. While the Internet Archive has already collected much that would be of interest to future historians of medicine, a preliminary analysis of its current holdings suggest that significant content or functionality may be missing.
- There is, therefore, an urgent need for a Web archiving initiative that would have a specific focus on preserving the medical Web. The Wellcome Library is well placed to facilitate this and such an initiative would nicely complement its existing strategy with regard to preserving the record of medicine past and present.

#### 1.1.3. Why should the JISC be interested in this?

The Joint Information Systems Committee of the Higher and Further Education Funding Councils (JISC) has a number of areas where Web archiving initiatives would directly support its mission. These include:

- JISC funds a number of development programmes. It therefore has an interest in ensuring that the Web based outputs of these programmes (e.g. project records, publications) persist and remain available to the community and to JISC. Many of the Web sites of projects funded by previous JISC programmes have already disappeared.
- JISC also supports national development of digital collections for HE/FE and the Resource Discovery Network (RDN) services that select and describe high-quality Web resources judged to be of relevance to UK further and higher education. A Web archiving initiative could underpin this effort by preserving copies of some of these sites, e.g. in case the original sites change or disappear. The expertise and subject knowledge of the RDN could in turn assist development of national and special collections by bodies such as the national libraries or Wellcome Trust. These collections would be of long-term value to HE/FE institutions.
- JISC also funds the JANET network used by most UK further and higher education institutions and, as its operator, UKERNA has overall responsibility for the ac.uk domain.

#### 1.1.4. Collaboration

- Collaboration will be the key to any successful attempt to collect and preserve the Web.
- The Web is a global phenomenon. Many attempts are being made to collect and preserve it on a national or domain level, e.g. by national libraries and archives. This means that no one single initiative (with the exception of the Internet Archive) can hope for total coverage of the Web. Close collaboration between different Web archiving initiatives, therefore, will be extremely important, e.g. to avoid unnecessary duplication in coverage or to share in the development of tools, guidelines, etc.
- More specifically, there is a need for all organisations involved in Web archiving initiatives in the UK to work together. In particular there is the opportunity to work closely with the British Library as it develops its proposals for Web-archiving as part of the national archive of publications. Potentially, many different types of organisation have an interest in collecting and preserving aspects of the UK Web, while the British Library (BL), the Public Record Office (PRO) and the British Broadcasting Corporation (BBC) have already begun to experiment with Web archiving. The Digital Preservation Coalition (DPC) is well placed to provide the general focus of this collaboration, although there may be a need for specific communications channels.

#### 1.1.5. Challenges

The Web poses preservation challenges for a number of reasons:

- The Web's fast growth rate and 'fluid' characteristics mean that it is difficult to keep up-to-date with its content sufficiently for humans to decide what is worth preserving.
- Web technologies are immature and evolving all the time. Increasingly, Web content is delivered from dynamic databases that are extremely difficult to collect and preserve. Other sites use specific software (e.g. browser plug-ins) that may not be widely available or use non-standard features that may not work in all browsers. Other Web sites may belong to the part of the Web that is characterised by the term 'deep Web' and will be hard to find using most Web search services and maybe even harder to preserve.
- Unclear responsibilities for preservation - the diverse nature of the Web means that a variety of different organisation types are interested in its preservation. Archives are interested in Web sites when they may contain records, libraries when they contain publications or other resources of interest to their target communities. The global nature of the Web also means that responsibility for its preservation does not fall neatly into the traditional national categories.
- Legal reasons issues relating to copyright, the lack of legal deposit mechanisms (at least in the UK), liability issues related to data protection, content liability and defamation. These represent serious problems and are dealt with in a separate report that has been prepared by Andrew Charlesworth of the University of Bristol.

#### 1.1.6. Approaches

Since the late 1990s, a small number of organisations have begun to develop approaches to the preservation of the Web, or more precisely, well-defined subsets of it. Those organisations that have developed initiatives include national libraries and archives, scholarly societies and universities. Perhaps the most ambitious of these initiatives is the Internet Archive. This US-based non-profit organisation has been collecting broad snapshots of the Web since 1996. In 2001, it began to give public access to its collections through the 'Wayback Machine.'

Current Web archiving initiatives normally take one of three main approaches:

- Deposit, whereby Web-based documents or 'snapshots' of Web sites are transferred into the custody of a repository body, e.g. national archives or libraries.
- Automatic harvesting, whereby crawler programs attempt to download parts of the surface Web. This is the approach of the Internet Archive (who have a broad collection strategy) and some national libraries, e.g. Sweden and Finland.
- Selection, negotiation and capture, whereby repositories select Web resources for preservation, negotiate their inclusion in co-operation with Web site owners and then capture them using software (e.g. for site replication or mirroring, harvesting, etc.). This is the approach of the National Library of Australia and the British Library's recent pilot project.

These are not mutually exclusive. Several Web archiving initiatives (e.g. the Bibliothèque nationale de France and the National Library of New Zealand) plan to use combinations of both the selective and harvesting based approaches. The selective approach can deal with some level of technical complexity in Web sites, as the capture of each can be individually planned and associated with migration paths. This may be a more successful approach with some parts of the so-called 'deep Web.' However, hardware issues aside, collection would appear to be more expensive (per Gigabyte archived) than the harvesting approach. Estimates of the relative costs vary, but the selective approach would normally be considerably more expensive in terms of staff time and expertise. This simple assessment, however, ignores related factors related to the cost of preservation over time (whole of life costs), the potential for automation, and quality issues (i.e., fitness for purpose).

## 1.2. Recommendations

1. Both the JISC and Wellcome Trust should attempt to foster good institutional practice with regard to the management of Web sites. For example, they could consider the development of Web site management guidelines for adoption by their user communities or for inclusion in grant conditions, etc.
2. Until the exact position is clarified by legislation, a *selective* approach to Web archiving - with appropriate permissions secured - would be the best way to proceed for both the JISC and the Wellcome Trust. Other methods of archiving will need to be approached with caution due to problems with copyright and other legal issues (see also the conclusions and recommendations in the associated legal study by Andrew Charlesworth).
3. If the Wellcome Trust is to meet its strategic objectives in extending its collecting activities into the digital environment, then it will need to consider undertaking some kind of Web archiving activity. To achieve this the following approach is recommended:

*Establish a pilot medical Web archiving project using the selective approach, as pioneered by the National Library of Australia (see also Recommendation 5).*

*This pilot should consider using the NLA's PANDAS software for this archiving activity. This pilot could be run independently or as part of a wider collaborative project with other partners.*

*The high-quality medical Web sites identified in the RDN gateway OMNI should be considered as the starting point for any medical Web archiving initiative.*

*The library will need to develop a Web archiving selection policy to help ensure that it can archive a broad, representative sample of medical Web sites. This policy should allow for the inclusion of 'low-quality' (e.g. medical quackery) sites that may be of interest to future historians.*

4. If the JISC is to meet its strategic objectives for management of JISC electronic records, in digital preservation and collection development then it will also need to consider

undertaking some form of Web archiving. To achieve this the following approach is recommended:

*Establish a pilot project to test capture and archiving of JISC records and publications on project Web sites using the selective approach, as pioneered by the National Library of Australia (see also Recommendation 5).*

*As part of this pilot, the JISC should define selection policies and procedures.*

*This pilot should consider using the NLA's PANDAS software for this archiving activity. This pilot could be run independently or as part of a wider collaborative project with other partners.*

*Work in collaboration with emerging initiatives from the British Library and Wellcome Trust. There are significant synergies with some existing JISC services. Web sites identified and described by the RDN gateways could be the starting points for any selective subject-based Web archiving initiatives in the UK. The RDN gateways contain (November 2002) descriptions of over 60,500 Internet resources available on the Web.*

5. Research: the current generation of harvesting technologies has limitations with regard to dealing with deep-Web sites. This has a particular impact on Web archiving approaches based on automatic harvesting. While some research is being carried out on this issue from a Web search perspective, there is a need for more collaborative research into this issue from the perspective of Web archives.
6. Collaboration: for both the JISC and the Wellcome Trust there is significant opportunity for partnership on Web-archiving. For example, there will be opportunities to collaborate on strategic, technical, organisational or content issues.

For the UK, both should attempt to work closely with the British Library, the other copyright libraries, the Public Record Office, data archives and the e-Science centres that have experience of managing large volumes of data. The focus for this collaborative activity could be within the Digital Preservation Coalition (DPC). On an international level, close co-operation with institutions like the US National Library of Medicine and the Internet Archive will be important.

As an exemplar of collaboration, it is recommended that the JISC and the Wellcome Library should seek to work together and with other partners to create their pilot Web archiving services. Not only will this realise economies of scale, but more importantly provide a model demonstrating how collaboration can work in practice.

## 2. Introduction

### 2.1. Report background

In March 2002, the Joint Information Systems Committee (JISC) and the Library of the Wellcome Trust invited proposals for an evaluation and feasibility study of Web archiving. The Wellcome Trust's interest in this subject is motivated by its interest in furthering medical research and the study of the history and public understanding of medicine. A proposal to extend the Wellcome Library's collecting activities to the Web has been endorsed by its Library Advisory Committee and the Medicine, Society and History Committee. The JISC's interest in Web archiving is prompted by its dual roles as a provider of Web-based services to the UK further education (FE) and higher education (HE) communities and as a funder of research and development projects. Both organisations are members of the Digital Preservation Coalition (DPC) and therefore committed to supporting collaboration to advance a common agenda in digital preservation.

The aims of this study are to provide the JISC and Wellcome Trust with:

- An analysis of existing Web archiving arrangements to determine to what extent they address the needs of the UK research and FE/HE communities. In particular this is focused on an evaluation of sites available through the Internet Archive's Wayback Machine, to see whether these would meet the needs of their current and future users.
- To provide recommendations on how the Wellcome Library and the JISC could begin to develop Web archiving initiatives to meet the needs of their constituent communities.

This study will first outline the urgent need for Web archiving initiatives and indicate the benefits these would have for the user communities of the JISC and Wellcome Trust. This will be followed by an attempt to define the nature of the World Wide Web (and the UK part of it) and an introduction and evaluation of existing Web archiving initiatives. There will follow a short section on implementation. Andrew Charlesworth of the University of Bristol is dealing with legal issues in a separate report.

### 2.2. Why preserve the World Wide Web?

Since its inception in the early 1990s, the World Wide Web has become a pervasive communication medium for many different kinds of information. Its importance can be gauged by Peter Lyman's recent comment that it has become "the information source of first resort for millions of readers" (Lyman, 2002, p. 38).

The Web had its origins in the academic world. It has been noted that the Web started out as a participatory groupware system, "a way for high-energy physicists to share their research data and conclusions" (O'Reilly, 2001, p. 53). The Web continues to play a major role in scholarly and scientific communication, it being used as a medium for disseminating information about institutions and research projects and as a means of distributing publications (e.g., through e-print archives or e-journals), data or learning resources. The Web also began to be used to provide popular interfaces to databases, including services like MEDLINE (e.g., through the National Library of Medicine's PubMed) and sequence databases like GenBank or the EMBL Nucleotide Sequence Database. Hendler (2003, p. 520) has accurately written that scientists have become "increasingly reliant" on the Web for supporting research.

*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.*

However, the Web's relative ease of use meant that it did not take very long before users outside of research institutions began to develop implementations in other areas. Thus, the Web as it exists today is 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 businesses (e.g., companies like eBay or Amazon.com) as well as a move by existing organisations (e.g., the news media, television companies, retailers, etc.) to develop a presence on the Web. On a smaller scale, many individuals have begun to use services like GeoCities.com (<http://geocities.yahoo.com/>) to create Web pages that focus on their personal interests and hobbies.

The rapid growth of the Web, however, has been largely chaotic. This means that while the Web contains much that would definitely be considered to be of continuing value (e.g., the outputs of scholarly and scientific research, the Web sites of political parties, etc.), there is much content that is of low-quality (or worse). Chakrabarti, *et al.*, (1999, p. 44) note that each Web page might "range from a few characters to a few hundred thousand, containing truth, falsehood, wisdom, propaganda or sheer nonsense."

The dynamic nature of the Web means that pages and whole sites are continually evolving, meaning that pages are frequently changed or deleted. Alexa Internet (<http://www.alexa.com/>) once estimated that Web pages disappear after an average time of 75 days (Lawrence, *et al.*, 2001, p. 30). This rate of decay means that without collection and preservation there is a danger that invaluable scholarly, cultural and scientific resources will be unavailable to future generations. Major concerns are 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 (Webb, 2001). The NLA's PANDORA archive, therefore, deliberately collected samples of these Web sites before they were deleted or changed (<http://pandora.nla.gov.au/col/c4006>).

One response to the problem of these vanishing Web sites is to provide Web site owners and publishers with best practice guidelines on the preservation of Web resources. These typically contain guidance on how Web sites should be implemented with regard to their long-term management, e.g. the consistent adoption of standards. For example, the National Library of Australia has published *Safeguarding Australia's Web resources: guidelines for creators and publishers* (<http://www.nla.gov.au/guidelines/webresources.html>), "to assist those creators and publishers who do not already have well established digital data management procedures in place." Similar guidelines have been produced for organisations with Web sites that contain official or public records, e.g. in the *Guidelines for UK Government Websites* published by the Office of the e-Envoy (2002). A set of Web site creation guidelines could also be included as part of grant conditions issued by funding institutions, as has happened with the technical standards and guidelines published to support the UK's NOF-Digitise programme (<http://www.peoplesnetwork.gov.uk/content/technical.asp>).

**Recommendation 1:** Both the JISC and Wellcome Trust should attempt to foster good institutional practice with regard to the management of Web sites. For example, they could consider the development of Web site management guidelines for adoption by their user communities or for inclusion in grant conditions, etc.

### 2.3. Challenges of preserving the Web

The Web, however, is a very difficult object to collect and preserve. The size and nature of the Web will be described in more detail in section 3 (and Appendices C & D), but it may be useful to summarise here some of the main reasons why Web-archiving can be problematic.

### 2.3.1. De-centralised organisation

Firstly, there is no single organisation (or set of organisations) responsible for the Web. It was developed in a decentralised way and has no governing body that can mandate the adoption of standards or Web site preservation policies. Instead, decisions about Web content and delivery are devolved down to Web site owners themselves. Bollacker, Lawrence & Giles (2000, p. 42) point out that "the Web database draws from many sources, each with its own organization."

With the exception of the Internet Archive, Web preservation initiatives naturally tend to concentrate on highly 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 those 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 perceptive, for example, viewing a whole national Web domain (however defined) as suitable for collection (e.g. through legal deposit legislation) 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. A decentralised approach would certainly be in accordance with the recommendations of the influential report (Garrett & Waters, 1996) produced by the Task Force on Archiving of Digital Information commissioned by the Commission on Preservation and Access (CPA) and the Research Libraries Group (RLG).

### 2.3.2. Dynamic nature

A second reason why the Web is difficult to preserve is its fluid nature (this is outlined in more detail in section 3, and in Appendix C). The Web takes the form of an ever-evolving information resource and, over time, pages, whole sites and even entire domains will appear and disappear, change, be restructured, taken over, etc. This process often leaves no trace. Casey (1998, p. 304) 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." Precise estimates of the average lifetime of a Web page are difficult to produce (and may be not be meaningful) but those figures that exist suggest that they should be counted in months rather than years. Lawrence, *et al.* (2001, p. 30) cited an Alexa Internet estimate that Web pages disappear on average after 75 days. Brewster Kahle of the Internet Archive extended this figure to 100 days in an article published in *USA Today* in October 2001 (Kornblum, 2001). Whatever the exact figure, all users of the Web will be aware of Web sites that have disappeared or have been restructured and will frequently find URL links that now only retrieve an HTTP 404 error.

### 2.3.3. Evolving technologies

A third 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. These may be extremely difficult to replicate in repositories without detailed information (metadata) about the software and database structures. Other sites use specific software (e.g. browser plug-ins) that may not be widely available, or 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.

### 2.3.4. Legal challenges

A fourth major problem that relates to Web archiving is its legal basis. Legal issues include copyright, the lack of legal deposit mechanisms (at least in the UK), liability issues related to data protection, content liability and defamation. These collectively represent serious impediments to Web archiving and are dealt with in a separate report prepared by Andrew Charlesworth of the University of Bristol. In this, Charlesworth investigated the legal

implications of different Web archiving strategies and recommended that the JISC and Wellcome Trust should pursue a Web archiving strategy based on the model adopted by the National Library of Australia's PANDORA initiative. In this, the national library expressly obtains the permission of the relevant rights holders before collecting the resource. This significantly reduces the risk of copyright infringement and allows for the content liability risk to be distributed between the Web site owner and the NLA, according to contract.

Charlesworth also looked at automatic harvesting-based approaches like those adopted by the Internet Archive and some national libraries. He argued that the Internet Archive approach might not be as risky as it first appeared, especially if adequate administrative precautions were taken. For example, the JISC or Wellcome Trust could set up a limited company to undertake harvesting, although Charlesworth noted this might be construed as encouraging or abetting copyright infringement. The national library initiatives were different in that they were usually carried out under some framework of legal deposit, but this did not apply at present in the UK. Charlesworth finally recommended, therefore, that the JISC and Wellcome Trust should consider future strategy to obtain the necessary changes in law that would allow the legal deposit and/or preservation of UK digital materials and specifically Web-based resources.

**Recommendation 2:** Until the exact position is clarified by legislation, a *selective* approach to Web archiving –with appropriate permissions secured – would be the best way to proceed for both the JISC and the Wellcome Trust. Other methods of archiving will need to be approached with caution due to problems with copyright and other legal issues (see also the conclusions and recommendations in the associated legal study by Andrew Charlesworth).

## 2.4. Responsibility for preserving the Web

In the absence of one single organisation that would be responsible for preserving the entirety of the World Wide Web, it is likely that the responsibility for preserving defined subsets of the Web will fall to a range of organisation types. Reflecting this, at the present time, there are a variety of different types of organisation pursuing Web archiving initiatives. These include archives, national libraries, historical data archives and even Web site producers 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. This privately funded organisation has been collecting Web pages since 1996 and has generated a huge database as well as co-operating with the Library of Congress and the Smithsonian Institution on creating special collections. National Libraries are probably responsible for some of the more visible and successful Web archiving initiatives that remain. Following the early examples of Sweden (Kulturarw<sup>3</sup>) and Australia (PANDORA), pilot Web archiving projects have been launched in other countries, including Austria, Finland, France, New Zealand, the US and the UK. In some countries (e.g. France, Sweden), 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 National Library of Australia's example in PANDORA, seek permission from Web site owners before adding them to the library's collections.

Amongst the other organisations that have experimented with Web archiving, many national archives are interested in the content of Web sites as public records. This is an issue that will become more important as government business is increasingly transacted over the Web. In response, some national archives have begun to issue guidelines for Web site managers, e.g. the National Archives of Australia (2001a; 2001b) and the UK's Public Record Office (2001). Some national archives have also begun to add Web sites to their collections. For example, the US National Archives and Records Administration (NARA) arranged for the deposit of snapshots of federal agency Web sites at the end of President Clinton's term of office in 2001.

In the same year, the Public Record Office (PRO) added a copy of the pre-election Number 10 Downing Street Web site to its collections.

Some universities and scholarly societies support 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 International Institute of Social History (<http://www.iisg.nl/occasio/>).

UK organisations have now begun to get involved in Web archiving initiatives. The Public Record Office has produced guidance for Web site managers and successfully tested the archiving of a key Web site. Some Web site producers, e.g. the British Broadcasting Corporation (BBC), manage their Web content for long-term preservation themselves.

The British Library has begun to experiment with a selective Web archiving pilot project, and collected 100 sites. So far, this effort has been relatively small-scale. However the British Library has plans to scale-up their pilot, if additional funding is forthcoming to selectively capture up to 10,000 Web sites and have indicated a strong desire to do this in partnership with other bodies.

Some of the reasons for the small-scale of existing efforts are legal. In his accompanying report on legal issues, Andrew Charlesworth notes that legal deposit in the UK does not at the moment apply to non-printed materials like Web sites.

*As such, the downloading and storage by one of the copyright depositories of material from a Web site, whether that site was based in the UK or elsewhere, would appear to be a straightforward infringement of copyright ...*

In accordance with this, there have been no attempts made to automatically harvest the UK public Web domain on the Swedish or Finnish model.

The relatively small-scale of the existing UK effort means that there many gaps in coverage. Some of these may be partly filled by the ongoing Web harvesting activities of the Internet Archive, which is probably the largest collection of archived UK Web pages in existence. However, a selective evaluation of the Internet Archive's holdings (see Appendices A and B) suggests that there remains a need for further Web archiving initiatives focused on particular user communities. This is where the Wellcome Trust and the JISC may have opportunities to initiate Web archiving themselves, and/or work in partnership with others to achieve this.

## **2.5. The role of the Wellcome Trust and the JISC**

### **2.5.1. The Wellcome Trust: archiving the medical Web**

The Wellcome Trust's interest Web archiving is motivated by its interest in furthering medical research and the study of the history and understanding of medicine.

The Web has had a huge impact on the availability of medical information for both patients and practitioners. This is partly a matter of quantity, in that the Web provides free access to a large (and expanding) volume of information that was previously inaccessible (Jadad & Gagliardi, 1998, p. 611). Foote (2001, p. 233) has noted that "Universities, government agencies, libraries, and pharmaceutical companies have all put health-related information on the Web, in the spirit of the free-exchange of ideas, of establishing good will, of self-aggrandizement, and of salesmanship." The Web, however, also has more subtle effects on health care. For example, Internet technologies (including the Web) can be used by patients with rare diseases to communicate with fellow sufferers and to provide focused consumer health care information (Patsos, 2001). The medical Web, therefore, has potential long-term documentary value for historians of medicine.

### Wayback Machine evaluation

In order to help determine whether the JISC and Wellcome Library should embark upon their own Web archiving programmes, or whether they could rely upon the existing Web holdings of the Internet Archive to meet the needs of their current and future uses, UKOLN undertook a small evaluation of sites available through the Wayback Machine. UKOLN looked at a sample of 31 key medical Web sites and a smaller sample of 14 JISC eLib project Web sites and evaluated to what extent older versions of the site were accessible through the Internet Archive. Full details of the evaluation can be found in Appendices A and B. The key findings of the evaluation were:

- Early JISC eLib Web sites examined were relatively simple and well represented in the Wayback Machine while more complex sites demonstrated a range of difficulties.
- Redirection problems: one of the most common problems found were links that redirected to current versions of Web pages. All site and database search facilities that were tested did this, as did entering passwords, some 'button' functions (e.g., those based on cgi-scripts) and internal navigation features that used JavaScript. On a few pages, even plain HTML hyperlinks did the same thing (e.g. Sanger Institute, 27 June 1997).
- Missing images: another common problem that was frequently encountered was missing images, especially in earlier versions of some Web sites. Sometimes this had a negative effect on site navigation.
- Missing links: many links tested did not retrieve any information. Very few Web sites appeared to be complete in the archive, although small sites based almost entirely on simple linked HTML pages fared much better than more complex sites. On one site (the evidence-based-health e-mail list on JISCmail), all of the significant links with content did not work.
- Non-functional multimedia content: the existence of interactive content or multimedia often effected the way in which archived pages would display.

To summarise, the Internet Archive provides an interesting window onto a great deal of the Web. It is a resource that will repay serendipitous surfing, despite the frequency of missing pages. In its existing form, however, it is not a substitute for a focused Web collection and preservation initiative for the purposes of JISC and the Wellcome Trust. However it does provide a valuable and extensive snapshot which could be used to complement any focussed initiative

#### Box 2.1 JISC and Medical Web sites in the Internet Archive

To date, however, there has been no specific focus on collecting and preserving the medical Web sites. Although the Internet Archive contains pages that will be of interest to future medical historians, a preliminary analysis of its 'Wayback Machine' highlights a number of problems (Box 2.1) that compromises its full usefulness to the research community.

Consequently, if the Wellcome Library is to meet its strategic objectives in extending its collecting activities into the digital environment, then it will need to consider undertaking some form of Web archiving activities.

### **2.5.2. Archiving the medical Web: how?**

The medical Web is large, although there is no easy way of defining its limits. To give a very rough idea of scale, in October 2002 a simple search of the term "medicine" on the Google search service (<http://www.google.com/>) retrieved over fifteen and a half million Web pages. Obviously a more selective approach will be required for a successful medical Web archiving initiative.

One approach would be to concentrate on particular Internet domains; e.g. that used by the UK's National Health Service (<http://www.nhs.uk/>) but that would only be useful for identifying a small subset of resources.

An alternative approach would be to concentrate on the resources selected for inclusion in selective services like the JISC-funded Resource Discovery Network (RDN) or the National electronic Library for Health (Muir Gray & de Lusignan, 1999). This would help limit focus to Web sites selected according to some quality criteria. To give an idea of the current scale of the 'high-quality' medical Web, OMNI (<http://omni.ac.uk/>) - the RDN's subject gateway for health and medicine - contained descriptions of 6,818 selected Web resources in October 2002.

In addition to providing any medical Web archiving initiative with a preliminary list of selected 'high-quality' sites, collaboration with the RDN may make it possible to share existing metadata. All records in OMNI, for example, contain short descriptions, subject terms (NLM classifications and MeSH) and some administrative metadata. If this can be shared, then a Wellcome-based medical Web archiving initiative would not have to invest time and money recreating all of this data.

Using the sites identified in OMNI as a starting point for a selective collection of the medical Web would make the problem more manageable, but would automatically exclude 'low-quality' sites that may be of interest to historians of medicine. Low-quality sites may include the Web pages of unqualified medical practitioners, those who sell alternative medicine remedies and drugs without prescription. These may help support future historical research into popular medical beliefs, in the same way that Porter (1989) once undertook a detailed historical study of medical quackery in England.

To address this type of use the Wellcome Library will need to develop a robust Web archiving selection policy to ensure that a representative sample of the medical Web - the bad and the ugly, as well as the good - is effectively collected.

**Recommendation 3:** If the Wellcome Trust is to meet its strategic objectives in extending its collecting activities into the digital environment, then it will need to consider undertaking some kind of Web archiving activity. To achieve this the following approach is recommended:

*Establish a pilot medical Web archiving project using the selective approach, as pioneered by the National Library of Australia. (See also Recommendation 5)*

*This pilot should consider using the NLA's PANDAS software for this archiving activity. This pilot could be run independently or as part of a wider collaborative project with other partners.*

*The high-quality medical Web sites identified in the RDN gateway OMNI should be considered as the starting point for any medical Web archiving initiative.*

*A robust Web archiving selection policy will be required to ensure that the Library can archive a broad range of medical Web sites.*

### **2.5.3. The Joint Information Systems Committee**

The mission of the Joint Information Systems Committee of the Higher and Further Education Funding Councils (JISC) is:

*to help further and higher education institutions and the research community realise their ambitions in exploiting the opportunities of information and communications technology by exercising vision and leadership, encouraging collaboration and co-operation and by funding and managing national development programmes and services of the highest quality.*

There are a number of areas where JISC should have an interest in Web archiving and where this will support its mission. These can be summarised as:

- Securing the electronic records and publications from its development programmes as disseminated on the Web
- Contributing to the development of services and collections based on quality Web resources within the sector
- Encouraging collaboration and co-operation with partners in the field of Web archiving
- Mobilising and developing relevant sectoral ICT skills and expertise

The cumulative results of JISC-funded projects have been one of the major factors responsible for determining the development of IT within FE and HE in the UK. The eLib Project for example is widely regarded as a landmark programme. According to the Public Record Office it represented "a watershed in how information is made available. It is likely that there will be studies of the programme undertaken in the future." However, out of seventy project Web sites created by the programme, fourteen have already been lost. The situation is even worse for the JISC Technology Applications Programme (JTAP) where only fifty-nine Web sites remain of the ninety-two originally created.

The reason for this loss is due to the JISC's current reliance on the institution hosting the project to continue to preserve and allow access to its results after the project has finished and the funding stopped. Often, within a matter of months, the project staff has moved on. Then when the institution's Web site is next revised the material is removed from the site - and effectively no longer available for use by the community.

The JISC also supports the development of electronic collections and Internet resource discovery services on behalf of UK HE and FE. A key JISC initiative in this area is the Resource Discovery Network or RDN (Dempsey, 2000). The RDN (<http://www.rdn.ac.uk/>) provides end-user access to high-quality Internet resources that have been selected, indexed and described by subject specialists in one of over sixty partner institutions. The service is a network made up of a central organisation - the Resource Discovery Network Centre (RDNC)

- and a number of independent service providers called 'hubs.' Subject-based hubs exist at the moment for hospitality, leisure, sport and tourism (ALTIS), the health and life sciences (BIOME), engineering, mathematics and computing (EEVL), the Humanities (Humbul Humanities Hub), the physical sciences (PSIgate) and the social sciences, business and law (SOSIG). Further hubs for the arts and creative industries (Artifact) and for geography and the environment (GESource) are currently under development.

RDN hubs typically provide one or more subject gateways (or Internet resource catalogues) that give access to descriptions of Internet resources that can be searched or browsed according to subject classification terms assigned by a subject specialist. All Internet resources included in RDN services are selected according to particular quality criteria. For example, the BIOME hub has published documents that explain and delineate the process of evaluating resources for inclusion in BIOME gateways (<http://biome.ac.uk/guidelines/eval/>). The RDN gateways collectively contain (November 2002) descriptions of over 60,500 Internet resources. All of these can be searched simultaneously from the main RDN Web pages using 'ResourceFinder.'

The focus of the RDN means that there are potential synergies with any UK-based Web archiving initiative. All resources included in RDN gateways conform to published quality criteria and have been judged by subject experts as being of use to UK FE or HE. The gateways also create basic resource descriptions (or metadata) for each resource. The Web sites identified and described by the RDN gateways, therefore, would be a logical starting point for selective subject-based Web archiving initiatives in the UK. A list of RDN resource URLs could also be used to 'seed' a crawler-based initiative.

JISC also funds the JANET network used by most UK HE and FE institutions and, as its operator, UKERNA has overall responsibility for the ac.uk domain. HE and FE Web sites tend to have a high informational value. They do not just contain information about the institution itself, but often give access to the minutes of internal meetings (e.g. through an Intranet), staff publications (e.g. through institutional e-print archives), research data, etc. Another focus of JISC activity might be to take added responsibility for the preservation of the ac.uk domain materials within a national framework.

Any initiative in this area would need to be undertaken in collaboration with the HE and FE institutions themselves and other partners.

### ***Archiving the Web: next steps for JISC***

The strengths and weaknesses of different approaches to Web archiving are explored elsewhere in the report. Consideration of legal issues by Andrew Charlesworth and consideration of the research value of records created by different methods suggests JISC should adopt a selective approach to harvesting and securing its project outputs at this time. The PANDAS software developed by the NLA looks particularly promising in this regard and close working relationships on preservation between the UK and Australia are already in place.

Using this Web archiving feasibility report as a starting point it is suggested the JISC should undertake a pilot project to test the use of the PANDAS software developed by NLA. This could create and test a Digital Project Archive for the capture and storage of JISC-funded project Web sites.

This pilot would need to define selection policy and collection procedures in conjunction with the JISC Electronic Records Manager. This will be complicated by the fact that many JISC project Web sites will include a mixture of different resource types, including some that would have traditionally been classed as 'publications' or 'records' (e.g., e-mail archives, meeting minutes, software specifications, etc.). In addition, sites may include a range of interactive content, multimedia, software, images and databases of content or metadata that have characteristics and properties which require different approaches to capture, hosting and preservation to records and publications.

The project would need to agree with JISC the appropriate metadata schema(s) and unique identifiers that would be adopted. It is likely that some additional metadata fields will be needed for recording archival context of the documents for JISC records management purposes. It could then seek to transfer content from selected programmes e.g., JTAP, eLib and JCALT to the archive, explore rights clearance procedures and procedures to trace and retrieve 'lost' content wherever possible. There is potential to undertake a pilot independently or with a wider brief and on shared funding basis with other partners. It should also be borne in mind that there may be differences between the collection of actively managed project Web sites and those where projects have ended and key staff have left the host institution. In the latter cases, it may be difficult to get 'on-site support' for collection or even a general agreement that the site should be preserved.

Preservation of online publications and Web sites for the UK domain is not covered by voluntary legal deposit and this area is of interest to the British Library and other deposit libraries, and the national archives. The UK Web domain is very large and selective archiving is intensive and requires subject knowledge. It could not be done for the whole of the UK by any single institution acting in isolation. It is ideally suited therefore to collaborative approaches. In addition to JISC other partners in the Digital Preservation Coalition have expressed interest in selective archiving and potential joint projects. It would be worthwhile encouraging co-ordination and collaboration of activity in the UK and exploring options for joint activity.

**Recommendation 4:** If the JISC is to meet its strategic objectives for management of JISC electronic records, in digital preservation and collection development then it will also need to consider undertaking some form of Web archiving. To achieve this the following approach is recommended:

*Establish a pilot project to test capture and archiving of JISC records and publications on project Web sites using the selective approach, as pioneered by the National Library of Australia (see also Recommendation 5).*

*As part of this pilot, the JISC should define selection policies and procedures.*

*This pilot should consider using the NLA's PANDAS software for this archiving activity.*

*This pilot could be run independently or as part of a wider collaborative project with other partners.*

*Work in collaboration with the British Library and Wellcome Trust. There are significant synergies with some existing JISC services and in particular with the (Resource Discovery Network (RDN)). the Web sites identified and described by the RDN gateways could be the starting points for any selective subject-based Web archiving initiatives in the UK.*

### 3. The size and nature of the World Wide Web

This section briefly investigates various attempts to characterise and quantify the World Wide Web by looking at specific challenges related to its preservation. The large size and dynamic nature of the Web make it a challenging object to preserve. More detailed information on these topics - including a more detailed attempt to characterise the 'UK Web' - is available in Appendices C and D.

#### 3.1. Size and growth of the Web

The first problem is that the Web is large and still growing. Estimates of its size and growth rate vary (and these to some extent depend on what is being counted) but all agree that the Web is large with a consistent year on year growth. A rapid growth rate is attested by studies undertaken between 1997 and 1999 at the NEC Research Institute (Lawrence & Giles, 1998; 1999), by the annual statistics collected by the Web Characterization Project of OCLC Research (<http://wcp.oclc.org/>), and by a survey undertaken by Cyveillance (Murray & Moore, 2000). In 2000, Lyman & Varian (2000) collated these (and other) figures and concluded that the total amount of information on the 'surface' Web was somewhere between 25 and 50 terabytes.

#### 3.2. The dynamic nature of the Web

Another important characteristic of the Web that has implications for its preservation is its dynamic nature. Web sites frequently appear and disappear, are updated and restructured. These factors lead to the Web's 'broken-link' problem, symbolised by the well-known HTTP Error 404. Lawrence, *et al.* (2001, p. 30) cited an Alexa Internet (<http://www.alexa.com/>) estimate that Web pages disappear after an average time of 75 days.

In addition, Internet domain names will sometimes disappear or change ownership. This means that at any one time, a proportion of all links on the Web will not work or (even worse) will link to the wrong site. This causes a particular problem with URL references in scientific research.

In its early stages, the Web was largely made up of hyperlinked HTML pages, sometimes with inline images stored separately in GIF or as files or using JPEG compression. To this was gradually added a number of other formats for text (e.g. TeX, PostScript or PDF) and multimedia. Many of these - especially multimedia formats - require the use of browser 'plugins.'

#### 3.3. The 'deep Web'

In 2001, Bar-Ilan (2001, p. 9) argued that all size estimates of the Web were misleading because they only tended to count "static pages, freely accessible to search engines and Web users." She pointed to the existence of a large number of other pages that were not so accessible; chiefly those created dynamically from databases, or with other accessibility barriers (e.g. password protection) or format problems. For example, increasing numbers of Web sites are now dynamically served e.g., through things like Microsoft's ASP. Others provide a 'front-end' to large databases, many of which actually predate the development of the Web itself. A much-cited white paper produced by the search company BrightPlanet (Bergman, 2001) estimated that this subset of the Web - known as the 'invisible,' 'hidden' or 'deep Web' - could be up to 400 to 500 times bigger than the so-called 'surface Web.'

BrightPlanet's survey attempted to identify some of the largest deep Web sites (Bergman, 2001). The largest public sites identified in 2001 were the National Climatic Data Center of the US National Oceanic and Atmospheric Administration (NOAA) and NASA's Earth Observing System Data and Information System (EOSDIS). Other large deep Web sites identified included the products of key digital library projects (e.g., the Informedia Digital

Video Library, the Alexandria Digital Library, the UC Berkeley Digital Library, JSTOR), bibliographic databases (e.g., PubMed), databases of scientific data (e.g. GenBank) and library catalogues. The largest fee-based sites included database services for law and business (Lexis-Nexis, Dun & Bradstreet, etc.), bibliographic databases (e.g., Ovid, INSPEC) and e-journal services provided by scholarly publishers (e.g., Elsevier, EBSCO, Springer-Verlag).

The deep Web poses a severe challenge to Web preservation initiatives, and in particular to those based on harvesting technology. Much database-driven Web information will be as invisible to Web harvesting robots as they are to the existing generation of search engines. Search companies (like Google) and computing science researchers are already investigating these issues from a Web search perspective (e.g., Raghavan & Garcia-Molina, 2001) but there is a need for more collaborative research into this issue from the perspective of Web archives.

**Recommendation 5:** Research: the current generation of harvesting technologies has limitations with regard to dealing with deep-Web sites. This has a particular impact on Web archiving approaches based on automatic harvesting. While some research is being carried out on this issue from a Web search perspective, there is a need for more collaborative research into this issue from the perspective of Web archives.

### 3.4. The future of the Web

It is perhaps also worth emphasising that the Web is a 'moving-target' for preservation initiatives. In the near future, there are likely to be changes as it evolves to take account of the W3C's vision of a 'Semantic Web,' whereby information is given well-defined meaning 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.

### 3.5. Defining and characterising the UK Web space

The size and nature of the UK Web is as difficult to quantify as the World Wide Web itself. Firstly, it is difficult to understand exactly what is meant by the "UK Web space". It could mean Web sites with a .uk domain name or sites physically hosted in the UK. The definition could also be extended to Web sites that belong to an UK organisation, sites in which the intellectual content belongs to an UK organisation or individual, or sites in which the intellectual content is of relevance to the UK. Whichever of these definitions is chosen will influence exactly which sites are included within the UK domain. For example, using the .uk domain will exclude UK Web sites that hosted in other domains, e.g., sites that use .com, .org, etc. Using the physical location of the Web server as the definition will exclude relevant sites hosted in other countries and may include sites that would not otherwise be considered to be part of the UK domain.

While acknowledging these problems, we can get some idea of the relative size of the UK Web by comparing national statistics derived from OCLC's ongoing Web Characterization Project. In 2002, these figures (Table 3.1) revealed that the US was by far the largest domain (55%), distantly followed by Germany (6%), Japan (5%), the UK, and Canada (both at 3%). The UK Web is, therefore, one of the larger national domains, but remains significantly smaller than that of the US.

If we take the .uk domain by itself, we can also begin to see how this breaks down into sub-domains by looking at the statistics in Table 3.2 that were provided by Netcraft (Table 3.2), a Web server monitoring company (<http://www.netcraft.com>). These reveal that co.uk is the largest UK sub-domain, containing over 90% of Web sites in the .uk domain. The sub-

domains largely used by schools (`sch.uk`), research and FE/HE institutions (`ac.uk`), and the National Health Service (`nhs.uk`) are much smaller.

**Table 3.1: Web characterisation: country of origin, 2002**

| Country     | Percent of Public Sites |
|-------------|-------------------------|
| US          | 55%                     |
| Germany     | 6%                      |
| Japan       | 5%                      |
| UK          | 3%                      |
| Canada      | 3%                      |
| Italy       | 2%                      |
| France      | 2%                      |
| Netherlands | 2%                      |
| Others      | 18%                     |
| Unknown     | 4%                      |

Source: OCLC Research (<http://wcp.oclc.org/stats/intnl.html>)

**Table 3.2: Numbers of Web sites in the .uk domain, March 2002**

| Domain       | Total            |
|--------------|------------------|
| .co.uk       | 2,750,706        |
| .org.uk      | 170,172          |
| .sch.uk      | 16,852           |
| .ac.uk       | 14,124           |
| .ltd.uk      | 8,527            |
| .gov.uk      | 2,157            |
| .net.uk      | 580              |
| .plc.uk      | 570              |
| .nhs.uk      | 215              |
| ...          |                  |
| <b>Total</b> | <b>2,964,056</b> |

Source: Based on figures supplied by Netcraft (<http://www.netcraft.com/>)

### 3.6. Conclusions

The Web is a very large resource that is still growing at a very rapid rate. The rate of growth and the amount of information that is 'hidden' from automated search tools, mean that it is very difficult to get accurate figures on its size. Comparative statistics from OCLC Research suggest that the UK is one of the larger national domains, although significantly smaller than the US domain. A closer look at the .uk domain shows that over 90% of almost three million Web servers are in the commercial (`co.uk`) sub-domain, while less than 0.5% of this makes up the `ac.uk` domain.

## 4. Review of existing Web archiving initiatives

Since the advent of the Web as a key global information resource in the mid-1990s, several organisations have attempted to deal with its preservation. Most of these initiatives are selective, covering particular country domains, subject areas, individual Web sites, etc.

Initiatives to date have taken one of two broad technical approaches. Firstly, there are initiatives that select and replicate Web sites on an individual basis, an approach exemplified by the National Library of Australia's PANDORA archive (<http://pandora.nla.gov.au/>) and by some projects developed by national archives. A second group of initiatives use crawler programs to automatically gather and store publicly available Web sites. The most ambitious of these initiatives is the Internet Archive (<http://www.archive.org/>), which has been taking periodic snapshots of the entire Web since 1996. Other crawler-based initiatives have focussed on national domains, e.g. the Swedish Royal Library's Kulturarw<sup>3</sup> project (<http://kulturarw3.kb.se/>) and the Austrian On-Line Archive (AOLA)

The remainder of this chapter will introduce and evaluate some well-known Web preservation initiatives. This is based on publicly available information and the responses to a questionnaire that was sent to selected initiatives (Appendix E).

Table 4.1 attempts to provide a summary of the main initiatives described here. It includes an indication of whether the initiative is based on automatic harvesting or gathering or on the selective approach pioneered by PANDORA. It also gives an indication on whether the archives are publicly available and to facilitate comparison gives an indication of scale and cost, where these figures are available.

**Table 4.1 Summary of Web Archiving Initiatives**

| Country   | Initiative/Lead Organisation     | Method <sup>1</sup> | Access <sup>2</sup> | Size <sup>3</sup> | Cost (£) <sup>4</sup> |
|-----------|----------------------------------|---------------------|---------------------|-------------------|-----------------------|
| Australia | PANDORA (NLA)                    | S                   | Y                   | 353 Gb.           | 360 k pa              |
| Austria   | AOLA (ONB/TU Wien)               | H                   | N                   | 448 Gb.           |                       |
| Finland   | Helsinki University Library      | H                   | N                   | 401 Gb.           |                       |
| France    | Bibliothèque nationale de France | H & S               | N                   | < 1 Tb.           | > 1 m pa              |
| Sweden    | Kulturarw <sup>3</sup> (KB)      | H                   | L                   | 4.5 Tb.           |                       |
| UK        | Britain on the Web (BL)          | S                   | N                   | 30 Mb.            | £600 k pa             |
| USA       | Internet Archive                 | H                   | Y                   | > 150 Tb.         |                       |
| USA       | MINERVA (LoC)                    | S                   | N                   | 35 sites          |                       |

**Key:**

1. Method is split into automatic harvesting (H) and selective (S) approaches.
2. Access indicates whether archives are publicly available (Y or N), or where access is limited (L), e.g. from the premises of the collecting organisation.
3. Size is approximate, and given in bytes. Where this figure is not available, the number of sites collected is given.
4. Costs (where available) are approximate and have been converted into Pounds Sterling using rates published in November 2002. The BL and BnF figures are future projections, based on the scaling-up of the pilot project to 10,000 sites (BL) and the setting-up of an initiative employing around 20 FTEs (BnF). These figures do NOT relate to the cost of the current archive.

#### 4.1. The Internet Archive

The Internet Archive (<http://www.archive.org/>) was one of the first attempts to collect the Web for future use. In 1997, Kahle (1997, p. 72) argued that the falling costs of digital storage meant "that a permanent record of the Web and the rest of the Internet can be preserved by a small group of technical professionals equipped with a modest complement of computer workstations and data storage devices."

The Internet Archive (<http://www.archive.org/>) started collecting Web pages in 1996. The archive itself was established as a not-for-profit organisation, but collected pages that had been crawled and analysed by a commercial company called Alexa Internet (<http://www.alexa.com/>), who were developing Web navigation tools. The size of this ongoing initiative is immense. As of the summer of 2002, the Internet Archive had collected Web pages totalling around 150 terabytes, making it one of the largest collections of data in the World. By way of comparison, the total printed holdings of the Library of Congress have been estimated to approximate to around 20 terabytes of text (Kahle, 1997).

In 2001, the Internet Archive made its collections available through the Web itself. A tool, called the 'Wayback Machine' (<http://web.archive.org/>), allows users to enter URLs and retrieve the over ten billion pages stored in the Internet Archive's collections. The database can also be made available to researchers, who would obtain a user account from the Internet Archive.

The Internet Archive has also created and hosts 'special collections' on specific topics. For example, they have created a small collection called 'Web Pioneers,' which uses the Wayback Machine to retrieve key sites from the early history of the Web. This includes a 1997 version of the 'Trojan Room Coffee Machine' page, a popular Web page hosted by the Computer Laboratory at the University of Cambridge that was one of the very first demonstrations of Web cams. The images have not been updated since August 2001, when the laboratory moved to new premises and the Web cam was retired (Stafford-Fraser, 2001). Other collections have been undertaken in collaboration with other organisations. These include the Election 1996 collection undertaken in association with the Smithsonian Institution and the 'Election 2000 Internet Library' that was (<http://web.archive.org/collections/e2k.html>) commissioned by the Library of Congress. More recently, the Internet Archive has helped create the 'September 11 Web Archive' (<http://september11.archive.org/>), in co-operation with the Library of Congress, WebArchivist.org and the Pew Internet & American Life Project.

The 'Web Pioneers' special collection explicitly uses the Wayback Machine to locate older versions of Web sites. The 'Election 2000 Internet Library' and the 'September 11 Web Archive' use broadly the same technologies, but the sites have been chosen for capture on a particular date. Access to the 'Election 2000' collection is via a directory of 797 sites, broadly grouped by category. Access to the 'September 11' collection is via a user interface designed by WebArchivist.org. The Pew Internet & American Life Project (2002) published a very detailed analysis of responses to September 11 on the Internet, and this contains many links to Web resources held by the Internet Archive. Some individual chapters also contain links to 'Webscapes' (collections of example Web sites) that are located on the 'September 11 Web Archive' server.

The Internet Archive is actively seeking co-operation with other organisations, including national libraries, so that it can focus its future activities on the requirements of partner organisations. This would initially be done by the setting-up of an Internet Archive Consortium.

*Collection:* The majority of the Internet Archive's holdings are received from Alexa Internet. This means that the collection has a broad focus across the Web rather than a narrow focus on any particular subset of it (Stata, 2002). This means that the collections will include sites of all levels of quality and in all subject areas, which may be valuable for future research.

The Internet Archive's crawler programs respect the Robots Exclusion Protocol and will not harvest Web sites or parts of Web sites that are protected by robots.txt files (<http://www.robotstxt.org/wc/exclusion.html>). The Internet Archive FAQ explains that if Web site owners do not want their past or present day sites included in the Internet Archive they should create a simple robots.txt file, and gives instructions on how this can be done. For the opposite case, the FAQs also explain how site owners can get their site included in the collection (<http://www.archive.org/about/faqs.php>).

Some dynamic sites can be successfully collected. Leonard DiFranza said that some are easily stored in the archive, while others fall apart completely.

*When a dynamic page renders standard html, the archive works beautifully.  
When a dynamic page contains forms, JavaScript, or other elements that  
require interaction with the originating host, the archive will not contain the  
original site's functionality.*

*Access:* The philosophy of the Internet Archive is that there cannot really be preservation without some kind of access (Stata, 2002). Kahle (2002b) has argued that "preservation without access is dangerous - there's no way of reviewing what's in there." Most current access to the Internet Archive's collections is through the Wayback Machine. Individual researchers, however, can get access to the 'raw' data by obtaining a user account on application to the Archive. An experimental search facility was being tested in November 2002.

*Hardware:* The Internet Archive is stored on a continually expanding cluster of interlinked PCs. These typically run on the FreeBSD operating system (or Linux), have 512 Mb. of memory and can hold over 300 Gb. of data on IDE disks.

*Software:* The crawler programs are rewritten periodically. Kahle (2002a) says that every 12 to 18 months "Alexa throws out its existing crawler and starts over from scratch." Files are stored in ARC files, and have some associated metadata (27 fields).

*Costs:* The Internet Archive is expensive to run because it has very high requirements for bandwidth and storage. Kahle (2002a), however, notes that it operates with a high level of frugality.

*Our technology has to be state-of-the-art to keep up. We collect 10 Tb per month. Our approach has to be cost-effective and flexible.*

The Archive has few full-time staff, in total about 30 people. Hardware costs have estimated at approximately \$US 3,000 per terabyte (Stata, 2002).

## **4.2. Selective approaches**

### **4.2.1. The National Library of Australia (PANDORA)**

One of the first national-based Web preservation initiatives was the National Library of Australia's PANDORA (Preserving and Accessing Networked Documentary Resources of Australia) archive (<http://pandora.nla.gov.au/>). Work on the project started at the beginning of 1997 with the development of a 'proof-of-concept' archive and a conceptual framework for a permanent service (Cameron & Pearce, 1998). Even before the PANDORA project itself was set up, the National Library of Australia (NLA) had begun to establish guidelines for the selection of online publications that it would seek to preserve.

The selective approach is pragmatic. Law (2001, p. 13) explains that the main reason for this is that "it is complex, time consuming and expensive to collect and archive digital publications; therefore the Library has chosen at this stage to concentrate resources on those publications considered to have current and future research value."

PANDORA is also a model of collaboration. The NLA works extensively in co-operation with its partner institutions - chiefly state libraries and ScreenSound Australia (the national screen and sound archive). The collaboration process itself is expensive in terms of effort and resources but remains important to the NLA.

*Collection:* PANDORA's selection guidelines are based on the idea that a "higher degree of selectivity is ... necessary for online publications than is the case with print" (<http://pandora.nla.gov.au/selectionguidelines.html>). General criteria include a 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.

Once sites have been selected and agreement secured with the site owners, they are collected using gathering software or the NLA makes arrangements with the publisher to receive the files on physical media or via ftp or email attachment. For robotic gathering, PANDORA initially used a modified version of the Harvest indexing software developed by the University of Colorado (Phillips, 1999), but more recently has adopted a twin approach using HTTrack and Teleport Pro. An evaluation undertaken by the NLA in 2000 concluded that at that time, there were few sites where a complete gathering would not be possible using one of these more recent packages (McPhillips, 2002). Using more than one gatherer program improves the chances of a successful replication of the Web site.

The frequency of capture depends on the Web site being considered for inclusion in PANDORA. Monographs only need to be captured once, while at least one serial is archived on a weekly basis. The depth of capture also depends on the Web site. If the Web site is very large, the NLA may collect only a part of it or select individual publications. External links are not collected.

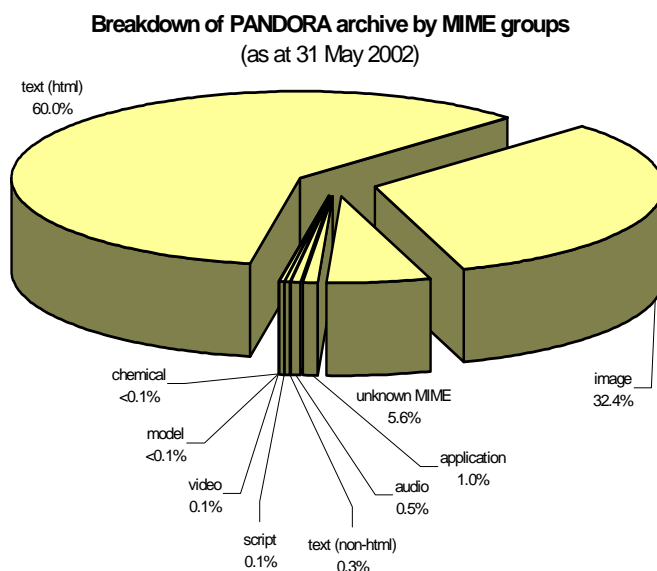
In January 2003, PANDORA contained 3,300 titles (representing 14,400,000 files and 405 Gb.). It is growing at an average rate of 21.5 Gb. per month, and around 125 titles per month.

*Composition of archive:* The NLA was able to provide detailed information about the composition of PANDORA:

| MIME groups     | No. Objects |
|-----------------|-------------|
| text (HTML)     | 2,373,025   |
| image           | 1,280,108   |
| unknown MIME    | 222,712     |
| application     | 40,924      |
| audio           | 19,047      |
| text (non-HTML) | 12,075      |
| script          | 4,787       |
| video           | 3,003       |
| model           | 119         |
| chemical        | 11          |

*Notes:* "text" is split into HTML (including CSS, XML) and non-HTML (plain text, RTF); "image" could be split into GIF (810,000) and JPEG (466,000) and other (4,108) other image items. There are also 28,360 PDFs.

Source: Gerard Clifton (Preservation Services, NLA), 2002



*Source: Gerard Clifton (Preservation Services, NLA), 2002*

**Management:** The NLA's IT Branch built an archive management system called PANDAS (PANDORA Digital Archiving System) This system manages the metadata collected about titles that have been selected (or rejected) for inclusion in PANDORA. It also initiates the gathering process, undertakes quality control and manages access to a gathered resource. All titles in PANDORA have MARC catalogue records that are included in the National Bibliographic Database and the local catalogues of PANDORA partners.

Preservation and access master copies are stored on Unix file systems in a consolidated format on the Library's Digital Object Storage System.

**Access:** Access is available through the PANDORA Web site, and the sites available can be searched or browsed by 15 top level subject domains and an alphabetical title list.

Most of the resources available in PANDORA were (or still are) freely available on the Web. The NLA obtains permission from the resource owner to collect the resource and to make it publicly available. PANDORA also contains around 90 commercial publications. Access to these is restricted (except through partner library reading rooms) for a period of time nominated by the publisher.

**Software:** The initial technical approach used a mixture of proprietary hardware and software, open source tools and some applications developed by the library itself (Webb, 1999). This has since been supplemented by the in-house development of PANDAS, first implemented in 2001 and now in version 2.

**Costs:** The NLA receives no additional funding from government to undertake PANDORA so the library funds it entirely from its existing allocation. The partner institutions contribute staff resources ranging from 0.5 to 2 full time operational staff. Margaret Phillips reports that a costing exercise undertaken earlier in 2002 revealed that the NLA is spending in excess of AUSS\$1 million (ca. £360,000) per year to build the archive (excluding staff costs of partner agencies). The NLA's effort is spread between several sections of the library.

#### 4.2.2. The Library of Congress (Web Preservation Project, Minerva prototype)

In 2000, the Library of Congress (LoC) started working on a prototype system called Minerva (Mapping the Internet the Electronic Resources Virtual Archive) that would collect and preserve open-access Web resources (<http://www.loc.gov/minerva/>). The purpose of Minerva was to gain experience with the practical issues of collecting and maintaining selected Web sites and to inform the development of a larger-scale preservation programme. A pilot took a small number (35) of selected Web sites and attempted to gather snapshots of these using a mirroring program (HTTrack). The snapshots were then evaluated and catalogued using OCLC's CORC (Cooperative Online Resource Catalog) software. In parallel with this pilot, the Library of Congress also co-operated with the Internet Archive on the creation of the 'Election 2000 Internet Library.'

The outcome of these experiments was the development of recommendations for a larger-scale preservation programme. This concluded that the large size of the Web would mean that most processes would have to be automated, but carefully monitored by library staff. Some topics would require further study: notably selection criteria, legal issues (copyright) and cataloguing (Arms, 2001b).

*Access:* Access to the sites that were gathered by the Minerva pilot is restricted to staff of the Library of Congress. The 'Election 2000' and later collections are available through the Internet Archive.

*Costs:* Arms, *et al.*, (2001) estimate that the selective approach, as carried out in the Minerva pilot, is "at least 100 times as expensive as bulk collection" on the Internet Archive model.

#### 4.2.3. The British Library (Britain on the Web)

Britain on the Web (formerly domain.uk) is a small-scale project organised by the British Library (BL) from 2001 to 2002 (Bury, 2002; Duffy, 2002). A decision was taken to select 100 Web sites as a snapshot of UK Web activity. Selection was made by curatorial staff, and included sites that were newsworthy (the 2001 General Election, foot and mouth disease, etc.) and a sample based on the highest level Dewey Decimal Classification divisions - to ensure a wide coverage of subject areas. Excluded were Web sites produced by the government (on the assumption that the PRO would have a role in collecting these), publishers and those with specific technical problems (e.g., large size, password protection, copyright problems, etc.). Permission was sought from site owners, and then sites were captured using Blue Squirrel Web Whacker or HTTrack. Stephen Bury estimated that the BL eventually accumulated about 30 Mb. The capture process generated some metadata while URLs were also fed through the dc-dot Dublin Core metadata generator (<http://www.ukoln.ac.uk/metadata/dcdot/>). Sites were revisited periodically, in particular during the time of the General Election. There is currently no public access to the copies of the Web sites that were collected.

At the end of the pilot, it was recommended that the Britain on the Web project should be scaled up over the next few years, including a trial of harvesting the .uk domain (possibly using the NEDLIB harvester). Partnership with other organisations was seen as important in undertaking any selective Web-archiving alongside harvesting approaches. Suggested collaborations included the PRO on government Web sites, the RDN for the selection of Web sites of academic merit, and other national libraries. In 2002, the BL also applied for substantial additional government funding (ca. £600,000) in order to be able to scale up this initiative. This would allow something in the region of 10,000 resources to be selectively gathered and archived in addition to an annual snapshot using the harvesting approach. Compared with estimates of the whole UK Web (see section 3.5 and Appendix D), the selective archiving can only cover a relatively small percentage of the Web sites available.

The BL now plans to fund (from April 2003) a new Web archiving project that can be scaled-up if the proposal for additional funding from government is successful. The library is also keen for legal deposit to be extended to cover digital material, including some Internet-based publications. In 1998, the report of a Working Party on Legal Deposit set up by the Secretary

of State for Culture, Media and Sport noted the growing importance of the Web and recommended that at least part of it should be considered for deposit (Working Party on Legal Deposit, 1998).

*Most of this [the Web] would be out of scope of UK deposit arrangements, either because the material falls outside the definition of 'publication' (e.g. company and institutional sites, entertainment material), or is a foreign rather than UK publication. In addition much is ephemeral, or of little or no long term research significance. UK documents which do fall within the definition of 'publications' should be considered for deposit, with depositories having the right to select and retain only that material considered to be of relevance and significance for the national published archive.*

Following the publication of this report, a code of practice was drawn up for the voluntary deposit of non-print publications, which came into effect in January 2000. The BL has also led an ongoing campaign for the extension of legal deposit to non-print publications. One recent outcome of this is the recent attempt by Chris Mole, the Labour Member of Parliament for Ipswich, to introduce a Private Members Bill in Parliament that would extend legal deposit to non-print publications, including Web publications (e.g., British Library, 2003; Mole, 2003).

### 4.3. Harvesting-based approaches

#### 4.3.1. The Swedish Royal Library (Kulturarw<sup>3</sup>)

In 1996, the Swedish Royal Library (KB) initiated a Web archiving project called Kulturarw<sup>3</sup>. Like the Internet Archive, the project decided from the start to use an 'active' collection policy that was based on the deployment by KB of a harvesting robot (Arvidson & Lettenström, 1998). The harvester was first run in 1997, when it downloaded 6.8 million URLs from 15,700 Web sites. A more recent harvest (in 2001) retrieved 30 million objects from 126,000 sites (Arvidson, 2002). Over 90 per cent of the document types collected are either HTML pages or JPEG and GIF images.

*Collection:* The collection is done by a harvesting robot; a modified version of the Combine harvester developed by NetLab at Lund University Library as part of the DESIRE project (e.g., Ardö & Lundberg, 1998). This collects sites in the .se geographical domain and sites identified (using WHOIS) as being located in Sweden. In addition, sites in the .nu domain (Niue Island) are targeted because they are widely used in Sweden. Arvidson (2002) estimates that around 50 per cent of the Swedish Web is not registered in the .se domain. The selection criteria are not written down in any detail, although there is a list of Swedish domains registered under other top-level domains. The harvester is run two to three times each year. The archive currently holds around 4.5 Tb, and is estimated to grow at about 2-3 Tb per year.

*Access:* At the start of the project, the KB did not plan any public access. The situation changed in May 2002 when the Swedish government, in response to a legal challenge, issued a special decree that authorised the KB to allow public access to the Kulturarw<sup>3</sup> collection within the library premises ([http://www.kb.se/Info/Pressmed/Arkiv/2002/020605\\_eng.htm](http://www.kb.se/Info/Pressmed/Arkiv/2002/020605_eng.htm)). Allan Arvidson reports that the only current access to the archive is a cgi-script that allows the surfing of the archive in time and 'space.'

*Management:* For metadata, Kulturarw<sup>3</sup> stores the HTTP headers that come with the object and automatically add metadata from the crawl: URL, time of acquisition, a checksum of content and the version of crawler used.

*Software:* Kulturarw<sup>3</sup> uses a modified version of the open-source Combine, home made software for maintenance and access, and a commercial hierarchical storage management (HSM) system.

*Hardware:* Harvesting runs on a Sun 450 server (4 cpus and 4 Gb. of memory) and the archive is maintained on a Sun 4500 server (4 cpus and 4 Gb. of memory). This is attached to a disk array with around 1.5 Tb disk space and an ADIC AML/J tape robot. This hardware is also used for other purposes by the library.

*Costs:* Kulturarw<sup>3</sup> is funded from the KB's ordinary budget. Staff effort is two full-time and one part time (0.2 FTE).

#### 4.3.2. Helsinki University Library (EVA)

In 1997, Helsinki University Library (the National Library of Finland) and its partners began a project to investigate the collection and preservation of documents published on the Finnish Internet (<http://www.lib.helsinki.fi/eva/english.html>). Like Kulturarw<sup>3</sup>, this project, called EVA, was one of the first to suggest the harvesting of all "freely available, published, static HTML-documents [together] with their inline material like pictures, video and audio clips, applets, etc." (Lounamaa & Salonharju, 1998). The project itself was limited to the .fi domain, although it was recognised that significant Finnish-related content existed elsewhere. The technical approach involved the development of a harvester robot that would capture pages and then analyse the content for link integrity. A snapshot made in March 1998 contained around 1.8 million documents from about 7,500 Web sites in the .fi domain.

#### 4.3.3. The NEDLIB Harvester

The NEDLIB (Networked European Deposit Library) project was a collaboration between eight European national libraries, one national archive, two IT organisations and three major publishers and was funded by the European Commission between 1997 to 2000. One of the tasks of the project was to develop and test a crawler program specifically designed for harvesting Web data. The NEDLIB partners produced a specification and the NEDLIB harvester was developed by the Finnish IT Center for Science (CSC). The harvester consists of a number of interrelated daemons and works with the MySQL relational database management system (Hakala, 2001c). The software undertakes some duplicate control and compression and can do incremental harvesting, downloading only documents that have been modified or newly created. The source code of the software is freely available from CSC (<http://www.csc.fi/sovellus/nedlib/>)

The NEDLIB harvester was first tested on the Icelandic Web. In January 2001, the crawler covered all registered Icelandic domains and downloaded over 565,169 documents from 1,426,371 URLs (Hakala, 2001a). This test (which produced 4.4 Gb. of data once compressed) demonstrated that a modest workstation would be sufficient for harvesting and storing many smaller national domains. There was a large amount of duplication (identified by using MD5 checksum calculations), possibly explained by the fact that Web servers often have more than one name.

The Finnish Web was harvested in June 2002, including the whole .fi domain and additional servers identified in other domains (Hakala, 2002). On completion, this consisted of 11.7 million files, originating from 42 million locations. In compressed form, the collection produced 401 Gb. of data. As with the Swedish project, the vast majority of files collected (over 96%) were in a very limited number of formats: HTML (48%), GIF (25%), JPEG (20%) and PDF (3%). The maintenance of the collection was outsourced to CSC, where it is "stored on their tape robot, alongside with the Finnish climate data and other stuff which will be preserved for a long time" (Hakala, 2002).

Tests with the NEDLIB harvester raised a number of problems. Hakala (2001a) explains that most were related to bad data or poorly developed HTTP server applications. There was a specific problem with some CGI scripts, which sometimes set up endless loops.

#### 4.3.4. The Austrian On-Line Archive (AOLA)

The Austrian On-Line Archive (AOLA) is an initiative of the Austrian National Library (*Österreichische Nationalbibliothek*) and the Department of Software Technology and Interactive Systems at the Vienna University of Technology. As with the Nordic projects, AOLA also adopts a crawler-based collection strategy. The scope of the project covers the entire `.at` domain, servers located in Austria but registered under other domains and sites dedicated to topics of Austrian interest. A first, experimental, crawl was made in May 2001 using the NEDLIB harvester, but this was aborted after it hit problems (Aschenbrenner, 2001, pp. 62-64). For a second test run in June 2001, AOLA used the Combine harvester (as used by the Swedish Kulturarw<sup>3</sup> project). The data from these experimental crawls and a second one carried out in spring 2002 consists of about 488 Gb. of data, which the AOLA project team have since investigated various ways of analysing (Rauber, Aschenbrenner & Witvoet, 2002; Rauber, et al., 2002).

*Collection:* The selection criteria for AOLA is defined as anything that is deemed to be part of the Austrian cultural heritage, i.e. the `.at` domain, Web servers based in Austria and sites in other domains with Austrian content. Collection frequency is irregular, but approximately once a year on a project basis, not necessarily covering the whole Austrian Web space. The National Library separately receives some Web resources directly from publishers, but this is not part of AOLA. Combining the selective and harvesting approaches is being considered for the future.

*Management:* Some simple metadata is captured as part of the collection process.

*Software:* Combine, as adapted by Kulturarw<sup>3</sup> and AOLA.

### 4.4. Combined approaches

#### 4.4.1. The Bibliothèque nationale de France

The French national library, the *Bibliothèque nationale de France* (BnF), has investigated the preservation of the French Web as part of its responsibilities for the legal deposit of French publications. From 2001, some experiments with collecting the Web were conducted by the BnF in association with INRIA (the French National Institute for Research in Computer Science and Automatic Control).

*Collection:* The BnF initiative makes several assumptions. Firstly, that while legal deposit (e.g. for publications in print or on physical media) would traditionally use a 'push' model, it was assumed that the size and nature of the Web would mean that its collection would require some level of automatic processing. Secondly, that the BnF's main role would be to focus on the 'French Web.' However, as with other initiatives based on crawler technologies, it was acknowledged that it is very difficult to define exactly what this means. Suggested criteria include language, the domain name or the site's physical location, but none of these are completely satisfactory. In the experiments, the perimeter selected was sites within the `.fr` domain, starting with sites selected by BnF librarians (Abiteboul, *et al.*, 2002b).

Web crawling was based on techniques developed at INRIA for Xyleme, a project concerned with creation of large dynamic data warehouses of XML data (<http://www.xyleme.com/>). Like other robots, the Xyleme crawler is designed to retrieve HTML and XML pages from the Web to a place where they can be monitored and stored. A key feature of Xyleme is that certain pages can be refreshed regularly. Decisions about refresh frequency can be guided by user specification or by an estimate of change frequency or page importance (Abiteboul, *et al.*, 2002a).

Abiteboul (2002) has argued elsewhere that the size of the Web is probably a lesser problem for crawler-based Web preservation initiatives than the limitations of network bandwidth. The amount of time that it takes for the current generation of crawlers to read and download the

surface Web would mean that many pages would be out-of-date by the time that the process ended. Many Web sites, e.g., those provided by newspapers, change on a daily basis. Abiteboul suggests that a solution would be to use bandwidth more intelligently, for example, to take automatic account of a Web page's change frequency or importance.

- *Change frequency.* The experiments with collecting the French Web collected information about when each page was retrieved and whether there had been any updating. This information was stored in an XML-based 'site-delta' that could be used to judge the required frequency of crawl (Abiteboul, *et al.*, 2002b, pp. 12-13).
- *Page importance.* Page importance can be calculated in a similar way to that used by search services like Google. The original PageRank algorithm used by Google assumed that a page was important if it was pointed to by many pages, or if it was linked to by pages that themselves had a high PageRank (Brin & Page, 1998). The INRIA and Xyleme research teams developed an algorithm for computing Web page importance without the need to download and store beforehand the whole graph of the Web (Abiteboul, Preda & Cobèna, 2002).

For the BnF, the automatic calculation of page importance provides one way of identifying that subset of the Web that needs to be preserved. Masanès (2002b) has noted that the low-cost nature of Internet publishing has broadened the notion of what constitutes 'publishing.' The BnF, therefore, regards the calculation of importance based on link structures as a way of focusing attention on the part of the Web that is most well-used. Masanès (2002b) says that if "making something 'publicly available' on the net becomes insignificant in and of itself, let's try to refine and focus on some part of it - the most 'public' one, which in a hypertext space is the most linked one." An initial evaluation, comparing a sample of automated rankings with evaluations of site relevance by library staff, showed a good degree of correlation (Masanès, 2002b).

Despite a strong focus on automatic processing, the BnF initiative also acknowledges the importance of human input into the collection process (Abiteboul, *et al.*, 2002b, p. 10).

*... librarians are vital in order to "correct" errors by positive action (e.g., forcing a frequent crawl of 00h00.com) or negative one (e.g., blocking the crawl of www.microsoft.fr). Furthermore, librarians are also vital to correct the somewhat brutal nature of the construction of the archive.*

Human input is also important with regard to the identification, selection and collection of 'deep Web' sites. The BnF have proposed a strategy based on the evaluation of sites by library staff and then liaising with the site owners over their deposit into the national library (the deposit track). While recognising that this approach could be problematic - both organisationally and technically - the BnF had undertaken a pilot project in 2002. In this, over 100 selected Web site owners were contacted and asked whether they would be prepared to deposit their sites with the BnF. Of these, about 50 signed the agreement, while only 34 actually made a deposit. Sites were delivered either on physical media (e.g., CD or DVD) or through FTP. The library can then undertake the potentially time-consuming tasks of validation and adding metadata (Masanès, 2002a). In the longer term, the library will also be responsible for the preservation (e.g. through migration) of the databases that form the basis of deep-Web sites. This is likely to be an even more time-consuming task.

The BnF notes that Web robots or harvesting technologies may have some uses in helping to support the deposit track. For example, robots could be used to analyse the technical features of crawled material, helping to detect deep-Web sites for which deposit may be required.

*Software:* The BnF have experimented with both of the Combine and NEDLIB harvesters, but these do not fulfil requirements for continuous archiving. For smaller scale site harvesting, BnF uses HTTrack driven by home made scripts with a MySQL database of URLs.

*Costs:* The BnF's Web archiving initiative is funded from the library's own budget.

#### 4.5. Other initiatives

This section has described many of the Web preservation initiatives that have been initiated by or collaborate with national libraries. There are others. For example, The National Library of Denmark has been operating a 'push' model for publishers to deposit online publications for some years. Since 2001, the Danish national library has also been involved in a project called *netarchive.dk* that tested different archiving strategies and the usability of the resulting collections for research (<http://www.netarchive.dk/>). Although not part of PANDORA, the State Library of Tasmania has itself collected Tasmanian Web sites as part of the 'Our Digital Island' initiative (<http://odi.statelibrary.tas.gov.au/>). The National Library of New Zealand is also beginning to experiment with Web archiving, initially based on the capture of selected Web sites using HTTrack.

National archives also have begun to get involved in the collection and preservation of Web sites, especially where Web sites are believed to contain public records. The National Archives of Australia (2001a; 2001b) and Public Record Office (2001) have issued detailed electronic records management (ERM) guidelines for Web site managers. Other national archives have begun to collect 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 in January 2001 for deposit with the Electronic and Special Media Records Services Division (ESMRSD). The date was chosen so that NARA could document (at least in part) agency use of the Internet at the end of the Clinton Administration (Bellardo, 2001). In association with this project, NARA issued some technical guidance on how agency staff should prepare the snapshot. Once prepared, the snapshot was sent to NARA on physical media (3480-class tape cartridges, 9-track tapes, CD-ROM or DLT) accompanied by technical documentation and a Web site description form. As with the library-based projects, NARA's initiative was mostly concerned with Web site capture and transfer to the ESMRSD. While NARA has the ultimate responsibility to provide for preservation and access in the long term, the Deputy Archivist of the United States acknowledged that "NARA does not have the capability at this time to take or preserve all of the types of agency Web records" (Bellardo, 2001). Sprehe (2001) described the NARA snapshot initiative as bearing the marks of great haste. He argued that preparing Web site snapshots was expensive and not essential for recordkeeping purposes. In any case, federal agencies were given very little time to respond to the request.

As well as having responsibility for the electronic records of federal departments and agencies, NARA also is responsible for the official records of national research centres. Deken (2002) has described some experiences as archivist of the Stanford Linear Accelerator Center (SLAC). This centre (<http://www.slac.stanford.edu/>) played an important part in the early history of the Web. On the 12 December 1991, <http://slacvm.slac.stanford.edu/> became the first Web server to go live in the US (Gillies & Cailliau, 2000, p. 219). Deken described the process of documenting the early phases of the Web site, revealing that some individuals had retained paper documentation in their own files. These included "copies of e-mail messages that were exchanged about setting up the first server, getting it to work properly, and negotiating changes and additions to the SLAC home page" (Deken, 2002). Attempts were also made to recreate earlier versions of the Web site from retained backups, but this was a challenging and time-consuming process. Deken was also concerned that transferring Web sites in the physical form required by NARA would lead to loss of functionality, e.g. with regard to external hyperlinks.

In the UK, the Public Record Office (PRO) transferred a snapshot of the No. 10 Downing Street Web site (<http://www.number-10.gov.uk/>) as it existed just before the General Election of June 2001. This was the first government Web site to be transferred to the national archive. Ryan (2002) has described some of the technical problems that the PRO had with transferring the Web site, as it had to be made to work in a different technical environment. The snapshot has been given the code PREM/18/1 and is freely available on the Web (<http://www.records.pro.gov.uk/documents/prem/18/1/default.asp>).

## 4.6. Evaluation

Broadly speaking, to date, three main approaches to collecting Web sites have been used:

- *Automatic harvesting*: the Internet Archive and a number of national initiatives use an automated approach based on crawler technology.
- *Selection*: the NLA and some other national libraries and archives adopt a selective approach that use mirroring-type tools to replicate complete Web sites periodically (a 'pull' model).
- *Deposit*: some national libraries and archives request site owners or publishers to deposit Web sites (or individual files) on physical media or by FTP (a 'push' model).

Less work has been undertaken on the related challenges of long-term preservation and access.

Specific software tools have been developed or adapted to support both collecting approaches. The Swedish Royal Library's Kulturarw<sup>3</sup> initiative adapted the Combine crawler, while other countries have used or evaluated the NEDLIB harvester developed by the Finnish CSC. 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 with Harvest, but currently has adopted a twin approach, using HTTrack and Teleport Pro/Exec. The British Library, the Library of Congress and the BnF 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.

It is difficult to fully evaluate the two main approaches to collecting the Web. Supporters of the automated crawler-based approach argue that it is a relatively cheap way to collect Web content, especially when compared with the selective approach. Thus Mannerheim (2001, p. 6) says that paradoxically, "it is a fact that the selective projects use more staff than the comprehensive ones." However, existing Web crawler technology cannot deal with many database-driven sites and can run into difficulty with items that need plug-ins or use scripting techniques. The selective approach allows more time to address and rectify these problems but severely limits the range of resources that can be collected.

The harvesting approach has proved its usefulness in the various Nordic projects and the Internet Archive. It would appear to be a particularly useful approach when a domain is relatively small and easy to identify, in particular where sites are static pages linked by standard HTML hyperlinks. The harvesting approach is much cheaper than the more labour-intensive approach adopted for PANDORA and similar initiatives. However, the changing nature of the Web means that the harvesting approach will over time become less effective unless crawlers can begin to deal adequately with active databases, software plug-ins (like Flash) and scripting techniques like JavaScript. The automated approach also only tends to deal with those parts of the Web that are publicly available. For 'deep Web' sites, the more labour-intensive selective approach may be the only way to realistically preserve them.

In addition it should be recognised that a significant element of the additional cost of the selective approach is occurred in rights clearance. Although this incurs additional cost it does allow most materials gathered in this way (for example in PANDORA), to be publicly accessible from the archive via the Web. This generates substantially higher use and gives wider accessibility than other methods. It also generates significantly lower legal risks (in the absence of any legal exceptions).

For the reasons noted above, initiatives are increasingly emphasising the need to consider using a combined approach of harvesting and selection to utilise the relative strengths of both approaches and address their limitations.

## **4.7. Other issues**

### **4.7.1. Collection policies**

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 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 (<http://www.robotstxt.org/wc/exclusion.html>). 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 and 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.

### **4.7.2. User access**

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) has investigated the issue of access to collections of Web documents (Bryggfeld, 2002). The result is 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/>).

### **4.7.3. Authenticity and trusted digital repositories**

As suggested before, many current Web archiving initiatives are largely focused on the collection of resources rather than on preservation techniques. In the short-term, there is nothing wrong with this, but there remains a longer-term need to consider how those Web sites being collected at the moment should be preserved over time. This may include assessments of various 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)* (ISO 14721:2002; CCSDS 650.0-B-1, 2002). 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 (Lynch, 2000, pp. 39-40). 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 (2000, p. 20) has said that "the fact that digital information is found within a trusted repository may become the base upon which all further assessment of action builds."

Ways of defining trusted repositories have recently been investigated by a working group established by the 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" (OCLC/RLG Working Group, 2002, p. 5). 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 become firmly embedded into the core activities of their host institutions.

#### **4.8. Conclusions**

It is hoped that this review of major 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 real implementations. 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 normally depends 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.

## 5. Implementation

This report has looked at several different approaches to Web archiving. To date, most projects have either chosen an approach based on automatic harvesting, or one based on the selective replication of whole Web sites. It is clear that the best Web preservation strategies depend upon a combination of these approaches. Arms (2001a) argues that comprehensive strategies "will combine bulk collection and selective collection. Technically both are sensible approaches. For future scholars, selective collection offers the benefits of depth whereas bulk collection provides breadth of coverage at much lower cost." The consideration of legal issues by Andrew Charlesworth and consideration of the research value of records created by different approaches (e.g., the Internet Archive evaluation in Appendices A and B) suggests that the selective approach would be best suitable for the needs of JISC and Wellcome Trust.

This report has, therefore, recommended that both the Wellcome Trust and the JISC set up a pilot project together or including other partners that uses the selective approach. These may in time, lead to more permanent initiatives, and their development should help inform the development of more detailed cost models.

Specifically, this report recommends that these pilot projects license the use of and test the PANDAS (PANDORA Digital Archiving System) software developed by the National Library of Australia. This would provide the benefit of tried and tested software for management of and access to the archive.

The remainder of this chapter looks at various issues relating to the implementation of the pilot phase of these projects.

### 5.1. Cost estimates

The exact costs of these pilot projects are hard to estimate and would depend on their scale and duration. The selective approach is far more labour intensive than the automated approach. Large initiatives like PANDORA have five FTEs, with additional IT support. Others may need to be bigger. The Library of Congress's Web Preservation Project noted that collecting and preserving the Web would be a substantial undertaking, "not worth beginning without a dedicated team of librarians and technical staff, about ten people initially" (Arms, 2001a). Of course, a pilot project would not necessarily need to be of this scale (note that the LoC project was investigating the collection of 30,000 Web sites). The exact number of FTEs required would depend on the exact scope of any pilot, but staff would be made up of chiefly of library and information professionals or archivists, supplemented by staff with IT and legal expertise.

Web archiving projects run on a wide variety of hardware and under several different operating systems. As we recommend the use and evaluation of PANDAS, we include here part of a technical overview of that system reproduced from a NLA document entitled: *Availability of the digital archiving software developed by the National Library of Australia.*

#### 5.1.1. Technical requirements of PANDAS

*The PANDAS user interface is web-based. It requires Internet Explorer 5.5 or higher. Collection managers will also require a range of web browser plug-ins and associated software to view publications being archived. PANDAS also uses WebDAV (web-based distributed authoring and versioning software) which ships with Windows 2000 and is an installable option for some earlier releases of Windows.*

*PANDAS is built around existing tools wherever possible. It consists of:*

- *a workflow/management system written in Java using the WebObjects application framework;*

- a metadata repository using Oracle 8i RDMS;
- a web-based gathering service based on a freely available Web site offline browser and mirroring tool called HTTrack <http://www.httrack.com/index.php> ; and
- a reporting facility using Microsoft Access Reports.

The public delivery system (PANDORA's Lid) is also built using Apache, WebObjects, Java and Oracle to provide resource discovery, navigation, and access control services. The actual items of digital content are delivered as static content through Apache. This service is hosted on a Sun Solaris server. The search service is provided by the mnoGoSearch (former UDMSearch) web search engine software which is freely available under the GNU Public Licence. <http://search.mnogo.ru/>

### **WebObjects**

A commercial Web Application Framework, WebObjects from Apple was chosen to support the development and web deployment of this system. This can be purchased from Apple for a cost of \$US699. See <http://www.apple.com/webobjects/> for details.

### **Operating System requirements**

It is possible to run all of the subsystems on the same server. However, the Library has chosen to run the pgather, pdisplay and psearch systems on separate Linux servers running Red Hat Linux 7.2/7.3 (<http://www.redhat.com>). The PANDAS workflow system runs on a Solaris 8 system. Although Solaris/MacOS X/Windows 2000 are the only platforms officially supported by WebObjects, the Library has had no problem running it on the Red Hat Linux platform.

### **Other Software Requirements**

To deploy all of the aspects of the system the following free software is also required:

- Apache 1.3+ (<http://www.apache.org>) with some additional modules `mod_speling` and `mod_rewrite`
- Perl 5+ (<http://www.cpan.org>)
- JDK 1.4+ (<http://java.sun.com>)

### **Database requirements**

This software has been developed using an Oracle back-end database (version 8 or 9). Although it is theoretically possible to use another database back-end (JDBC is used by the system to connect to the database) there may need to be some changes if certain features are not supported. Many of the perl scripts used rely on the Perl DBI Oracle module and would require a significant re-working if the database does not feature multi-nested SQL queries.

## **5.2. Staff skills**

Most Web archiving initiatives are supported by a mixture of IT specialists and staff from the library or archives professions. The exact mix of these skills depends upon the nature of the project. Harvesting-based initiatives tend to be smaller in scale and the technical nature of harvesting means that the majority of staff will tend to have an IT background. Selective projects need a more diverse range of staff skills to succeed. Taking a basic task plan, we can see the following responsibilities:

- *Selection.* This requires the initial development of a collection policy and then its implementation. This would normally be the task of information professionals.
- *Negotiation.* The owner of each selected site selected would then need to be contacted and asked to agree to its inclusion in the archive and asked about access restrictions. The initial drawing up of deposit agreements would need to be done by legal specialists. Once this has been done, the actual task of negotiation could be delegated to other staff although some continued legal input would be valuable. The experiences of some initiatives suggest that this stage can be extremely time-consuming.
- *Collection.* Once agreement has been obtained, the initiative's staff can begin to use software to archive the site. This process, which has been automated in PANDAS, includes some kind of validation and the capture or creation of metadata. The software tools used would need to be developed (or adapted) by IT specialists, while the day-to-day operation of the collection process could be delegated to information professionals.
- *Preservation.* Long term preservation may require the more detailed attention of IT specialists. While information or archives professionals could advise on what content needs to be retained and on technical preservation strategies, IT specialists would be required to undertake periodic migrations, unless tools can be developed to support this process.

Looking at these, the majority of the tasks of a Web archive could be undertaken by library and information professionals or by archivists. However, there is an additional need for continued technical support (including the production and revision of software tools) and legal advice.

To give a general idea of the mature set-up at the National Library of Australia; Web archiving (i.e. collection and ingest) is done within the Technical Services Branch of the Library and all five staff are qualified librarians (although this is not a requirement). The NLA's initiative also has the services of a person with an IT background for 20 hours a week to solve some of the trickier technical problems encountered.

*Staff are selected to work in the Electronic Unit for their high level of tech services skills, aptitude for and interest in things digital, their initiative, enthusiasm. They learn the work on the job. We do find that it is a longer and steeper training curve than other tech services work. (e-mail from Margaret Phillips, 21 November 2002).*

A separate team of 2 FTE preservation staff with some support from the IT section investigates long-term preservation issues and reports to a professional preservation manager.

### 5.3. Collaboration

While the collection and preservation of online publications and Web sites in the UK domain is not covered by voluntary legal deposit, Web sites remain of interest to the British Library, the other copyright libraries, and the national archives. The UK Web domain is very large (although small in comparison with the US domain) and selective archiving is by nature intensive and requires specialist subject knowledge. It could not be done for the whole of the UK by any single institution acting in isolation. It is ideally suited therefore to collaborative approaches. In addition to JISC and Wellcome Library, other partners in the Digital Preservation Coalition have expressed interest in selective archiving and participating in potential joint projects. It would be worthwhile to encourage the co-ordination and collaboration of activity in the UK and to explore options for joint activity.

Areas of possible interest for collaboration might cover:

- Strategic issues - e.g., lobbying for changes in copyright law and other relevant legislation, the joint funding of projects and/or services, working together in externally-funded projects, liaising with Web masters and content owners, etc.
- Technical issues - e.g., developing software and supporting its implementation, detailed research into specific problem areas (database-driven sites, the use of software 'plug-ins'), interoperability and standards development, etc.
- Organisational issues - e.g., training, sharing of expertise, etc.
- Content issues - e.g., the division of responsibilities for different content types based on provenance, subject matter, etc., the relative roles of archives, libraries and other types of organisation, collaboration on developing selection criteria, metadata sharing (e.g. with subject gateways), etc.

The type of collaboration required would depend on the nature of the Web archiving initiatives proposed. Some initiatives will be primarily driven by institutional interest (e.g., the collection of government Intranets by national archives, project Web sites by research funding bodies, the content of e-print repositories by research libraries, etc.). Here any collaboration will tend to be focused on strategic, technical or organisational issues. For example, a pilot devoted to the collection of JISC project Web sites may require some collaboration over the implementation of the PANDAS software, some additional research into specific problem areas, the development of standards and (possibly) some shared infrastructure.

Other initiatives may have a much wider focus, e.g. on building national or subject collections. As with institutional initiatives there may be a need for collaboration on strategic, technical and organisational issues, but there will be an additional need to deal with content issues, e.g. collaborative content building and the shared development of selection criteria. This may include some liaison with organisations not primarily involved in Web archiving, e.g. subject gateways like the RDN.

Other forms of collaboration might be possible, especially with national initiatives. Initiatives could share effort by one taking responsibility for selection and capture and another undertaking its long-term preservation. There is scope for this type of collaboration between the JISC and Wellcome Trust initiatives proposed by this report and any future British Library Web archiving initiative. This type of collaboration would have to be negotiated in detail with respective responsibilities agreed. Other potential collaboration partners could be professional data repositories (e.g., the National Data Repository at ULCC), data archives and the e-Science centres who have long-experience of managing large volumes of data.

The Web is also a global phenomenon. As we have seen, many attempts are being made to collect and preserve it on a national or domain level, e.g. by national libraries and archives. This means that no one single initiative (with the exception of the Internet Archive) can hope for total coverage of the Web. Close collaboration between different Web archiving initiatives, therefore, will be extremely important, e.g. to avoid unnecessary duplication in coverage or to share in the development of tools, guidelines, etc. Examples of international collaboration in this area include the proposed Internet Archive Consortium, a way for the Internet Archive to collaborate more deeply with national libraries, and the European Web Archive proposal for the European Union's Sixth Framework Programme (FP6).

**Recommendation 6:** Collaboration - for both the JISC and the Wellcome Trust there is significant opportunity for partnership on Web-archiving. For example, there will be opportunities to collaborate on strategic, technical, organisational or content issues.

For the UK, both should attempt to work closely with the British Library, the other copyright libraries, the Public Record Office, data archives and the e-Science centres that have

experience of managing large volumes of data. The focus for this collaborative activity could be within the Digital Preservation Coalition (DPC). On an international level, close co-operation with institutions like the US National Library of Medicine and the Internet Archive will be important.

As an exemplar of collaboration, it is recommended that the JISC and the Wellcome Library should seek to work together and with other partners to create their pilot Web archiving services. Not only will this realise economies of scale, but more importantly provide a model demonstrating how collaboration can work in practice.

#### **5.4. Sustainability**

Collaboration will eventually be the key to the long-term sustainability of Web collection and preservation initiatives. To date, the various initiatives have made a case for Web archiving and demonstrated several different technical approaches to its collection and preservation. In the future, these approaches will have to evolve in step with the rapid development of the Web itself, as it gears up for the inclusion of more semantic content, Web Services and the data glut promised by the upcoming generation of e-science projects.

The scale and nature of the existing Web means that it is beyond the control of any one single organisation or initiative. Even the broadest current initiative (the Internet Archive) only focuses on those parts of the Web that are publicly available. Even if all of the technical problems of collection can be solved, the global nature of the Web and the difficulty of defining 'national' domains make it difficult to know who exactly should be responsible for its preservation. National libraries will certainly have a major role, but there is room for many others: research libraries, archives, funding agencies, standards organisations, publishers, computer science departments, art galleries, etc.

Successful collaboration will be able to help reduce redundancy and duplication and enable the sharing of the technological tools that are essential for preservation. It may also be able to help with ensuring unified access to the content of Web archives. Managing an ongoing collaboration between these diverse types of organisation will not be easy but, as with other areas of the digital preservation agenda, will be the only way to ensure the long-term sustainability of Web archiving initiatives.

## 6. References

- Abiteboul, S. (2002). "Issues in monitoring Web data." In: Cicchetti, R., Hameurlain, A. & Traunmüller, R., eds., *Database and expert systems applications: the 13th International Conference on Database and Expert Systems Applications, September 2–6, 2002, Aix-en-Provence, France*. Lecture Notes in Computer Science, 2453. Berlin: Springer, 1-8. Also available at: <ftp://ftp.inria.fr/INRIA/Projects/verso/gemo/GemoReport-264.pdf>
- Abiteboul, S., Cluet, S., Ferran, G. & Rousset M.-C. (2002a). "The Xyleme project." *Computer Networks*, 39 (3), 225-238.
- Abiteboul, S., Cobéna, G., Masanès, J. & Sedrati, G. (2002b). "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, 2458. Berlin: Springer, 1-15. Also available at: <ftp://ftp.inria.fr/INRIA/Projects/verso/gemo/GemoReport-229.pdf>
- Abiteboul, S., Preda, M. & Cobéna, G. (2002). "Computing Web page importance without storing the graph of the Web (extended abstract)." *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, 25 (1), 27-33. Available at: <http://www.research.microsoft.com/research/db/debull/A02mar/issue.htm>
- Ammen, C. (2001). "MINERVA: Mapping the INternet Electronic Resources Virtual Archive -Web preservation at the Library of Congress." What's next for digital deposit libraries? ECDL Workshop, Darmstadt, Germany, 8 September 2001. Available at: <http://www.bnf.fr/pages/infopro/ecdl/usa/sld001.htm>
- Ardö, A. & Lundberg, S. (1998). "A regional distributed WWW search and indexing service - the DESIRE way." *Computer Networks and ISDN Systems*, 30 (1-7), 173-183. Also available at: <http://www7.scu.edu.au/programme/fullpapers/1900/com1900.htm>
- Arms, W.Y. (2001a). *Web Preservation Project: interim report*. Washington, D.C.: Library of Congress, 15 January. Available at: <http://www.loc.gov/minerva/webpresi.pdf>
- Arms, W.Y. (2001b). *Web Preservation Project: final report*. Washington, D.C.: Library of Congress, 3 September. Available at: <http://www.loc.gov/minerva/webpresf.pdf>
- Arms, W.Y., Adkins, R., Ammen, C. & Hayes, A. (2001). "Collecting and preserving the Web: the Minerva prototype." *RLG DigiNews*, 5 (2), 15 April. Available at: <http://www.rlg.org/preserv/diginews/diginews5-2.html#feature1>
- Arvidson, A. & Lettenström, F. (1998). "The Kulturarw<sup>3</sup> project: the Swedish Royal Web Archive." *The Electronic Library*, 16 (2), 105-108.
- Arvidson, A. & Persson, K. (2001). "Harvesting the Swedish Web space." What's next for digital deposit libraries? ECDL Workshop, Darmstadt, Germany, 8 September 2001. Available at: <http://www.bnf.fr/pages/infopro/ecdl/sweden/sld001.htm>
- Arvidson, A. (2002). "The collection of Swedish Web pages at the Royal Library: the Web heritage of Sweden." 68th IFLA Council and General Conference, Glasgow, UK, 18-24 August 2002. Available at: <http://www.ifla.org/IV/ifla68/papers/111-163e.pdf>
- Arvidson, A., Persson, K. & Mannerheim, J. (2000). "The Kulturarw<sup>3</sup> Project - the Royal Swedish Web Archiw<sup>3</sup>e: an example of "complete" collection of Web pages." 66th IFLA Council and General Conference, Jerusalem, Israel, 13-18 August 2000. Available at: <http://www.ifla.org/IV/ifla66/papers/154-157e.htm>

Arvidson, A., Persson, K. & Mannerheim, J. (2001). "The Royal Swedish Web Archive: a "complete" collection of Web pages." *International Preservation News*, 26, December, 10-12.

Aschenbrenner, A. (2001). *Long-term preservation of digital material - building an archive to preserve digital cultural heritage from the Internet*. Diplomarbeit, Technische Universität Wien, Institut für Softwaretechnik und Interaktive Systeme. Available at: <http://www.ifs.tuwien.ac.at/~aola/publications.html>

Bar-Ilan, J. (2001). "Data collection methods on the Web for informetric purposes: a review and analysis." *Scientometrics*, 50 (1), 7-32.

Bellardo, L.J. (2001). *Memorandum to Chief Information Officers: snapshot of public Web sites*. Washington, D.C.: National Archives & Records Administration, 12 January. [http://www.archives.gov/records\\_management/cio\\_link/memo\\_to\\_cios.html](http://www.archives.gov/records_management/cio_link/memo_to_cios.html)

Bergman, M.K. (2001). "The deep Web: surfacing hidden value." *Journal of Electronic Publishing*, 7 (1), August. Available at: <http://www.press.umich.edu/jep/07-01/bergman.html>

Bergmark, D. (2002). "Automatic collection building." 2nd ECDL Workshop on Web Archiving, Rome, Italy, 19 September 2002. Available at: <http://bibnum.bnf.fr/ecdl/2002/>

Berners-Lee, T. & Hendler, J. (2001). "Publishing on the Semantic Web." *Nature*, 410 (6832), 26 April, 1023-1024.

Berthon, H. (2002). "Nurturing our digital memory: digital archiving and preservation at the National Library of Australia." Paper submitted to *International Preservation News*. Available at: <http://www.nla.gov.au/nla/staffpaper/2002/berthon1.html>

Berthon, H., Thomas, S. & Webb, C. (2002). "Safekeeping: a cooperative approach to building a digital preservation resource." *D-Lib Magazine*, 8 (1), January. Available at: <http://www.dlib.org/dlib/january02/berthon/01berthon.html>

Bollacker, K.D., Lawrence, S. & Giles, C.L. (2000). "Discovering relevant scientific literature on the Web." *IEEE Intelligent Systems*, 15 (2), 42-47.

Boudrez, F. & Van den Eynde, S. (2002). *Archiving Websites*. Digitale Archivering in Vlaamse Instellingen en Diensten (DAVID) project report. Antwerp: Stadsarchief Stad Antwerpen; Leuven: Interdisciplinair Centrum voor Recht en Informatica, July. Available at: <http://www.antwerpen.be/david/>

Brin, S. & Page, L. (1998). "The anatomy of a large-scale hypertextual Web search engine." *Computer Networks and ISDN Systems*, 30 (1-7), 107-117. Full version published in the proceedings of the 7th International World Wide Web Conference, Brisbane, Australia, 14-18 April 1998. Available at: <http://www7.scu.edu.au/programme/fullpapers/1921/com1921.htm>

British Library (2003). "Support legal deposit extension." British Library Press Release, 29 Jan. Available at: <http://www.bl.uk/cgi-bin/news.cgi?story=1322>

Brygfjeld, S.A. (2002). "Access to Web archives: the Nordic Web Archive Access Project." 68th IFLA Council and General Conference, Glasgow, UK, 18-24 August 2002. Available at: <http://www.ifla.org/IV/ifla68/papers/090-163e.pdf>

Burkard, T. (2002). *Herodotus: a peer-to-peer Web archival system*. MEng. Thesis, Massachusetts Institute of Technology, Department of Electrical Engineering and Computer Science. Available at: <http://www.pdos.lcs.mit.edu/papers/chord:tburkard-meng.pdf>

Bury, S. (2002). "Domain.uk: interim report." London: British Library, March (internal report).

Cameron, J. & Pearce, J. (1998). "PANDORA at the crossroads: issues and future directions." In: *Sixth DELOS Workshop: Preservation of Digital Information, Tomar, Portugal, 17-19 June 1998*. Le Chesnay: ERCIM, 1998, 23-30. Available at: <http://www.ercim.org/publication/ws-proceedings/DELOS6/pandora.pdf>

Casey, C. (1998). "The cyberarchive: a look at the storage and preservation of Web sites." *College & Research Libraries*, 59 (4), 304-310.

Cathro, W., Webb, C. & Whiting, J. (2001). "Archiving the Web: the PANDORA archive at the National Library of Australia." *Preserving the Present for the Future Web Archiving Conference, Copenhagen, 18-19 June 2001*. Available at: <http://www.nla.gov.au/nla/staffpaper/2001/cathro3.html>

CCSDS 650.0-B-1. (2002). *Reference Model for an Open Archival Information System (OAIS)*. Blue Book, Issue 1. Washington, D.C.: Consultative Committee on Space Data Systems. Available at: <http://www.classic.ccsds.org/documents/pdf/CCSDS-650.0-B-1.pdf>

Chakrabarti, S., Dom, B.E., Kumar, S.R., Raghavan, P., Rajagopalan, S., Tomkins, A., Kleinberg, J. & Gibson, D. (1999a). "Hypersearching the Web." *Scientific American*, 280 (6), 44-52.

Cobèna, G., Abiteboul, S. & Preda, M. (2002). "Crawling important sites on the Web." 2nd ECDL Workshop on Web Archiving, Rome, Italy, 19 September 2002. Available at: <http://bibnum.bnf.fr/ecdl/2002/>

Cole, T.W., Kaczmarek, J., Marty, P.F., Prom, C.J., Sandore, B. & Shreeves, S. (2002). "Now that we've found the 'hidden Web,' what can we do with it?" *Museums and the Web 2002*, Boston, Mass., 17-20 April 2002. Available at: <http://www.archimuse.com/mw2002/papers/cole/cole.html>

Crook, E. (2002). "Libraries and erotica." *NLA Newsletter*, 12 (11), August. Available at: <http://www.nla.gov.au/pub/nlanews/2002/aug02/article3.html>

Cundiff, M. (2002). "METS and Websites." 2nd ECDL Workshop on Web Archiving, Rome, Italy, 19 September 2002. Available at: <http://bibnum.bnf.fr/ecdl/2002/>

Day, M. (2002). "2nd ECDL Workshop on Web Archiving." *Cultivate International*, 8, November. Available at: <http://www.cultivate-int.org/issue8/ecdlws2/>

Deken, J.M. (2002). "First in the Web, but where are the pieces?" arXiv: physics/0208059, 14 August. Available at: <http://www.arxiv.org/html/physics/0208059>

Dempsey, L. (2000). "The subject gateway: experiences and issues based on the emergence of the Resource Discovery Network." *Online Information Review*, 24 (1), 2000, 8-23. Available at: <http://www.rdn.ac.uk/publications/ior-2000-02-dempsey/>

Dickins, J. (2002). "Library's porn plan opposed." *Herald Sun* [Melbourne], 27 August. Available at: <http://www.heraldsun.news.com.au/printpage/0,5481,4975841,00.html>

Dollar Consulting. (2001). *Archival preservation of Smithsonian Web resources: strategies, principles, and best practices*. Washington, D.C.: Smithsonian Institution Archives, 20 July. Available at: <http://www.si.edu/archives/archives/dollar%20report.html>

Duffy, J. (2002). "The Internet, volume one." BBC News Online, 27 March. Available at: <http://news.bbc.co.uk/1/hi/uk/1896620.stm>

Europemedia.net (2002). "Entire French Web to be archived." europemedia.net, 22 June. Available at: <http://www.europemedia.net/shownews.asp?ArticleID=4075>

Ferran, P. (2002). "Smart Crawling, XML Storage and Query." 2nd ECDL Workshop on Web Archiving, Rome, Italy, 19 September 2002. Available at: <http://bibnum.bnf.fr/ecdl/2002/>

Foote, S. (2000). "Medical reference tools." *Journal of Library Administration*, 30 (3/4), 231-270.

Garrett, J. & Waters, D., eds. (1996). *Preserving digital information: report of the Task Force on Archiving of Digital Information*. Washington, D.C.: Commission on Preservation and Access, 1 May. Available at: <http://www.rlg.org/ArchTF/>

Gatenby, P. (2001). "Digital continuity: the role of the National Library of Australia." Digital Continuity: a Forum for Australian Universities, Swinburne University of Technology, 19 November 2001. Available at: <http://www.nla.gov.au/nla/staffpaper/2001/pgatenby4.html>

Gatenby, P. (2002). "Legal deposit, electronic publications and digital archiving: the National Library of Australia's experience." 68th IFLA Council and General Conference, Glasgow, UK, 18-24 August 2002. Available at: <http://www.ifla.org/IV/ifla68/papers/071-124e.pdf>

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

Hakala, J. (2001a). "The NEDLIB Harvester." *Zeitschrift für Bibliothekswesen und Bibliographie*, 48 (3-4), 211-216.

Hakala, J. (2001b). "Collecting and preserving the Web: developing and testing the NEDLIB Harvester." *RLG DigiNews*, 5 (2), 15 April. Available at: <http://www.rlg.org/preserv/diginews/diginews5-2.html#feature2>

Hakala, J. (2001c). "Harvesting the Finnish Web space - practical experiences." What's next for digital deposit libraries? ECDL Workshop, Darmstadt, Germany, 8 September 2001. Available at: <http://www.bnf.fr/pages/infopro/ecdl/finland/sld001.htm>

Hakala, J. (2002). "Harvesting of the Finnish Web space completed." E-mail sent to list: <web-archive@cru.fr>, 19 August.

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

Henriksen, B. (2001). "Danish legal deposit on the Internet: current solutions and approaches for the future." What's next for digital deposit libraries? ECDL Workshop, Darmstadt, Germany, 8 September 2001.

Henriksen, B. (2002). "The Danish project netarchive.dk" 2nd ECDL Workshop on Web Archiving, Rome, Italy, 19 September 2002. Available at: <http://bibnum.bnf.fr/ecdl/2002/>

Hirtle, P.B. (2000). "Archival authenticity in a digital age." In: *Authenticity in a digital environment*, Washington, D.C.: Council on Library and Information Resources, 8-23. Available at: <http://www.clir.org/pubs/abstract/pub92abst.html>

ISO 14721:2002. *Space data and information transfer systems -- Open archival information system -- Reference model*. ISO TC 20/SC 13. Geneva: International Organization for Standardization.

Jadad, A.R. & Gagliardi, A. (1998). "Rating health information on the Internet: navigating to knowledge or to Babel?" *JAMA*, 279 (8), 611-614.

Kahle, B. (1997). "Preserving the Internet." *Scientific American*, 276 (3), 72-73.

- Kahle, B. (2002a). "Editors' interview: the Internet Archive." *RLG DigiNews*, 6 (3), 15 June. Available at: <http://www.rlg.org/preserv/diginews/diginews6-3.html#interview>
- Kahle, B. (2002b). "Way back when ..." *New Scientist*, 176 (2370), 23 November, 46-49.
- Kelly, B. (2002). "Approaches to the preservation of Web sites." Online Information 2002, Olympia, London, 3-5 December 2002. Available at: <http://www.ukoln.ac.uk/web-focus/events/conferences/online-information-2002/paper.pdf>
- Kenney, A.R., McGovern, N.Y., Botticelli, P., Entlich, R., Lagoze, C. & Payette, S. (2002). "Preservation risk management for Web resources: virtual remote control in Cornell's Project Prism." *D-Lib Magazine*, 8 (1), January. Available at: <http://www.dlib.org/dlib/january02/kenney/01kenney.html>
- Kornblum, J. (2001). "Web-page database goes Wayback when." *USA Today*, 30 October. <http://www.usatoday.com/life/cyber/tech/2001/10/30/ebrief.htm>
- Law, C. (2001). "PANDORA: the Australian electronic heritage in a box." *International Preservation News*, 26, December, 13-17.
- Lawrence, S. & Giles, C.L. (1998). "Searching the World Wide Web." *Science*, 280, 98-100.
- Lawrence, S. & Giles, C.L. (1999a). "Searching the Web: general and scientific information access." *IEEE Communications Magazine*, 37 (1), 116-122.
- Lawrence, S. & Giles, C.L. (1999b). "Accessibility of information on the Web." *Nature*, 400, 107-109.
- 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), February, 26-31.
- Lazar, J., McClure, C.R. & Sprehe, J.T. (1998). *Solving electronic records management (ERM) issues for government Websites: policies, practices, and strategies: conference report on questionnaire and participant discussion, April 22, 1998, Washington, D.C.* Syracuse, N.Y.: Syracuse University, School of Information Studies. Available at: <http://istweb.syr.edu/~mcclure/ermreport5.htm>
- Leger, D. (2001). "Legal deposit and the Internet : reconciling two worlds." What's next for digital deposit libraries? ECDL Workshop, Darmstadt, Germany, 8 September 2001. Available at: <http://www.bnf.fr/pages/infopro/ecdl/quebec/sld001.htm>
- Liegmann, H. (2001). "Collection of German online resources by Die Deutsche Bibliothek." What's next for digital deposit libraries? ECDL Workshop, Darmstadt, Germany, 8 September 2001. Available at: <http://www.bnf.fr/pages/infopro/ecdl/germany/sld001.htm>
- Liegmann, H. (2002). "Submission and delivery interface for online publications" 2nd ECDL Workshop on Web Archiving, Rome, Italy, 19 September 2002. Available at: <http://bibnum.bnf.fr/ecdl/2002/>
- Lounamaa, K. & Salonharju, I. (1998). "EVA: the acquisition and archiving of electronic network publications in Finland." In: *Sixth DELOS Workshop: Preservation of Digital Information, Tomar, Portugal, 17-19 June 1998*. Le Chesnay: ERCIM, 1998, 31-34. Available at: <http://www.ercim.org/publication/ws-proceedings/DELOS6/eva.pdf>
- Lyman, P. & Varian, H.R. (2000). *How much information?* Berkeley, Calif.: University of California at Berkeley, School of Information Management and Systems. Available at: <http://www.sims.berkeley.edu/research/projects/how-much-info/internet.html>

- 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 and Library of Congress. Available at: <http://www.clir.org/pubs/abstract/pub106abst.html>
- Lynch, C. (2000). "Authenticity and integrity in the digital environment: an exploratory analysis of the central role of trust." In: *Authenticity in a digital environment*, Washington, D.C.: Council on Library and Information Resources, 32-50. Available at: <http://www.clir.org/pubs/abstract/pub92abst.html>
- Mannerheim, J. (2001). "The new preservation tasks of the library community." *International Preservation News*, 26, December, 5-9.
- Masanès, J. (2001). "The BnF's project for Web archiving." What's next for digital deposit libraries? ECDL Workshop, Darmstadt, Germany, 8 September 2001. Available at: <http://www.bnf.fr/pages/infopro/ecdl/france/slg001.htm>
- Masanès, J. (2002a). "Archiving the deep Web." 2nd ECDL Workshop on Web Archiving, Rome, Italy, 19 September 2002. Available at: <http://bibnum.bnf.fr/ecdl/2002/>
- Masanès, J. (2002b). "Towards continuous Web archiving: first results and an agenda for the future." *D-Lib Magazine*, 8 (12), December. Available at: <http://www.dlib.org/dlib/december02/masanes/12masanes.html>
- McClure, C.R. & Sprehe, J.T. (1998) *Analysis and development of model quality guidelines for electronic records management on State and Federal Websites: final report*. Syracuse, N.Y.: Syracuse University, School of Information Studies. Available at: [http://istweb.syr.edu/~mcclure/nhprc/nhprc\\_title.html](http://istweb.syr.edu/~mcclure/nhprc/nhprc_title.html)
- McClure, C.R. & Sprehe, J.T. (1998). *Guidelines for electronic records management on State and Federal Agency Websites*. Syracuse, N.Y.: Syracuse University, School of Information Studies. Available at: <http://istweb.syr.edu/~mcclure/guidelines.html>
- McPhillips, S.C. (2002). *PANDAS - a PANDORA project*. Canberra: National Library of Australia, 30 January.
- Mole, C. (2003). "MP in bid to capture electronic publications for future generations." Chris Mole MP Press Release, 6 Feb. Available at: [http://www.chrismolemp.org.uk/news\\_60203.htm](http://www.chrismolemp.org.uk/news_60203.htm)
- Le Monde. (2002). "Le dépôt légal du Web, terrain de compétition à la française." *Le Monde*, 6 April. Also available at: <http://www-rocq.inria.fr/~abitebou/pub/lemonde02.html>
- Muir Gray, J.A. & de Lusignan, S. (1999). "National electronic Library for Health (NeLH)." *BMJ*, 319, 1476-1479
- Murray, B. & Moore, A. (2000) *Sizing the Internet*. Cyveillance White Paper, July. Available at: [http://www.cyveillance.com/web/downloads/Sizing\\_the\\_Internet.pdf](http://www.cyveillance.com/web/downloads/Sizing_the_Internet.pdf)
- National Archives of Australia. (2001a). *Archiving Web resources: a policy for keeping records of Web-based activity in the Commonwealth Government*. Canberra: NAA, January. Available at: [http://www.naa.gov.au/recordkeeping/er/web\\_records/archweb\\_policy.pdf](http://www.naa.gov.au/recordkeeping/er/web_records/archweb_policy.pdf)
- National Archives of Australia. (2001b). *Archiving Web resources: guidelines for keeping records of Web-based activity in the Commonwealth Government*. Canberra, NAA, March. Available at: [http://www.naa.gov.au/recordkeeping/er/web\\_records/archweb\\_guide.pdf](http://www.naa.gov.au/recordkeeping/er/web_records/archweb_guide.pdf)

Office of the e-Envoy. (2002). *Guidelines for UK Government Websites*, Annex L: Archiving Websites. Available at: [http://www.e-envoy.gov.uk/oe/OeE.nsf/sections/webguidelines-handbook-annexes/\\$file/annexl.htm](http://www.e-envoy.gov.uk/oe/OeE.nsf/sections/webguidelines-handbook-annexes/$file/annexl.htm)

O'Reilly, T. (2001). "Remaking the peer-to-peer meme." In: Oram, A., ed., *Peer-to-peer: harnessing the benefits of a disruptive technology*. Sebastapol, Calif.: O'Reilly, 38-58.

Patsos, M. (2001). "The Internet and medicine: building a community for patients with rare diseases." *JAMA*, 285 (6), 805.

Pew Internet & American Life Project. (2002). *One year later: September 11 and the Internet*. Washington, D.C.: Pew Internet & American Life Project, 5 September. Available at: [http://www.pewinternet.org/reports/pdfs/PIP\\_9-11\\_Report.pdf](http://www.pewinternet.org/reports/pdfs/PIP_9-11_Report.pdf)

Phillips, M. (1998). "The preservation of Internet publications." 7th International World Wide Web Conference, Brisbane, Australia, 14-18 April 1998. Available at: <http://www.nla.gov.au/nla/staffpaper/www7mep.html>

Phillips, M. (1999). "Ensuring long-term access to online publications." *Journal of Electronic Publishing*, 4 (4), June. Available at: <http://www.press-umich.edu/jep/04-04/phillips.html>

Phillips, M. (2002). "Archiving the Web: the national collection of Australian online publications." International Symposium on Web Archiving, National Diet Library, Tokyo, Japan, 30 January 2002. Available at: <http://www.nla.gov.au/nla/staffpaper/2002/phillips1.html>

Porter, R. (1989). *Health for sale: quackery in England, 1660-1850*. Manchester: Manchester University Press.

Public Record Office. (2001). *Managing Web resources: management of electronic records on Websites and Intranets: an ERM toolkit*, v. 1.0. Kew: Public Record Office, December.

Raghavan, S. & Garcia-Molina, H. (2001). "Crawling the hidden Web." In: *VLDB 2001: Proceedings of 27th International Conference on Very Large Data Bases, September 11-14, 2002, Roma, Italy*. San Francisco, Calif.: Morgan Kaufmann. Available at: [http://www.dia.uniroma3.it/~vldbproc/017\\_129.pdf](http://www.dia.uniroma3.it/~vldbproc/017_129.pdf)

Rauber, A. & Aschenbrenner, A. (2001). "Part of our culture is born digital: on efforts to preserve it for future generations." *Trans: Internet-Zeitschrift für Kulturwissenschaften*, 10, July. Available at: <http://www.inst.at/trans/10Nr/rauber10.htm>

Rauber, A., Aschenbrenner, A. & Schmidt, A. (2001). "Austrian on-line archive: current status and next steps." What's next for digital deposit libraries? ECDL Workshop, Darmstadt, Germany, 8 September 2001. Available at: <http://www.bnf.fr/pages/infopro/ecdl/austria/austria.pdf>

Rauber, A., Aschenbrenner, A. & Witvoet, O. (2002). "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, 2458. Berlin: Springer, 16-31.

Rauber, A., Aschenbrenner, A., Witvoet, O., Bruckner, R.M. & Kaiser, M. (2002). "Uncovering information hidden in Web archives: a glimpse at Web analysis building on data warehouses." *D-Lib Magazine*, 8 (12), December. Available at: <http://www.dlib.org/dlib/december02/rauber/12rauber.html>

RLG/OCLC Working Group on Digital Archive Attributes. (2002). *Trusted digital repositories: attributes and responsibilities*. Mountain View, Calif.: Research Libraries Group. Available at: <http://www.rlg.org/longterm/repositories.pdf>

- Ryan, D. (2002). "Preserving the No 10 Web site: the story so far." Web-archiving: managing and archiving online documents, DPC Forum, London, 25 March 2002. Presentation slides available at: <http://www.jisc.ac.uk/dner/preservation/presentations/pdf/Ryan.pdf>
- Sherman, C. & Price, G. (2001). *The invisible Web: uncovering information sources search engines can't see*. Medford, N.J.: CyberAge Books.
- Sprehe, J.T. (2001). "Federal Web sites: half-cocked snapshots." *Federal Computer Week*, 5 February. <http://www.fcw.com/fcw/articles/2001/0205/pol-sprehe-02-05-01.asp>
- Stafford-Fraser, Q. (2001). "The life and times of the first Web cam." *Communications of the ACM*, 44 (7), 25-26.
- Stata, R. (2002). "Presentation of the Internet Archive." 2nd ECDL Workshop on Web Archiving, Rome, Italy, 19 September 2002. Available at: <http://bibnum.bnf.fr/ecdl/2002/>
- Voerman, G., Keyzer, A. & Hollander, F. den. (2000). "Digital incunabula." English translation of article published in: *De Nieuwste Tijd*, 15, 125-131. Available at: <http://www.archipol.nl/english/project/publications/incunabula.html>
- Webb, C. (1999). "Preservation of electronic information: what we should be thinking about now." *International Preservation News*, 18, March, 8-13.
- Webb, C. (2000). "Towards a preserved national collection of selected Australian digital publications" *New Review of Academic Librarianship*, 6, 179-191. Also available at: <http://www.nla.gov.au/nla/staffpaper/2000/webb6.html>
- 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. Available at: <http://www.oclc.org/events/presentations/symposium/preisswebb.shtm>
- Webb, C. (2002). "Digital preservation: a many-layered thing: experience at the National Library of Australia." In: *The state of digital preservation: an international perspective*. Washington, D.C.: Council on Library and Information Resources, 65-77. Available at: <http://www.clir.org/pubs/abstract/pub107abst.html>
- Wilson, M.I. (2001). "Location, location, location: the geography of the dot com problem." *Environment and Planning B: Planning and Design*, 28 (1), 59-71.
- Woodyard, D. (2002). "Britain on the Web." 2nd ECDL Workshop on Web Archiving, Rome, Italy, 19 September 2002. Available at: <http://bibnum.bnf.fr/ecdl/2002/>
- Working Party on Legal Deposit (1998). *Report of the Working Party on Legal Deposit*. London: British Library. Available at: <http://www.bl.uk/about/policies/workreplegdep.html>

## Appendix A: Evaluation of selected medical sites

A sample of thirty-one medical Web sites was chosen for review as part of this study (listed in the Invitation to Tender). Each one of these was searched for in the Internet Archive's Wayback Machine and manually checked to see whether the version collected there represented a version of the Web site sufficient for preservation. In the time available, it was not possible to evaluate all of the versions of Web sites accessible via the Wayback Machine, but just to look at two or more versions of the home page and to check whether links could be followed and whether most images were present. This brief evaluation showed that while a large amount of Web information is available from the Wayback Machine, almost all Web sites were missing some content or aspects of its functionality.

Potential problems with the evaluation methodology were:

- The browser used for the evaluation was Netscape Navigator v 6, did not have all plug-ins required, e.g. Java, VRML. This meant that the functionality of all pages could not be fully evaluated.
- The Wayback Machine retrieves individual Web pages searched for by URL, while the evaluation was concerned with looking at the whole Web site. This means that the figures given for the number of pages held (and their dates) only apply to the URL that was first searched for.
- A better sense of the completeness of the Internet Archive's holdings may be possible by running automatic link-checker software. However, the Wayback Machine seamlessly links versions of Web sites collected on different dates, so even this could have been misleading.

Main findings:

- Not in archive: only one Web site (the Nuffield Council on Bioethics) appeared not to be retrievable from the Wayback Machine (this may be because of a recent change in URL). Older versions of several sites could only be found when the URL was changed to an older form.
- Redirecting outside Wayback Machine: one of the most common problems found were links that redirected to current versions of Web pages. All site and database search facilities that were tested did this, as did passwords, some 'button' functions (e.g., based on cgi-scripts) and internal navigation features that used JavaScript. On a few pages, even plain HTML hyperlinks did the same thing (e.g. Sanger Institute, 27 June 1997).
- Missing images: another common problem that was frequently encountered was missing images. In some cases the existence of ALT tags meant that the Web pages could be used OK. In others, the lack of images meant that site navigation was impossible.
- Missing links: the Wayback Machine made a variety of different file types available. Excel spreadsheets, PDFs, PowerPoint files, etc. were all successfully downloaded during the evaluation. However, many links tested did not retrieve any information. Sometimes this was due to the existence of robots.txt, sometimes the archive promised that it would try to retrieve these pages on its next crawl, at other times because the page had been archived but not indexed. There was sometimes no clear way of explaining why the crawler had followed one link while ignoring an adjacent one. Very few Web sites appeared to be complete in the archive, although those based almost completely on simple linked HTML pages fared better than others. On one site (the evidence-based-health e-mail list), all the significant links with content did not work, possibly because the URLs included a cgi query.

- Interactive content and multimedia: the existence of interactive content or multimedia often effected the way in which archived pages would display. Some of this, however, may be due to the browser type used in the evaluation.

To summarise, the Wayback Machine provides an interesting window onto a great deal of the medical Web. It is a resource that will repay serendipitous surfing, despite the frequency of missing pages. In its existing form, however, it is not a substitute for a focused Web collection and preservation initiative.

## 1. Dynamic sites

### British National Formulary (BNF.org)

*Description:* The British Medical Association (BMA) and the Royal Pharmaceutical Society of Great Britain jointly publish the *British National Formulary* (BNF) in book form every six months. It is currently in its 43rd edition (March 2002). *WeBNF* was developed by the Clinical and Biomedical Computing Unit at the University of Cambridge and can be customised for local use. The BNF Web site contains some JavaScript and claims to be optimised for viewing "at a screen resolution of 1024 x 768 with 256 or more colours using Microsoft® Internet Explorer version 5 or above." The BNF data itself is held in a searchable database and a plug-in (Java Run Time Environment (JRE) v1.3?) is required to view results.

*Internet Archive:* A search of the Wayback Machine retrieved 37 versions of this page, dating from December 1998 to October 2001. The earliest pages are 'holding' pages; the first with content is dated 18 October 2000, and links to *WeBNF* 39. The links to HTML pages that were tested worked (although some images were missing or loaded incorrectly) but an attempt to link to the database itself produced an alert: "WeBNF is temporarily unavailable. Please try later." Attempting to link to *WeBNF* 41 in a more recent version of the page (that dated 25 September 2001) loads an HTTP 404-error message in one of its frames. Attempting to search also loads a 404-error message. Some of the other documents in frames will also not load.

To summarise, the Internet Archive version of the BNF Web site makes available some aspects of the general look and feel of the site (including the product information), but does not preserve any of the underlying data.

*URL:* <http://www.bnf.org/>

*Versions of Web site evaluated:*

<http://web.archive.org/web/19981212012912/http://bnf.org/>

<http://web.archive.org/web/20001018030130/http://www.bnf.org/>

<http://web.archive.org/web/20010925071426/http://bnf.org/>

### NeLH Guidelines database

*Description:* The NeLH Guidelines database is part of the National electronic Library for Health (NeLH). These guidelines change over time. An online archive of this site would allow historians to plot these. The site requires the use of Netscape 3 or Internet Explorer 4 upwards.

*Internet Archive:* The URL for this site has changed since the production of the ITT. Searches of the Wayback Machine shows that the Internet Archive provides access to four versions of the Web site under its old URL ([http://www.nelh.nhs.uk/guidelines\\_database.asp](http://www.nelh.nhs.uk/guidelines_database.asp)) from 2001. However, searches under its current URL did not match any results.

The earliest version in the Internet Archive is from March 2001. The images on the home page are 'greyed out' in the version available, but those external links that were checked did work. However, the button link to the database itself did not work. It attempted to link to IP address 195.152.56.15 (in a separate browser window) but without success. The link *did* work in the latest version of the site available from the Wayback Machine (dated October 2001), but the way the link was constructed meant that it linked to the *current* version of the NeLH *Index of Guidelines* rather than one in the Internet Archive. A separate search of the Wayback Machine for this particular URL (<http://www.nelh.nhs.uk/guidelinesdb/html/glframes.htm>) showed that access to this page is blocked via robots.txt.

To summarise, while versions of the NeLH Guidelines database Web pages are successfully replicated in the Internet Archive (but without images); the database itself is not available.

*URL:* <http://www.nelh.nhs.uk/guidelinesfinder/>

*Versions of Web site evaluated:*

[http://web.archive.org/web/20010305134630/http://www.nelh.nhs.uk/guidelines\\_database.asp](http://web.archive.org/web/20010305134630/http://www.nelh.nhs.uk/guidelines_database.asp)

[http://web.archive.org/web/20011026001720/http://www.nelh.nhs.uk/guidelines\\_database.asp](http://web.archive.org/web/20011026001720/http://www.nelh.nhs.uk/guidelines_database.asp)

### **Department of Health Circulars on the Internet (COIN)**

*Description:* This is a database of circulars published by the UK Department of Health (DoH). The full-text of circulars are made available in PDF, sometimes also in HTML. Unlike other official publications (e.g. Statutory Instruments) circulars are issued by individual government departments rather than formally published by The Stationery Office (TSO) or Her Majesty's Stationery Office (HMSO). Traditionally, circulars used to be issued in printed form and would normally be supplied to the copyright libraries through legal deposit. Circulars are now routinely made available on departmental Web sites. A report of the UK Office of the e-Envoy (2002) notes that circulars have been identified as a category of information "that have been identified as being of particular interest to the Parliamentary libraries, university libraries, and the major public reference libraries that maintain collections of official material." Long-term responsibility for the preservation of departmental circulars would probably either belong to the copyright libraries (e.g., through legal deposit) or - if classed as official records - the Public Record Office (PRO).

*Internet Archive:* A search of Wayback Machine for the entry point from the Department of Health Web pages (<http://www.doh.gov.uk/coinh.htm>) retrieved 15 versions of this page from between 1999 and 2000. The first of these, from February 1999, linked successfully to a version of the COIN database that could be browsed by title and series. These links retrieved metadata about circulars and a link to a full-text version of documents in PDF. However, not all of these links were able to retrieve the full-text. A later version of the DoH entry point (May 2000) linked to the version of the COIN Web site that allowed browsing by series. Despite some of the navigation links not working, it was possible to browse through the site but many of the document links did not work. An attempt to use the search function sent an unsuccessful database enquiry directly to the [tap.ccta.gov.uk](http://tap.ccta.gov.uk) domain.

Note that there are no matches in the Wayback Machine for a search on the current main COIN search page (<http://www.info.doh.gov.uk/doh/coin4.nsf/Circulars?ReadForm>.) or for any versions of the site later than November 2000.

To summarise, the versions of COIN held in the Internet Archive retain some of the functionality of the original Web site but much of the content is missing. This did not seem to be the result of the Web site structure or the existence of robots.txt, but possibly because the Alexa crawler did not attempt to collect all pages in the Web site.

*URL:* <http://www.info.doh.gov.uk/doh/coin4.nsf/Circulars?ReadForm>

*Versions of Web site evaluated:*

<http://web.archive.org/web/19990127161516/tap.ccta.gov.uk/doh/coin4.nsf>

<http://web.archive.org/web/19990218042854/http://www.doh.gov.uk/coinh.htm>

### **ClinicalTrials.gov**

*Description:* ClinicalTrials.gov is a service provided by the US National Institutes of Health through the National Library of Medicine "to provide patients, family members and members of the public current information about clinical research studies" (<http://clinicaltrials.gov/>). The Web site is useful for getting a perspective on current research priorities. Looking at this in 2001 one can quickly see that most clinical trails are in the fields of cancer and immune system diseases. Over time these priorities may change.

*Internet Archive:* A search of the Wayback Machine showed that 36 versions of the site are held in the Internet Archive, the earliest from April 2000 and the latest from January 2002.

A more detailed look at the earliest version of ClinicalTrials.gov retrieved from the Wayback Machine (dated 8 April 2000) revealed that it was possible to browse through conditions alphabetically and retrieve full details of studies that were recruiting. The browse by disease heading, however, did not work. Also, use of the search facilities sends terms to the current database at ClinicalTrials.gov rather than the version in the Internet Archive.

The browse by disease heading does work in later versions of the archived Web site (e.g. that dated January 2002), as well as the additional browse by sponsor. There are several points in the browse interface that will direct users to the current database rather than the one held in the Internet Archive. There are tick boxes that allow all trials to be listed and for users to choose which ones they want to view in more detail. If either of these facilities are used, the browse redirects to the current database. Simple HTML links within the browse do appear to work but some other links (e.g. those to PubMed) open a new window and again direct users away from archive.org to the current version of the site.

To summarise, the versions of ClinicalTrials.gov held in the Internet Archive retain much of the functionality of the original site. The search facilities and parts of the browse interface will, however, redirect to the current version of the site.

*Current URL:* <http://clinicaltrials.gov/>

*Versions of Web site evaluated:*

<http://web.archive.org/web/20000408223033/clinicaltrials.gov/ct/gui/c/r>

<http://web.archive.org/web/20001201223800/http://www.clinicaltrials.gov/>

<http://web.archive.org/web/20020124043127/http://clinicaltrials.gov/>

## **2. Organisational sites**

### **The Wellcome Trust**

*Description:* The Wellcome Trust is an UK-based charity that funds scientific research in the areas of human and animal health. The Wellcome Trust's Web site.

*Internet Archive:* A search of the Wayback Machine showed that 42 versions of the site are held in the Internet Archive, the earliest from January 1997 and the latest from July 2001. No versions of the Wellcome Trust Web site, however have been downloaded since then.

Linking to the earliest site available in the Wayback Machine (dated 30 January 1997) gave access to two versions of the Wellcome Trust Web site, one using frames and graphics, the other being text only. Of these the text-based one appeared to work very well, while the one using frames and graphics did not deal with these very well leaving parts of the graphical navigation invisible. As with other sites, a attempt to search the site redirected the request to the current wellcome.ac.uk domain rather than the archived version. Naturally, the Telnet link to the Wellcome Library's Catalogue (uk.ac.ucl.wihm) also did not work.

Later versions of the site collected by the Internet Archive (e.g. that collected on 17 October 2001) have good representations of the home page including all images. As usual, site searches are automatically redirected to the current server, but much other content (including images) is present. The 'crosslinks' and 'quicklinks' boxes present on most pages also redirect users to the current Wellcome Web server. As for content, a few pages appeared not to be available, but the majority of links followed were successful.

To summarise, the versions of the Wellcome Trust Web pages held in the Internet Archive retain much of the content and some of the functionality of the original site. However, no version of the site has been downloaded since June 2001.

*URL:* <http://www.wellcome.ac.uk/>

*Versions of Web site evaluated:*

<http://web.archive.org/web/19970130050435/http://www.wellcome.ac.uk/>

<http://web.archive.org/web/20000303100417/http://www.wellcome.ac.uk/>

### **National Childbirth Trust**

*Description:* This is an example of pressure group and consumer health organisation.

*Internet Archive:* Five versions of the National Childbirth Trust are retrieved from a search of the Wayback Machine. These are all from 2001. Linking to the earliest (dated 23 February 2001) reveals that the content of the home page can be displayed (including images) but the page layout is not the same. This makes navigation more difficult in the version stored in the Internet Archive. The archived version does not include all images (some of which are links to site sponsors through doubleclick.net) and the 'pop-up' window that appears on opening the original site does not load properly. By contrast, all of the articles on aspects of pregnancy, childbirth and childcare do appear to be available in the Internet Archive versions.

*URL:* <http://www.nctpregnancyandbabycare.com/>

*Versions of Web site evaluated:*

<http://web.archive.org/web/20010223214151/http://www.nctpregnancyandbabycare.com/>

### **National Institutes of Health**

*Description:* The National Institutes of Health is the US Federal agency focusing on medical research. Its Web domain is huge and includes the Web sites of 27 national institutes and centres, including the National Library of Medicine (NLM), the National Cancer Institute and the National Human Genome Research Institute. This evaluation will not be able to cover all of these sites, so will concentrate on the NLM and the central NIH Web sites.

*Internet Archive:* A search of the Wayback Machine retrieves 167 versions of the NIH home page (<http://www.nih.gov/>) dating from December 1997 to January 2002 (131 of these, however, are from the year 2001). The earliest version (dated 10 December 1997) displays the home page with most images intact.

Evaluation of a later version of the Web site (dated 1 March 2000) showed that most links to public HTML pages worked, including many images. As usual, searches of the Web site and CGI-based 'buttons' redirected to the current NIH server. The archived version of the NIH Web site even retained the 'access denied' pages delivered to unauthorised users (e.g. on Millennium Bug information). Linking to the National Library of Medicine home page retrieved a slightly later version (dated 15 August 2000) but displayed correctly. However, access to some informational links, including the search page of Medline (via PubMed) was blocked by robots.txt while other pages were missing images. Searches of the Medical Subject Headings (MeSH) browser and other services defaulted to the current NLM Web domain while other search forms (e.g. for the NLM's LOCATORplus catalogue) had not been downloaded.

*URL:* <http://www.nih.gov/>

*Versions of Web site evaluated:*

<http://web.archive.org/web/19971210191959/http://www.nih.gov/>

<http://web.archive.org/web/20000301105907/http://www.nih.gov/>

<http://web.archive.org/web/20000815052859/www.nlm.nih.gov/>

### **Nuffield Council on Bioethics**

*Description:* The Nuffield Council on Bioethics is an independent body jointly funded by the Nuffield Foundation, the Wellcome Trust and the Medical Research Council (MRC). It deals with ethical issues arising from developments in medicine and biology. Historical versions of the Web site would be useful for seeing how attitudes to bio-ethical issues change over time.

*Internet Archive:* The Wayback Machine found no version of this Web site.

*URL:* <http://www.nuffieldbioethics.org/home/index.asp>

### **Human Fertilisation and Embryology Authority**

*Description:* Useful again for seeing how attitudes to IVF change over time.

*Internet Archive:* A search of the Wayback Machine showed that 17 versions of the site are held in the Internet Archive, the earliest from November 1998 and the latest from January 2002. No versions of the site were downloaded in 2001. The Web site has been comprehensively redesigned since the last version was downloaded by the Internet Archive on 25 January 2002.

The earliest version of the site (dated 11 November 1998) available on the Internet Archive was based on HTML FRAMES and the images essential to navigation did not display properly. Despite this problem, the links themselves did appear to work. Some documents were not completely downloaded; e.g. some chapters in *The Patients' Guide to Infertility & IVF* were not available.

*URL:* <http://www.hfea.gov.uk/>

*Versions of Web site evaluated:*

<http://web.archive.org/web/19981111191342/http://www.hfea.gov.uk/>

## World Health Organization

*Description:* The World Health Organization (WHO) is the United Nations agency with responsibility for health. Its Web site is available in English, French and Spanish

*Internet Archive:* A search of the Wayback Machine retrieves 492 versions of the WHO home page (<http://www.who.int/>) dating from December 1998 to January 2002 (112 of these are from the year 2001). The earliest version (dated 2 December 1998) displays the home page with most, but not all, images intact (some later versions included all images). A brief evaluation of the December 1998 version, showed that many of the top-level HTML pages were available in the Internet Archive, but there were problems with some search based pages (e.g. the catalogue of publications).

Linking to a later version of the site (dated 8 January 2001) successfully retrieved the contents of the *Bulletin of the World Health Organization*, with links to the full-text in PDF.

*URL:* <http://www.who.int/>

*Versions of Web site evaluated:*

<http://web.archive.org/web/19981202200903/http://www.who.int/index.html>

<http://web.archive.org/web/20010108153600/http://www.who.int/>

## British Medical Association

*Description:* The professional association

*Internet Archive:* A search of the Wayback Machine retrieved 26 versions of the current BMA home page (<http://web.bma.org.uk/>), dating from December 1998 to September 2001. The earliest version of this Web site (dated 12 December 1998) is interesting because it is a pilot (it reads "Welcome to the home page of the BMA Paperless Committee Trial Web Site"), but none of the links appear to work. This page persists (with some changes) until May 2001, when it is labelled the "new homepage" of the BMA. This version of the home page (dated 16 May 2001) displays well (with all images), but not all of the internal links work. The JavaScript based navigation buttons display OK, but using them redirect users to the current BMA Web domain.

A search on Wayback Machine retrieves 459 versions of the BMA home page at its older address (<http://www.bma.org.uk/>). These versions date from April 1997 to January 2002 (and the later ones are the same as: <http://web.bma.org.uk/>). The earliest version of this that can be retrieved from the Wayback Machine (dated 4 April 1997) is missing some images and not all links could be successfully followed.

*URL:* <http://web.bma.org.uk/>

*Versions of Web site evaluated:*

<http://web.archive.org/web/19981212022112/http://web.bma.org.uk/>

<http://web.archive.org/web/20010516000527/http://web.bma.org.uk/>

<http://web.archive.org/web/19970404184408/http://www.bma.org.uk/index.html>

### 3. Multimedia sites and resources using plug-ins

#### Ventricular Catheterisation v2

*Description:* The Ventricular Catheterisation simulator is one of the services hosted by the Manchester Visualization Centre (MVC) at the University of Manchester. The site makes use of VR technologies and requires the use of a VRML plug-in.

*Internet Archive:* A search of the Wayback Machine retrieved 10 versions of Ventricular Catheterisation simulator dating from March 2000 to October 2001. The browser used for the evaluation, however, did not have the VRML plug-in installed so it was not possible to compare its use with the version available from the MVC. Unlike the MVC version, no pop up browser window containing instructions appeared on retrieving the earliest version (dated 7 March 2000).

*URL:* [http://synaptic.mvc.mcc.ac.uk/Ventricular\\_audio.html](http://synaptic.mvc.mcc.ac.uk/Ventricular_audio.html)

*Versions of Web site evaluated:*

[http://web.archive.org/web/20000307000813/http://synaptic.mvc.mcc.ac.uk/Ventricular\\_audio.html](http://web.archive.org/web/20000307000813/http://synaptic.mvc.mcc.ac.uk/Ventricular_audio.html)

#### DIPEX

*Description:* DIPEX is the Database of Individual Patient Experiences provided by the University of Oxford, Institute of Health Sciences, Department of Primary Care. The site contains video and audio clips and other medical information. Technically, the site makes a major use of multimedia features requiring plug-ins like Flash and Real Video.

*Internet Archive:* A search of the Wayback Machine retrieved 5 versions of the DIPEX home page, all dating from 2001. The opening page from the earliest version listed (dated 9 February 2002) was missing its main image (Flash based?) but allowed users to link through to a prototype page on the prostate. However all links on this page were dead.

A later version (dated 26 September 2001) of the home page displayed better, despite missing background colour. Some links worked (a few pages had been archived but not indexed) but much of the functionality of the site was not present and the Flash animations essential for more detailed navigation did not load.

*URL:* <http://www.dipex.org/>

*Versions of Web site evaluated:*

<http://web.archive.org/web/20010209030127/http://www.dipex.org/>

<http://web.archive.org/web/20010926231322/http://www.dipex.org/>

#### UN Aids

*Description:* This is the *Report on the global HIV/AIDS epidemic* published in June 2000 by the Joint United Nations Programme on HIV/AIDS. Chapters are in a variety of file formats: HTML, PDF and MS Excel spread sheets.

*Internet Archive:* A search of the Wayback Machine retrieved 4 versions of this page, dating from August 2000 to October 2001. The main contents page, and all of those linked files that were tested, were retrieved successfully in the Internet Archive versions. This means that the Internet Archive successfully collected the different file formats used in this report.

*URL:* [http://www.unaids.org/epidemic\\_update/report/](http://www.unaids.org/epidemic_update/report/)

*Versions of Web site evaluated:*

[http://web.archive.org/web/20011031190104/http://www.unaids.org/epidemic\\_update/report/](http://web.archive.org/web/20011031190104/http://www.unaids.org/epidemic_update/report/)

### **University of Edinburgh, Department of Orthopaedic Surgery**

*Description:* A departmental Web site from an UK University. Chosen for the number of PowerPoint files included on the site. These currently run as presentations within Internet Explorer v. 4 (and above) or can be downloaded if the user does not use the relevant versions of this browser.

*Internet Archive:* A search of the Wayback Machine retrieved 32 versions of this page, dating from February 1998 to July 2001. The earliest viewable version of the home page displayed OK, but some images were not retrieved - including the navigation bar at the top of the screen. The PowerPoint presentations are available from later versions of the Web site. We were able to successfully download them from the version dated 1 February 2001.

*URL:* <http://www.orthopaedic.ed.ac.uk/>

*Versions of Web site evaluated:*

<http://web.archive.org/web/19981205044846/http://www.orthopaedic.ed.ac.uk/>

<http://web.archive.org/web/20010201151900/http://www.orthopaedic.ed.ac.uk/>

### **Anatomy.tv**

*Description:* This is a product from Primal Pictures (<http://www.primalpictures.com/about.htm>) The Web site uses the national Internet domain of Tuvalu (.tv), although use of this domain is currently licensed to a company that is part of VeriSign. The images in anatomy.tv are made up from body scans to produce three-dimensional anatomical pictures that can be rotated, layers can also be removed to expose underlying structures. There are requirements for screen resolution and mode. [Nothing from this page loads on Netscape 6].

*Internet Archive:* A search of the Wayback Machine retrieved 2 versions of this page, both dating from 2001. Unlike the current version of the page, the archived home page loaded OK in Netscape 6 (but prompted requests to change screen resolution, etc.). The anatomy.tv service, however, requires users to login with username and password and anyone attempting to do this will proceed to the current version of the Web site.

*URL:* <http://nhs.anatomy.tv/>

*Versions of Web site evaluated:*

<http://web.archive.org/web/20010816190415/http://nhs.anatomy.tv/>

## **4. E-journals and newsletters, etc.**

### **Bandolier**

*Description:* *Bandolier* is a monthly health care journal edited and produced by staff based at the Pain Relief Unit in Oxford. It uses evidence-based medicine techniques to provide advice about particular treatments or diseases for both healthcare professionals and consumers. The

Web site contains a large quantity of value-added information not available in the printed versions of the journal. For example, there are essays on specific topics (Bandolier Extra) and collections of good quality evidence (Bandolier Knowledge). The site deliberately keeps each issue simple (e.g., no frames) because the producers want it "to be rapidly accessed even by folks at home with the slowest modem" (<http://www.jr2.ox.ac.uk/bandolier/aboutus.html>). The Internet edition is made available from a University of Oxford server based at the John Radcliffe Hospital. It is, therefore, part of the ac . uk domain. There is also a 'redirect' from the URL <http://www.ebandolier.com/>.

*Internet Archive:* A search of the Wayback Machine retrieved 31 versions of this page, dating from January 1998 to January 2002; 17 of these from 2001. The earliest viewable version of the home page (dated 26 January 1998) displayed OK, but did not include images. It was possible to navigate through to individual issues and articles, but the search function redirected the search (unsuccessfully) to the jr2.ox.ac.uk domain. Later versions of the site stored in the Internet Archive allow PDFs of individual issues to be downloaded.

*URL:* <http://www.jr2.ox.ac.uk/bandolier/>

*Versions of Web site evaluated:*

<http://web.archive.org/web/19980126202741/http://www.jr2.ox.ac.uk/bandolier/>

<http://web.archive.org/web/20010124065100/http://www.jr2.ox.ac.uk/bandolier/>

## **bmj.com**

*Description:* The electronic *BMJ* (bmj.com) is the online version of the weekly journal published by the British Medical Association, formerly known as the *British Medical Journal*. It is published with the assistance of Stanford University Libraries' HighWire Press. The full-text of *BMJ* is made available in HTML, with the additional option of downloading PDF versions, where these exist. The full-text of papers published since June 1998 are also available from the PubMed Central service (<http://www.pubmedcentral.nih.gov/>) provided by the NIH.

The bmj.com service also contains many features not available in the printed versions of the *BMJ*. These include:

- A combination of a short paper published in the printed edition with a longer one available on bmj.com - known in *BMJ* parlance as ELPS (electronic long, paper short).
- Supplementary material published only on bmj.com, e.g., additional references, illustrations, audio and video clips, etc.
- Rapid responses to articles published in the journal. An editorial published in May 2002 reported that these rapid responses made up 40% of the searchable content of the online *BMJ* (Delamothe & Smith, 2002).

The bmj.com Web site also contains a range of other services: e.g., general information for potential contributors or advertisers, a digest of UK health news (<http://bmj.com/uknews/>), the *Student BMJ* (<http://www.studentbmj.com/>) and NetPrints, a repository of non-peer reviewed research in clinical medicine published by the BMJ Publishing Group, again in association with HighWire Press (<http://clinmed.netprints.org/>).

*Internet Archive:* A search of the Wayback Machine retrieved 693 versions of this page, dating from December 1996 to January 2002; the majority of these from 2000 and 2001. The earliest page dates from December 1996 (i.e., before bmj.com began to offer full-text access to all issues). Some of the internal links work (e.g. 'about the *BMJ*' in pre-1998 versions of the Web site), but the vast majority of content (including all published papers) appear to be

protected by the robots exclusion protocol (also for the content made available through PubMed Central). Browsing through the archived sites is difficult because many of the navigation buttons are 'greyed out'.

*URL:* <http://bmj.com/>

*Versions of Web site evaluated:*

<http://web.archive.org/web/19961219183138/http://www.bmj.com/>

<http://web.archive.org/web/20010104201100/http://www.bmj.com/>

## **The Lancet**

*Description:* Internet site of the weekly medical journal that contains information not included in the paper version. Many services are only available to subscribers.

*Internet Archive:* A search of the Wayback Machine retrieved 106 versions of this page, dating from December 1996 to January 2002; the majority of these from 2000 and 2001. The earliest version of the home page archived (dated 26 December 1996) displayed OK with all images included. Not all the internal links on this page worked, however, and some of those that did work linked to much later versions of the page (from 2000). It was not possible to consistently link to issue tables of contents, although some articles in The Lancet's Electronic Research Archive (ERA) e-prints service (set up in 1999) could be retrieved.

*URL:* <http://www.thelancet.com/>

*Versions of Web site evaluated:*

<http://web.archive.org/web/19961226110437/http://www.thelancet.com/>

<http://web.archive.org/web/20000229151113/http://www.thelancet.com/index.html>

## **Eurosurveillance**

*Description:* A weekly journal that uses RTF and makes pages available in different languages. The URL of this site is now: <http://www.eurosurveillance.org/index-02.asp>

*Internet Archive:* A search of the Wayback Machine on the old URL retrieved 12 versions of the page, dating from October 1999 to November 2001. The earliest version of the home page archived (dated 12 October 1999) displayed OK. It was possible to link through to the latest editions of the journal and to access previous issues both in HTML and RTF. A Wayback Machine search for the newer URL (<http://www.eurosurveillance.org/>) retrieved 5 versions from December 2000 to September 2001. While the home page itself displayed OK, other pages were missing navigation icons and the text of some issues of the journal were not available.

*URL:* <http://www.ceses.org/eurosurveillance.htm>

*Versions of Web site evaluated:*

<http://web.archive.org/web/19991012215253/http://www.ceses.org/eurosurveillance.htm>

<http://web.archive.org/web/20010208093849/http://eurosurveillance.org/>

### **Clinical Governance Bulletin**

*Description:* Bulletin published 6 times per year by the Royal Society of Medicine.

*Internet Archive:* A search of the Wayback Machine retrieved 7 versions of the page, dating from August 2000 to June 2001. The earliest of these (dated 23 August 2000) linked to the 1st issue of the bulletin - which could be downloaded in PDF - while the latest version of the bulletin available in the archive (to date) is that published July 2001.

*URL:* <http://www.roysocmed.ac.uk/pub/cgb.htm>

*Versions of Web site evaluated:*

<http://web.archive.org/web/20000823040940/http://www.roysocmed.ac.uk/pub/cgb.htm>

<http://web.archive.org/web/20010617004312/http://www.roysocmed.ac.uk/pub/cgb.htm>

### **Evidence-based-health discussion list**

*Description:* This is the Web site of the evidence-based-health e-mail discussion list, hosted by the JISCmail service. The Web site consists of some background material on the purpose of the list together with mail archives dating back to September 1998. Older messages, dating back to February 1998, are available from the Mailbase Web site (<http://www.mailbase.ac.uk/lists/evidence-based-health/archive.html>).

*Internet Archive:* A search of the Wayback Machine retrieved 4 versions of this page, all dating from 2001. The versions in the Internet Archive contain general information about the list and its associated files but none of the list content is available, probably because they are accessed via a CGI search.

*URL:* <http://www.jiscmail.ac.uk/lists/evidence-based-health.html>

*Versions of Web site evaluated:*

<http://web.archive.org/web/20010208091614/http://www.jiscmail.ac.uk/lists/evidence-based-health.html>

<http://web.archive.org/web/20011124205251/http://www.jiscmail.ac.uk/lists/evidence-based-health.html>

## **5. Large sites**

### **Atlas of human anatomy**

*Description:* A large site, but largely based on simple HTML and images.

*Internet Archive:* A search of the Wayback Machine retrieved 20 versions of this page, dating from October 1999 to January 2002. Those versions of the page that were tested demonstrated that most links tested worked OK. The occasional page would not retrieve an image, but this appeared to be the exception rather than the rule.

*URL:* <http://www.vh.org/Providers/Textbooks/HumanAnatomy/CrossSectionAtlas.html>

*Versions of Web site evaluated:*

<http://web.archive.org/web/19991009233057/http://www.vh.org/Providers/Textbooks/HumanAnatomy/CrossSectionAtlas.html>

<http://web.archive.org/web/20010203165400/http://www.vh.org/Providers/Textbooks/HumanAnatomy/CrossSectionAtlas.html>

### **BBC Health**

*Description:* A large and complex Web site that changes on a regular basis and is full of links to other BBC Web resources (e.g. programme schedules) and external sites.

*Internet Archive:* A search of the Wayback Machine retrieved 351 versions of this page, dating from August 2000 to February 2002; 205 of these from 2001. The earliest page available (dated 15 August 2000) displayed OK, although some images were 'greyed out.' Some internal links worked fine, while others (e.g. news items) were blocked by robots.txt.

*URL:* <http://www.bbc.co.uk/health/>

*Versions of Web site evaluated:*

<http://web.archive.org/web/20000815053415/http://www.bbc.co.uk/health/>

### **HebsWeb**

*Description:* The Web site of the Health Education Board of Scotland (HEBS). It contains information on a variety of health issues and would be useful for identifying contemporary health concerns.

*Internet Archive:* A search of the Wayback Machine retrieved 16 versions of this page, dating from December 1997 to October 2001. The home pages of all the versions linked to displayed OK, but not all of the internal links worked, e.g. PDF versions of publications. Some pages that used HTML FRAMES also would not display properly.

*URL:* <http://www.hebs.scot.nhs.uk/pro.htm>

*Versions of Web site evaluated:*

<http://web.archive.org/web/19971212001920/http://www.hebs.scot.nhs.uk/pro.htm>

<http://web.archive.org/web/20000304022359/http://www.hebs.scot.nhs.uk/pro.htm>

<http://web.archive.org/web/20010204052200/http://www.hebs.scot.nhs.uk/pro.htm>

### **Sanger Institute**

*Description:* The Wellcome Trust Sanger Institute is the largest genetic sequencing centre in the UK. It is also responsible for Web sites that are not in the sanger.ac.uk domain (e.g., <http://www.yourgenome.org/>) and has some overlap with others (e.g., <http://www.ebi.ac.uk/>, <http://www.ensembl.org/>). The site would be of interest to any future historian working on the history of genome research, including the development of the human genome.

*Internet Archive:* A search of the Wayback Machine retrieved 103 versions of this page, dating from February 1997 to January 2002; 72 of these from 2001. The earliest version that could be retrieved (dated 27 June 1997) displayed OK, but all links defaulted to the current version of the Web site rather than the copy in the Internet Archive. Links in later versions worked as expected, and much content was available including e-mail list archives. As expected, however, attempted searches of the Web site and of genome databases redirected to the current Web server. Some pages appeared to be protected by robots.txt.

*URL:* <http://www.sanger.ac.uk/>

*Versions of Web site evaluated:*

<http://web.archive.org/web/19970627055353/http://www.sanger.ac.uk/>

<http://web.archive.org/web/19980129074002/http://www.sanger.ac.uk/>

<http://web.archive.org/web/20010119032500/http://www.sanger.ac.uk/>

### **AidsMap**

*Internet Archive:* A search of the Wayback Machine retrieved 89 versions of this page, dating from December 1998 to January 2002. The site depends on images (and a Java plugin) for navigation and these features did not work very well in those Internet Archive versions of the site that were evaluated. The opening image in the version dated 6 December 1998 did not display at all, while other pages were missing most of their navigation features.

*URL:* <http://www.aidsmap.com/>

*Versions of Web site evaluated:*

<http://web.archive.org/web/19981206074015/http://www.aidsmap.com/>

<http://web.archive.org/web/20000301063738/http://www.aidsmap.com/>

<http://web.archive.org/web/20020125090945/http://www.aidsmap.com/>

## **6. Genetics and bioethics sites**

### **Office of genetics and disease prevention**

*Internet Archive:* A search of the Wayback Machine retrieved 48 versions of this page, dating from December 1998 to January 2002. In the earliest of these (dated 3 December 1998), many images did not display, and this had a negative effect on navigation. Some internal links (e.g. for newsletters) did not work. Later versions of the site were also missing many images, but better use of HTML ALT tags meant that navigation was possible, although some links defaulted to the current version of the Web site rather than the copy available in the Internet Archive.

*URL:* <http://www.cdc.gov/genetics/default.htm>

*Versions of Web site evaluated:*

<http://web.archive.org/web/19981203102302/http://www.cdc.gov/genetics/default.htm>

<http://web.archive.org/web/20010805230229/http://www.cdc.gov/genetics/default.htm>

<http://web.archive.org/web/20020126193531/http://www.cdc.gov/genetics/default.htm>

### **Blazing the genetic trail**

*Description:* This is the Web version of a popular report produced by the Howard Hughes Memorial Institute.

*Internet Archive:* A search of the Wayback Machine retrieved 21 versions of this page, dating from December 1998 to October 2001. Linking to the earliest version available (dated 2 December 1998) showed that the page displayed OK, although some images were missing.

Those internal links that were tested appeared to work OK, although later versions of the Web site were missing some of these. It was also possible to download a PDF version of the full-report.

*URL:* <http://www.hhmi.org/GeneticTrail/>

*Versions of Web site evaluated:*

<http://web.archive.org/web/19981206155952/www.hhmi.org/GeneticTrail/>

<http://web.archive.org/web/20000815220031/www.hhmi.org/GeneticTrail/>

<http://web.archive.org/web/20000817185033/www.hhmi.org/GeneticTrail/front/bagt.pdf>

### **Gene Therapy Advisory Committee**

*Description:* This is the Web site of a committee of the UK Department of Health. Its Web page contains information on the committee's function and composition together with meeting reports and publications.

*Internet Archive:* A search of the Wayback Machine retrieved 7 versions of this page under an older URL (<http://www.doh.gov.uk/genetics/gtac.htm>), dating from October 1999 to October 2000. These pages tested displayed OK (e.g. that dated 4 October 1999), although some links did not appear to work

*URL:* <http://www.doh.gov.uk/genetics/gtac/>

*Versions of Web site evaluated:*

<http://web.archive.org/web/19991004172039/http://www.doh.gov.uk/genetics/gtac.htm>

<http://web.archive.org/web/20000423101328/http://www.doh.gov.uk/genetics/gtac.htm>

### **Center for Bioethics and Human Dignity**

*Description:* The CBHD is an organisation that deals with bioethics, including end-of-life treatment, genetic intervention, euthanasia and suicide, reproductive technologies, etc. The Web site links to news, articles and information about relevant events. It also contains information on the CBHD, including a newsletter and membership details.

*Internet Archive:* A search of the Wayback Machine retrieved 23 versions of this page, dating from November 1999 to January 2002. The earliest of these retrieved (dated 28 November 1999) was missing some images, but HTML ALT tags aid navigation. Many of the links to external news services do not work (e.g. some are protected by robots.txt) but those internal links that were tested worked. The same was true of a later version of the site (dated 2 March 2001).

*URL:* <http://www.cbhd.org/>

*Versions of Web site evaluated:*

<http://web.archive.org/web/19991128232721/http://www.cbhd.org/>

<http://web.archive.org/web/20010302122516/http://cbhd.org/>

## **Reference**

Delamothe, T. & Smith, R. (2002). "Twenty thousand conversations." *BMJ*, 324, 1171-1172.  
Also available at: <http://bmj.com/cgi/content/full/324/7347/1171>

## Appendix B: Evaluation of selected eLib project Web sites

The Electronic Libraries (eLib) programme (<http://www.ukoln.ac.uk/services/elib>) was funded by JISC to develop aspects of digital library development in the UK Higher Education community. The programme funded over 70 projects, the majority of which hosted a Web site or home page. These were typically hosted by the lead University within a project consortia, although other partners sometimes had additional Web space dedicated to the project.

The eLib programme ended in 2001 and some project Web pages appear to be no longer available. As reported in *Ariadne* in January 2001 (Kelly, 2001):

*Of the 71 project Web sites which were surveyed, 3 Web sites (PPT, ACORN and ERIMS) are no longer available (i.e. the domain no longer exists) and 2 Web entry points (On Demand and TAPIN) are no longer available (i.e. the domain exists but not the project entry point).*

This does not necessarily mean, however, that the project Web site has disappeared forever. A more revisit of the Web sites revealed that the ACORN Web site (<http://acorn.lboro.ac.uk/>) is now available again. Others have just moved to a new URL, e.g. the HERON initiative now uses its own domain (<http://www.heron.ac.uk>) and information on M25 Link has moved to a different domain ([www.m25lib.ac.uk/M25link/](http://www.m25lib.ac.uk/M25link/)). However it was found that the given entry points for some project entry points (SEREN, ResIDe, Stacks and SKIP) or domains (NetLinks, ERIMS, DIAD and JEDDS) remained unavailable.

We, therefore, looked to see what could be found of these potentially problematic sites on the Internet Archive's Wayback Machine. Initial URLs for searching were taken from the descriptions of eLib projects available on the UKOLN Web pages (<http://www.ukoln.ac.uk/services/elib/projects/>).

To summarise our findings, most of the project Web pages had some presence in the Wayback Machine. One (SEREN) could not be found at all due to the use of robots.txt, while another (On-demand) only retrieved an archived 404-error page. The other project sites were at least partly available in the Wayback Machine, although in some cases there were not very many versions. The sites, on being checked, seemed to be fairly complete - at least at the very highest levels of directory structure. Most project sites were modest in size and tended to use simple HTML linking, sometimes with FRAMES. This meant that almost all home pages displayed OK (some without images) and that most internal links seemed to work. The very few sites that contained limited search facilities (e.g. ACORN) made unsuccessful attempts to contact servers outside the Internet Archive.

### ACORN: Access to Course Reading via Networks

Internet Archive: A search of the Wayback Machine retrieved 19 versions of this site, dating from December 1997 to May 2001. The earliest version of this displayed OK, including most images. All of the links that were tested also worked. The demonstration available in later versions of the pages worked until reaching the page that needed the entry of a test username and password. This then tried to link to the current version of the Web site.

URL searched for: <http://acorn.lboro.ac.uk/>

Versions of Web site evaluated:

<http://web.archive.org/web/19971221131204/http://acorn.lboro.ac.uk/>

<http://web.archive.org/web/20010516040726/http://acorn.lboro.ac.uk/>

**DIAD: Digitisation in Art and Design**

Internet Archive: A search of the Wayback Machine retrieved 7 versions of this site, dating from April 1997 to October 1999. The earliest versions of this displayed a 'directory' view of the Web site, giving easy access to backup and draft pages, and were missing some images. Later versions displayed OK (including FRAMES), but not all of the links tested worked.

URL searched for: <http://tdg.linst.ac.uk/tdg/research/diad/>

Versions of Web site evaluated:

<http://web.archive.org/web/19970422155842/http://tdg.linst.ac.uk/tdg/research/diad/>

<http://web.archive.org/web/19991003032214/http://tdg.linst.ac.uk/tdg/research/diad/>

**The Electronic Stacks Project**

Internet Archive: A search of the Wayback Machine retrieved 5 versions of this site, dating from February 1999 to November 2001. The earliest version evaluated (August 2000) displayed the main project page OK, but many of the links were to external sites (e.g. journals on the Oxford University Press site) and did not work. Some of these links, however, did work on later versions of the archived Web site, it even being possible to download PDF versions of journal articles.

URL searched for: <http://www.dcs.kcl.ac.uk/journals/stacks/>

Versions of Web site evaluated:

<http://web.archive.org/web/20000818184720/http://www.dcs.kcl.ac.uk/journals/stacks/>

<http://web.archive.org/web/20011101141704/http://www.dcs.kcl.ac.uk/journals/stacks/>

<http://web.archive.org/web/20010602073335/www3.oup.co.uk/igpl/current/>

**ERIMS: Electronic Readings in Management Studies**

Internet Archive: A search of the Wayback Machine retrieved just 2 versions of this site, dating from December 1998 to January 1999. Both versions displayed the home page OK, but some internal links had not been archived and some images were missing.

URL searched for: <http://erims.temp.ox.ac.uk/erimsproj/>

Versions of Web site evaluated:

<http://web.archive.org/web/19981207010928/http://erims.temp.ox.ac.uk/erimsproj/>

<http://web.archive.org/web/19990128084516/http://erims.temp.ox.ac.uk/erimsproj/>

**HERON: Higher Education Resources ON-demand**

Internet Archive: A search of the Wayback Machine retrieved 14 versions of this site, dating from February 1999 to November 2001. The earliest version of this that was evaluated (February 1999) displayed the home page OK, although some images (e.g. logos) were missing. Also, not all of the links that were checked worked properly. Most of these links did work OK on the later version of the site (January 2001) that was evaluated, although a few internal links further down the file hierarchy still would not work properly.

URL searched for: <http://www.stir.ac.uk/infoserv/heron/>

Versions of Web site evaluated:

<http://web.archive.org/web/19990224141334/http://www.stir.ac.uk/infoserv/heron/>

<http://web.archive.org/web/20010602034205/www.heron.ac.uk/>

### **JEDDS: Joint Electronic Document Delivery Software Project**

Internet Archive: A search of the Wayback Machine retrieved 3 versions of this site under the given URL, dating from September 2000 to February 2001. The earliest version of this page evaluated displayed reasonably well, but some images useful for navigation were missing. A later version (22 October 2001) retrieved a completely revised version of the Web site, but again images useful for navigation did not appear. This version of the site redirected users to the RLG Web site for downloading the Ariel software.

URL searched for: <http://jedds.mcc.ac.uk/website/index.html>

Versions of Web site evaluated:

<http://web.archive.org/web/20000914072318/http://jedds.mcc.ac.uk/website/index.html>

<http://web.archive.org/web/20010221195314/http://jedds.mcc.ac.uk/website/index.html>

### **M25 Link**

Internet Archive: No versions of the informational page once hosted by the London School of Economics (LSE) were available on the Wayback Machine. A search for the relevant page in the M25 Link domain (<http://www.m25lib.ac.uk/M25link/>) was slightly more successful, retrieving 3 versions of this site, dating from October 2000 to June 2001. The versions of this page that were evaluated displayed fine, while most the internal links (e.g. for project reports) worked OK. Naturally, the link to the prototype system tried (and failed) to connect to a server at LSE.

URL searched for: <http://www.lse.ac.uk/library/m25/>

Versions of Web site evaluated:

<http://web.archive.org/web/20010429014806/www.m25lib.ac.uk/M25link/>

<http://web.archive.org/web/20010617084310/www.m25lib.ac.uk/M25link/>

### **NetLinks: Collaborative Professional Development for Networked Learner Support**

Internet Archive: A search of the Wayback Machine retrieved 28 versions of this site, dating from March 1997 to April 2001. Early versions of the site evaluated displayed the home page and its images OK, and those internal links that were tested were working. The same was true of the later version of the page that was evaluated.

URL searched for: <http://netways.shef.ac.uk/>

Versions of Web site evaluated:

<http://web.archive.org/web/19970329005608/http://netways.shef.ac.uk/>

<http://web.archive.org/web/20010301223218/http://netways.shef.ac.uk/>

### **On-demand Publishing in the Humanities**

Internet Archive: A search of the Wayback Machine retrieved 4 versions of this site, all from 2001. Those linked to, however, displayed an archived HTTP 404 error message: "The web document you requested could not be found (Error: 404b)."

URL searched for: [http://cwis.livjm.ac.uk/on\\_demand/](http://cwis.livjm.ac.uk/on_demand/)

Versions of Web site evaluated:

[http://web.archive.org/web/20010502162916/http://cwis.livjm.ac.uk/on\\_demand/](http://web.archive.org/web/20010502162916/http://cwis.livjm.ac.uk/on_demand/)

[http://web.archive.org/web/20011101153853/http://cwis.livjm.ac.uk/on\\_demand/](http://web.archive.org/web/20011101153853/http://cwis.livjm.ac.uk/on_demand/)

### **PPT: Parallel Publishing for Transactions**

Internet Archive: A search of the Wayback Machine retrieved 7 versions of this site, dating from April 1997 to February 1999. The earliest version of the site that was evaluated displayed the home page and its images OK, and all of those internal links that were tested worked OK. It was even possible to retrieve PDF versions of the articles published in a sample issue of the journal. Later versions of the Web site in the Wayback Machine also worked OK, but it appeared that PDF versions of articles were no longer available.

URL searched for: [http://ppt.geog.qmw.ac.uk/tibg/ppt\\_hom.html](http://ppt.geog.qmw.ac.uk/tibg/ppt_hom.html)

Versions of Web site evaluated:

[http://web.archive.org/web/19970405230130/http://ppt.geog.qmw.ac.uk/tibg/ppt\\_hom.html](http://web.archive.org/web/19970405230130/http://ppt.geog.qmw.ac.uk/tibg/ppt_hom.html)

[http://web.archive.org/web/19990203172916/http://ppt.geog.qmw.ac.uk/tibg/ppt\\_hom.html](http://web.archive.org/web/19990203172916/http://ppt.geog.qmw.ac.uk/tibg/ppt_hom.html)

### **ResIDe: Electronic reserve for UK Universities**

Internet Archive: A search of the Wayback Machine retrieved 7 versions of this site, dating from April 1997 to August 2000. Those versions of the Web site that were evaluated worked very well, and all of the internal links that were followed worked OK.

URL searched for: <http://www.uwe.ac.uk/library/itdev/reside/>

Versions of Web site evaluated:

<http://web.archive.org/web/19970428053749/http://www.uwe.ac.uk/library/itdev/reside/>

<http://web.archive.org/web/20000815111109/http://www.uwe.ac.uk/library/itdev/reside/>

### **SEREN: Sharing of Educational Resources in an Electronic Network in Wales**

Internet Archive: Access to [seren.newi.ac.uk](http://seren.newi.ac.uk) was blocked by robots.txt.

URL searched for: <http://seren.newi.ac.uk/>

**SKIP: Skills for new Information Professionals**

Internet Archive: A search of the Wayback Machine retrieved 20 versions of this site, dating from July 1997 to June 2001. Those versions of the Web site that were evaluated worked very well, and the majority of internal links followed worked OK.

URL searched for: <http://www.plym.ac.uk/faculties/research/skip1.htm>

Versions of Web site evaluated:

<http://web.archive.org/web/19970709011859/http://www.plym.ac.uk/faculties/research/skip1.htm>

<http://web.archive.org/web/20000504204938/http://www.plym.ac.uk/faculties/research/skip1.htm>

**TAPin: Training and Awareness Programme in networks**

Internet Archive: A search of the Wayback Machine retrieved 4 versions of this site, dating from March 1997 to February 1999. Early versions of the site were quite small (and completely HTML based) and all links followed worked OK. The latest version of the site evaluated, however, was missing images and some key pages.

URL searched for: <http://www.uce.ac.uk/tapin/>

Versions of Web site evaluated:

<http://web.archive.org/web/19970307052158/http://www.uce.ac.uk/tapin/>

<http://web.archive.org/web/19990203160838/http://www.uce.ac.uk/tapin/>

## Appendix C: The World Wide Web

Since its inception in the early 1990s, the World Wide Web has become a pervasive means of communication for all kinds of information resources. The Web has often been compared to a huge encyclopaedia (or library) but it is also the facilitator of more informal communication and a gateway to commercial enterprise. Because of this and the general ease of publishing information on the Web, a large proportion of the information that makes up the Web is considered to be ephemeral or is of unknown quality or provenance. It also tends to be very dynamic, with new services and Web pages appearing all the time, while older ones are constantly updated, restructured or deleted. The hypertextual nature of the Web also means that an individual document rarely stands alone but is normally part of an intricate network of linked documents or services. These factors make the Web a very difficult object to preserve.

### Background

The World Wide Web Consortium (W3C) simply defines the Web as "the universe of network-accessible information, the embodiment of human knowledge" (<http://www.w3.org/WWW/>). It originated from a project based at CERN, the European particle research laboratory based in Geneva, first as a way of managing its own organisational knowledge and then as a scalable means of sharing information with scientists working elsewhere in the world (Berners-Lee, *et al.*, 1994, p. 76). The team at CERN developed three key things that made the World Wide Web possible. The first was a means of specifying each machine connected to the Internet originally called the Universal Resource Identifier (URI) but later known as the Uniform Resource Locator (URL). The second development was a protocol for information transfer called the Hypertext Transfer Protocol (HTTP). The third was a uniform way of structuring hypertext documents with links - the Hypertext Markup Language (HTML), an implementation of the Standard Generalized Markup Language (SGML).

The World Wide Web was 'released' to the wider world in 1991 when CERN made its line-mode Web browser available by Anonymous FTP (Naughton, 2000, p. 241). This was followed by the development of other experimental browsers and in 1993 by the release of the graphics-based browser 'Mosaic,' developed at the US National Center for Supercomputing Applications (NCSA). The availability of Mosaic changed people's perceptions of the Web - which until then had been just one of several competing systems designed for sharing information (including WAIS and Gopher). Gillies & Cailliau (2000, p. 233) have called it the Web's 'killer application' - "the one that definitively dragged the Web out of academia and dropped it centre stage." From then on, the use of the Web began to grow at a very rapid rate, as did the development of new Web browser programs. Some of Mosaic's developers left NCSA to set up Netscape and in 1994 released the Netscape Navigator browser. In the same year Microsoft bought the company that had licensed Mosaic from NCSA and launched the Internet Explorer browser. The rapid growth of the Web (especially in the commercial sector) and the tensions raised by the 'browser wars' threatened its possible fragmentation. Already, different browser-specific extensions of HTML had begun to be defined, including the notorious <BLINK> tag that was supported by the Netscape browser (Gillies & Cailliau, 2000, p. 260). The threat of fragmentation eventually led to the formation of the W3C in late 1994 in order to co-ordinate the development and definition of the Web standards and protocols. In its own words, the purpose of the Consortium is "to lead the World Wide Web to its full potential by developing common protocols that promote its evolution and ensure its interoperability" (<http://www.w3.org/Consortium/>).

The Web from its earliest stages was not just made up of hyperlinked HTML pages. Some textual documents were made available in other formats, e.g. TeX, PostScript or Adobe's Portable Document Format (PDF). Inline images were stored separately in GIF format or as files using JPEG compression. In a relatively short time, however, the Web became the home of a much broader range of formats, for audio, still and moving images, multimedia, etc. Their adoption led to the increasing use of browser 'plug-ins,' (e.g. for multimedia) and scripting

technologies. Growing proportions of Web sites are generated 'on-the-fly' from content management systems or use back-end databases. Web content is increasingly hidden from crawler-based Web search services, leading to a growth in what is now known as the 'deep Web.'

The rapid growth of the Web also had a major impact on the content of the Web. It had been designed to support the sharing of information between scientists and most of the earliest Web sites were broadly academic in nature. After a short time, this began to change. For example, Chakrabarti, *et al.* (1999, p. 44) give an indication of the Web's intellectual range at the end of the 1990s.

*... because of the Web's rapid, chaotic growth, the resulting network of information lacks organization and structure. In fact, the Web has evolved into a global mess of previously unimagined proportions. Web pages can be written in any language, dialect or style by individuals with any background, education, culture, interest and motivation. Each page might range from a few characters to a few hundred thousand, containing truth, falsehood, wisdom, propaganda or sheer nonsense.*

This chapter will now look at various attempts to characterise and quantify the World Wide Web. The large and dynamic nature of the Web makes it a potentially difficult object to preserve.

### **The size of the publicly available Web**

Since its inception, the size and nature of the World Wide Web has become a subject for discussion and debate. For example, there have been various estimates made of its size and rate of growth. An early instance of this was the presentation of some statistics from the Open Text Index search engine at the 5th World Wide Web Conference. Although he acknowledged that very little was known about the Web at that time, Bray (1996, p. 995) used a snapshot taken in late 1995 to estimate that there were at that time over eleven million unique URLs and almost 224 thousand unique servers. Figures collected since then suggest a consistent year on year growth. In a 1998 article published in *Science*, researchers based at the NEC Research Institute in Princeton estimated that, as of December 1997, the size of the publicly indexable Web was at least 320 million pages (Lawrence & Giles, 1998). A follow up paper published in *Nature* the following year estimated that by February 1999, the same subset of the Web contained around 800 million pages on about three million servers (Lawrence & Giles, 1999). From an analysis of a random sample, the researchers observed that individual search engines only indexed a relatively small proportion of these sites. For example they estimated that the largest search engine (then, Northern Light) only indexed around sixteen per-cent of the publicly indexable Web. However, because there was (at that time) little overlap between search engines, the combined index from eleven search engines would have covered around about 42% of the indexable Web. From manually classifying their sample, they also noted that only around six per-cent of Web servers had scientific or educational content, concluding that "an index of all scientific information on the web would be feasible and very valuable."

**Table C.1: Number of unique Web sites, 1998-2002**

| Year | No. of unique Web sites |
|------|-------------------------|
| 1998 | 2,636,000               |
| 1999 | 4,662,000               |
| 2000 | 7,128,000               |
| 2001 | 8,443,000               |
| 2002 | 9,040,000               |

Source: OCLC Web Characterization Project (<http://wcp.oclc.org/>)

Other studies of Web size show a similar growth rate. The Web Characterization Project of the OCLC Research conducts an annual Web sample to analyse trends in the size and content of the Web (<http://wcp.oclc.org/>). This counts Web sites by sampling one per-cent of IP addresses and attempting to connect to each one. Duplicate sites are then subtracted to identify the total number of unique Web sites (Table C.1).

A survey undertaken by the company Cyveillance in 2000 suggested that there were over two billion unique publicly available Web pages and estimated that the Web would double in size by early 2001 (Murray & Moore, 2000). Lyman & Varian (2000) collated these and other figures and concluded that the total amount of information on the 'surface' Web was somewhere between 25 and 50 terabytes as of 2000.

### The 'deep Web'

However, these figures do not tell the whole story. In 2001, Bar-Ilan (2001, p. 9) concluded that studies attempting to characterise the 'whole Web' were no longer feasible.

*... the current estimated size of the indexable Web is around 2,200 million pages. What is meant by "indexable"? These are static pages, freely accessible by search engines and Web users ... There are lots of other pages on the Web, a large proportion of them are dynamically created pages which are based on querying some database not residing on the Web and being accessed through search forms.*

A white paper produced by the search company BrightPlanet (Bergman, 2001) has estimated that this part of the Web - sometimes known as the invisible, hidden or deep Web - may be up to 400 to 500 times bigger than the indexable (or surface) Web. The deep Web is made up of sites or resources that are not routinely indexed by search engines. This is sometimes due to the technological limitations of crawler-based search engines. Until relatively recently, many of these did not index well-used formats like PDF or Microsoft Word. Sometimes, inaccessibility is the responsibility of the target site, e.g. through password protection or use of the robots exclusion protocol. A bigger problem, however, has been the recent growth of Web sites that are dynamically served - e.g., through things like Microsoft's ASP - and the incorporation of large databases on the Web. Many of these databases actually predate the development of the Web itself, but with the development of Web-based interfaces, are now available to the wider world. A column in IEEE's *Computer* magazine in 1999 noted that many databases remained on mainframe computers (Lewis, 1999).

*So much time and effort have gone into building these mainframe databases that it would cost many companies more than their total worth to convert the databases to another format.*

BrightPlanet's 2000 survey attempted to identify some of the largest deep Web sites (Bergman, 2001). The largest public sites at that time were the National Climatic Data Center

of the US National Oceanic and Atmospheric Administration (NOAA) and NASA's Earth Observing System Data and Information System (EOSDIS). Other large deep Web sites included the products of key digital library projects (e.g., the Infromedia Digital Video Library, the Alexandria Digital Library, the UC Berkeley Digital Library, JSTOR), bibliographic databases (e.g., PubMed), databases of scientific data (e.g. GenBank) and library catalogues. The largest fee-based sites included database services for law and business (Lexis-Nexis, Dun & Bradstreet, etc.), bibliographic databases (e.g., Ovid, INSPEC) and e-journal services provided by scholarly publishers (e.g., Elsevier, EBSCO, Springer-Verlag).

The BrightPlanet white paper noted that it was not feasible for the deep Web to be searchable via a single database.

*It is infeasible to issue many hundreds of thousands or millions of direct queries to individual deep Web search databases. It is implausible to repeat this process across tens to hundreds of thousands of deep Web sites. And, of course, because content changes and is dynamic, it is impossible to repeat this task on a reasonable update schedule.*

The same difficulties would also apply to some Web-preservation initiatives; in particular those based on harvesting technology. Much database-driven Web information will be as invisible to Web harvesting robots as they are to the existing generation of search engines.

## The dynamic Web

Another important characteristic of the Web as it now exists is its dynamic nature. This is one reason why its preservation is such a difficult problem. Web sites appear and disappear, are updated and restructured. These factors lead to the Web's 'broken-link' problem, symbolised by the well-known HTTP Error 404 -- Not Found. Lawrence, *et al.* (2001, p. 30) cite an Alexa Internet (<http://www.alexa.com/>) estimate that Web pages disappear after an average time of 75 days. In addition, URLs are partly based on Internet domain names and these will sometimes disappear or change hands. For example, in 1999 the domain name of a popular (and much-linked) HTML validation site lapsed and was briefly taken over by a Web site with adult content (Kelly, 1999).

The Web's dynamism is one of the reasons for its success, but it means that at any one time, a large proportion of all links on the Web will not work or will link to the wrong Web site. This causes a particular problem with references in scientific research. Lawrence, *et al.* (2001, p. 28) identified several reasons why URLs become invalid.

*First, personal homepages tend to disappear when researchers move.  
Second, many who restructure Web sites fail to maintain old links.*

They recommend that Web citations need to include enough context information so that users can find the latest location using search engines; also that authors should place materials in centralised repositories like e-print archives. There are several other proposed solutions to the Web's 'broken-link' problem. Ashman (2000) outlines some of these including the use of versioning, dereferencing through some kind of resolver service (e.g., IETF's Uniform Resource Names proposal) and redirection.

## The 'small-world' Web

The structure of the World Wide Web has also become a topic of study in a branch of physics known as statistical mechanics. Despite its large size, studies of the nodes (Web pages) and edges (hyperlinks) in the Web reveal that it demonstrates the characteristics of what are now known as 'small-world networks.' These were first identified in social networks, e.g., the phenomenon popularly known as 'six degrees of separation,' (Milgram, 1967) and mathematically articulated in a much-cited paper by Watts & Strogatz (1998). Small-world networks are characterised by graphs that include a high degree of clustering but with a short

minimum distance between randomly chosen nodes (Adamic, 1999, p. 443). As well as the World Wide Web, many other networks have been identified as displaying small-world characteristics. These include electrical power grids, the nervous system of the nematode worm *Caenorhabditis elegans*, food webs in ecosystems and co-authorship patterns of scientists (e.g., Strogatz, 2001; Newman, 2001).

In the case of the Web, Adamic (1999) studied the average number of links to traverse between Web sites and discovered that it displayed the characteristics of a small-world network. In a separate study, Albert, Jeong & Barabási (1999) concluded that the distance between any two Web documents is around about 12. The results were summarised in a later paper by Barabási (2001, p. 35).

*These results clearly indicated that the Web represents a small world, i.e. the typical number of clicks between two Web pages is about 19, despite the fact that there are now over one billion pages out there.*

These studies also showed that the World Wide Web, the Internet and other networks were 'scale-free,' in that they evolved through time through the addition and removal of nodes and links.

Subsequent studies by computer scientists have suggested that the whole picture may be more complex than some of the small-world network proponents have suggested. An analysis by Broder *et al.* (2000) of large samples of the Web showed that although there was a highly connected Web core, there were equally large areas that were not as well linked. A news item in *Nature* (2000) summarised their conclusions.

*A central core contains pages between which users can surf easily. Another large cluster, labelled 'in', contains pages that link to the core but cannot be reached from it. These are often new pages that have not yet been linked to. A separate 'out' cluster consists of pages that can be reached from the core but do not link to it, such as corporate websites containing only internal links. Other groups of pages, called 'tendrils' and 'tubes', connect to either the in or out clusters, or both, but not to the core, whereas some pages are completely unconnected*

The resulting diagram looked like a bow tie with tendrils and tubes attached to the 'wings.' Accordingly, the press release that accompanied the announcement of these results referred to it as the "Bow Tie" Theory (IBM Research, 2000). The consequences of this for search engines is that they would need to crawl from a wide range of starting points in order to ensure breadth of coverage (Butler, 2000, p. 112).

Albert & Barabási (2002) and Dorogovtsev & Mendes (2002) have published detailed reviews of recent work in this branch of physics. Barabási (2002), Buchanan (2002) and Huberman (2001) have produced more popular accounts of small-world networks and outline their relevance to the World Wide Web and other networks.

These discoveries may have some practical uses. For example, the 'scale-free' nature of the Internet means that it is resistant to random attack, e.g. by hackers or computer viruses. Barabási (2001, p. 37) argued that the "random removal of nodes is most likely to affect small nodes rather than hubs with many links because nodes significantly outnumber hubs." Simultaneously, however, scale-free networks like the Internet **are** vulnerable to intentional attacks focused on a relatively small number of those 'hub' sites that have many links (Cohen, *et al.*, 2001).

Small-world networks may also have practical applications on the Web itself. Second-generation search services have already begun to make use of the link structure of the Web to help support searching and the ranking of results. The popular search service Google (<http://www.google.com/>) ranks Web pages according to the number and nature of incoming links (Brin & Page, 1998). Similar ideas have been implemented in the experimental Clever

search engine (<http://www.almaden.ibm.com/cs/k53/clever.html>) developed by IBM Research. Like Google, this uses the link structure of the Web to help discover and rank the most 'authoritative' pages on search topics (Kleinberg, 1999, p. 605). In addition, it also identifies 'hub' sites - Web pages that link to many 'authorities' and, therefore, provide good starting points for browsing. Chakrabarti, *et al.*, (1999b, p. 61) say that hub pages "appear in a variety of forms, ranging from professionally assembled resource lists on commercial sites to lists of recommended links on individual home pages." The Clever research team noted that a major difference between Google and Clever was the latter's ability to look backwards from an authoritative page to see which other pages link to it. They argued that the system took advantage "of the sociological phenomenon that humans are innately motivated to create hublike content expressing their expertise on specific topics" (Chakrabarti, *et al.*, 1999a, p. 52).

More recently, Flake, *et al.* (2002) have developed a new search algorithm that extracts meaning from the link structure of the Web to help identify Web communities. A related area where the investigation of Web link structure may have some applications is in the measurement of Web bibliometrics or impact factors. Those developing search services like Google and Clever have acknowledged their close ties with citation analysis (e.g., Chakrabarti, *et al.*, 1999a, pp. 50-51).

### The Semantic Web

The Semantic Web (<http://www.w3c.org/2001/sw/>) is the vision, most powerfully articulated by Tim Berners-Lee of the W3C, of an extension of the current Web in which information is given well-defined meaning so that machines can begin to understand it, and process it accordingly. This is not achieved through advanced AI techniques, but by relying "solely on the machine's ability to solve well-defined problems by performing well-defined operations on well-defined data" (Berners-Lee & Hendler, 2001, p. 1023). This means that content creators will need to use new languages that will make Web content understandable to machines. It's supporters note that the "challenge of the Semantic Web is to provide a language that expresses both data and rules for reasoning about the data and that allows rules from any existing knowledge-representation system to be exported onto the Web" (Berners-Lee, Hendler & Lassila, 2001).

Some of the technical building blocks for the Semantic Web are already in place. Semantic Web standards developed by the W3C include the Extensible Markup Language (XML), XML Schema, the Resource Description Framework (RDF) and RDF Schema. XML helps to add structure to documents, while RDF provides a simple data model (based on triples) and an XML-based syntax for the application of metadata about a resource. RDF differs from XML in that it uses Universal Resource Identifiers (URIs) to unambiguously denote objects and the properties of relationships between objects (Decker, Mitra & Melnik, 2000). The RDF Schema specification provides a mechanism to define properties and the relationships between these properties and other resources (Brickley & Guha, 2000).

### References

Adamic, L.A. & Huberman, B.A. (2001). "The Web's hidden order." *Communications of the ACM*, 44 (9), 55-59.

Adamic, L.A. (1999). "The small world web." In: Abiteboul, S. & Vercoistre, A.-M., eds., *Research and advanced technology for digital libraries: third European conference, ECDL'99, Paris, France, September 22-24, 1999*. Lecture Notes in Computer Science, 1696. Berlin: Springer, 443-452. Also available at: <http://www.hpl.hp.com/shl/papers/smallworld/>

Adamic, L.A. (2001). *Network dynamics: the World Wide Web*. PhD Thesis, Stanford University, Department of Applied Physics. Available at: [http://www.hpl.hp.com/shl/people/ladamic/thesis/ladamic\\_thesis.pdf](http://www.hpl.hp.com/shl/people/ladamic/thesis/ladamic_thesis.pdf)

- Albert, R. & Barabási, A.-L. (2002). "Statistical mechanics of complex networks." *Reviews of Modern Physics*, 74, 47-97.
- Albert, R., Jeong, H. & Barabási, A.-L. (1999). "Diameter of the World-Wide Web." *Nature*, 401, 130-131.
- Aloisio, G., Cafaro, M., Kesselman, C. & Williams, R. (2001). "Web access to supercomputing." *Computing in Science and Engineering*, 3 (6), 66-72.
- Arasu, A., Cho, J., Garcia-Molina, H., Paepke, A. & Raghavan, S. (2000). "Searching the Web." Technical report, Stanford University, Department of Computer Science. Available at: <http://dbpubs.stanford.edu/pub/2000-37>
- Ashman, H. (2000). "Electronic document addressing: dealing with change." *ACM Computing Surveys*, 32 (3), 201-212.
- Barabási, A.-L. (2001). "The physics of the Web." *Physics World*, 14 (7), 33-38.
- Barabási, A.-L. (2002). *Linked: the new science of networks*. Cambridge, Mass.: Perseus.
- Bar-Ilan, J. (2001). Data collection methods on the Web for informetric purposes: a review and analysis. *Scientometrics*, 50 (1), 7-32.
- Bergman, M.K. (2001). "The deep Web: surfacing hidden value." *Journal of Electronic Publishing*, 7 (1), August. Available at: <http://www.press.umich.edu/jep/07-01/bergman.html>
- Berners-Lee, T. & Hendler, J. (2001). "Publishing on the Semantic Web." *Nature*, 410 (6832), 26 April, 1023-1024.
- Berners-Lee, T., Cailliau, R., Luotonen, A., Nielsen, H.F. & Secret, A. (1994). "The World-Wide Web." *Communications of the ACM*, 37 (8), 76-82.
- Berners-Lee, T., Hendler, J. & Lassila, O. (2001). "The Semantic Web." *Scientific American*, 284 (5), May, 28-37. Available at: <http://www.sciam.com/2001/0501issue/0501bernbers-lee.html>
- Bollacker, K.D., Lawrence, S. & Giles, C.L. (2000). "Discovering relevant scientific literature on the Web." *IEEE Intelligent Systems*, 15 (2), 42-47.
- Bray, T. (1996). "Measuring the Web." *Computer Networks and ISDN Systems*, 28, 993-1005. Also published in the proceedings of the 5th International World Wide Web Conference, Paris, France, 6-10 May 1996. Available at: [http://www5conf.inria.fr/fich\\_html/papers/P9/Overview.html](http://www5conf.inria.fr/fich_html/papers/P9/Overview.html)
- Brickley, D. & Guha, R.V., eds. (2000). *Resource Description Framework (RDF) Schema specification, 1.0*. W3C Candidate Recommendation, 27 March. Available at: <http://www.w3.org/TR/2000/CR-rdf-schema-20000327/>
- Brin, S. & Page, L. (1998). "The anatomy of a large-scale hypertextual Web search engine." *Computer Networks and ISDN Systems*, 30 (1-7), 107-117. Full version published in the proceedings of the 7th International World Wide Web Conference, Brisbane, Australia, 14-18 April 1998. Available at: <http://www7.scu.edu.au/programme/fullpapers/1921/com1921.htm>
- Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A. & Wiener, J. (2000). "Graph structure in the Web." *Computer Networks*, 33 (1-6), 309-320. Also published in the proceedings of the 9th International World Wide Web Conference, Amsterdam, Netherlands, 15-19 May 2000. Available at: <http://www9.org/w9cdrom/160/160.html>

- Brunton, M. (2001). "Illuminating the Web." *Time Magazine*, 158 (2), 9 July, 52-54. Available at: <http://www.time.com/time/europe/biz/magazine/0,9868,166169,00.html>
- Butler, D. (2000). "Souped-up search engines." *Nature*, 405, 112-115.
- Chakrabarti, S., Dom, B., Raghavan, P., Rajagopalan, S., Gibson, D. & Kleinberg, J. (1998). "Automatic resource compilation by analyzing hyperlink structure and associated text." *Computer Networks and ISDN Systems*, 30 (1-7), 65-74. Also published in the proceedings of the 7th International World Wide Web Conference, Brisbane, Australia, 14-18 April 1998: Available at: <http://www7.scu.edu.au/programme/fullpapers/1898/com1898.html>
- Chakrabarti, S., Dom, B.E., Kumar, S.R., Raghavan, P., Rajagopalan, S., Tomkins, A., Kleinberg, J. & Gibson, D. (1999a). "Hypersearching the Web." *Scientific American*, 280 (6), 44-52.
- Chakrabarti, S., Dom, B.E., Kumar, S.R., Raghavan, P., Rajagopalan, S., Tomkins, A., Gibson, D. & Kleinberg, J. (1999b). "Mining the Web's link structure." *Computer*, 32 (8), 60-67.
- Chervenak, A., Foster, I., Kesselman, C., Salisbury, C. & Tuecke, S. (2000). "The data grid: towards an architecture for the distributed management and analysis of large scientific datasets." *Journal of Network and Computer Applications*, 23, 187-200.
- Cohen, R., Erez, K., ben-Avraham, D. & Havlin, S. (2001). "Breakdown of the Internet under intentional attack." *Physical Review Letters*, 86 (16), 3682-3685.
- Cole, T.W., Kaczmarek, J., Marty, P.F., Prom, C.J., Sandore, B. & Shreeves, S. (2002). "Now that we've found the 'hidden Web,' what can we do with it?" Museums and the Web 2002, Boston, Mass., 17-20 April 2002. Available at: <http://www.archimuse.com/mw2002/papers/cole/cole.html>
- De Roure, D., Jennings, N.R. & Shadbolt, N.R. (2002). "The Semantic Grid: a future e-Science infrastructure." In: Berman, F., Fox, G. & Hey, T., eds., *Grid computing: making the global infrastructure a reality*. Chichester: Wiley (forthcoming).
- Decker, S., Mitra, P. & Melnik, S. (2000). "Framework for the Semantic Web: an RDF tutorial." *IEEE Internet Computing*, 4 (6), November/December, 68-73.
- Dorogovtsev, S.N. & Mendes, J.F.F. (2002). "Evolution of networks." *Advances in Physics*, 51 (4), 1079-1187.
- Flake, G.W., Lawrence, S., Giles, S.L. & Coetzee, F.M. (2002). "Self-organization and identification of Web communities." *Computer*, 35 (3), 66-71.
- Foster, I. & Kesselman, C. (1999a). "Computational grids." In: Foster, I. & Kesselman, C. (eds.) *The Grid: blueprint for a new computing infrastructure*. San Francisco, Calif.: Morgan Kaufmann, 15-51.
- Foster, I. & Kesselman, C. (1999b). "The Globus toolkit." In: Foster, I. & Kesselman, C. (eds.) *The Grid: blueprint for a new computing infrastructure*. San Francisco, Calif.: Morgan Kaufmann, 259-278.
- Foster, I. (2002). "The Grid: a new infrastructure for 21st century science." *Physics Today*, 55 (2), 42-47.
- Foster, I., Kesselman, C. & Tuecke, S. (2001). "The anatomy of the Grid: enabling scalable virtual organisations." *International Journal of High Performance Computing Applications*, 15 (3), 200-222.

- Foster, I., Kesselman, C., Nick, J.M. & Tuecke, S. (2002). "Grid services for distributed system integration." *Computer*, 35 (6), 37-46.
- Gillies, J. & Cailliau, R. (2000). *How the Web was born: the story of the World Wide Web*. Oxford: Oxford University Press.
- Glover, E.J., Lawrence, S., Gordon, M.D., Birmingham, W.P. & Giles, S.L. (2001). "Web search - your way." *Communications of the ACM*, 44 (12), December, 97-102.
- Hey, T. & Trefethen, A. (2002). "The data deluge: an e-Science perspective." In: Berman, F., Fox, G. & Hey, T., eds., *Grid computing: making the global infrastructure a reality*. Chichester: Wiley (forthcoming).
- Huberman, B.A. (2001). *The laws of the Web: patterns in the ecology of information*. Cambridge, Mass.: MIT Press.
- IBM Research. (2000). *Researchers map the Web - press release*. San Jose, Calif.: IBM Almaden Research Center, 11 May. Available at: [http://www.almaden.ibm.com/almaden/webmap\\_release.html](http://www.almaden.ibm.com/almaden/webmap_release.html)
- Kelly, B. (1999). "News article: are you linking to a porn site?" *Exploit Interactive*, 1, April. Available at: <http://www.exploit-lib.org/issue1/webtechs/>
- Kelly, B. (2001). "Web focus: hot news from WWW10." *Ariadne*, 28, June. Available at: <http://www.ariadne.ac.uk/issue28/web-focus/>
- Kleinberg, J. & Lawrence, S. (2001). "The structure of the Web." *Science*, 294, 1849-1850.
- Kleinberg, J. (1999). "Authoritative sources in a hyperlinked environment." *Journal of the ACM*, 46(5), 604-632.
- Laszewski, G. von (2002). "Grid computing: enabling a vision for collaborative research." In: Fagerholm, J., Haataja, J., Järvinen, J., Lyly, M., Råback, P. & Savolainen, V., eds., *Applied Parallel Computing: Advanced Scientific Computing: 6th International Conference, PARA 2002, Espoo, Finland, June 15-18, 2002*. Lecture Notes in Computer Science, 2367. Berlin: Springer, 37-50.
- Lawrence, S. & Giles, C.L. (1998). "Searching the World Wide Web." *Science*, 280, 98-100.
- Lawrence, S. & Giles, C.L. (1999a). "Searching the Web: general and scientific information access." *IEEE Communications Magazine*, 37 (1), 116-122.
- Lawrence, S. & Giles, C.L. (1999b). "Accessibility of information on the Web." *Nature*, 400, 107-109.
- 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), February, 26-31.
- Leung, S.-T. A., Perl, S.E., Stata, R. & Wiener, J.L. (2001). *Towards Web-scale Web archaeology*. SRC Research Report, 174. Palo Alto, Calif.: Compaq Systems Research Center. <ftp://gatekeeper.research.compaq.com/pub/DEC/SRC/research-reports/SRC-174.pdf>
- Lewis, T. (1999). "Mainframes are dead, long live mainframes." *Computer*, 32 (8), 102-104.
- Lyman, P. & Varian, H.R. (2000). *How much information?* Berkeley, Calif.: University of California at Berkeley, School of Information Management and Systems. Available at: <http://www.sims.berkeley.edu/research/projects/how-much-info/internet.html>

- Milgram, S. (1967). "The small world problem." *Psychology Today*, 1, 60-67.
- Moore, R.W., Baru, C., Marciano, R., Rajasekar, A. & Wan, M. (1999), "Data-intensive computing." In: Foster, I. & Kesselman, C. (eds.) *The Grid: blueprint for a new computing infrastructure*. San Francisco, Calif.: Morgan Kaufmann, 105-129.
- Murray, B. & Moore, A. (2000) *Sizing the Internet*. Cyveillance White Paper, July. Available at: [http://www.cyveillance.com/web/downloads/Sizing\\_the\\_Internet.pdf](http://www.cyveillance.com/web/downloads/Sizing_the_Internet.pdf)
- Nature. (2000). "The Web is a bow tie." *Nature*, 405, 113.
- Newman, M.E.J. (2001). "The structure of scientific collaboration networks." *Proceedings of the National Academy of Sciences of the United States of America*, 98 (2), 404-409. Available from PubMed Central: <http://www.pubmedcentral.nih.gov/>
- Pennock, D.M., Flake, G.W., Lawrence, S., Glover, E.J. & Giles, C.L. (2002). "Winners don't take all: characterizing the competition for links on the Web." *Proceedings of the National Academy of Sciences of the United States of America*, 99 (8), 16 April, 5207-5211.
- Raghavan, P. (2002). "Social networks: from the Web to the enterprise." *IEEE Internet Computing*, 6 (1), 91-94.
- Raghavan, S. & Garcia-Molina, H. (2001). "Crawling the hidden Web." In: *VLDB 2001: Proceedings of 27th International Conference on Very Large Data Bases, September 11-14, 2002, Roma, Italy*. San Francisco, Calif.: Morgan Kaufmann. Available at: [http://www.dia.uniroma3.it/~vldbproc/017\\_129.pdf](http://www.dia.uniroma3.it/~vldbproc/017_129.pdf)
- Raghavan, S. & Garcia-Molina, H. (2002). "Representing Web graphs." Technical report, Stanford University, Department of Computer Science. Available at: <http://dbpubs.stanford.edu/pub/2002-30>
- Sherman, C. & Price, G. (2001). *The invisible Web: uncovering information sources search engines can't see*. Medford, N.J.: CyberAge Books.
- Strogatz, S. (2001). "Exploring complex networks." *Nature*, 410, 268-276.
- Thelwall, M. (2001a). "Exploring the link structure of the Web with network diagrams." *Journal of Information Science*, 27 (6), 393-401.
- Thelwall, M. (2001b). "Extracting macroscopic information from Web links." *Journal of the American Society for Information Science and Technology*, 52 (13), 1157-1168.
- Watts, D.J. & Strogatz, S.H. (1998). "Collective dynamics of 'small-world' networks." *Nature*, 393, 440-442.

## Appendix D: The UK Web space

By Brian Kelly, *UK Web Focus*, UKOLN, University of Bath

[A longer version of this can be found in: Brian Kelly, "Approaches to the preservation of Web sites," Online Information 2002, Olympia, London, 3-5 December 2002. Available at: <http://www.ukoln.ac.uk/web-focus/events/conferences/online-information-2002/>].

The size and nature of the UK Web is as difficult to quantify as the World Wide Web itself. However, some information can be of use when considering the approaches that can be taken to preserving UK Web space, in particular in making financial estimates. In addition it may be useful to examine the different approaches which have been taken to measuring the size of UK Web space as the different approaches, and the limitations which may be identified, may provide information which may help inform a preservation strategy.

### Definitions

The first issue to consider is what is meant by the "UK Web space". It could have several different meanings:

- Web sites with a .uk domain name.
- Web sites that are physically hosted in the UK.
- Web sites that are owned by an UK body.
- Web sites in which the intellectual content is owned by an UK body or British citizen.
- Web sites in which the intellectual content is of particular interest to the UK.

We can easily identify difficulties in making use of each of these definitions:

- Using the .uk domain name: this approach will fail to identify 'UK Web sites' that make use of .com, .org, etc. domain names.
- Using the physical location of the Web server: the location of Web servers may be based on hosting costs; server with identical content may be replicating around the world; the physical location may bear no relationships with the ownership of intellectual content of the Web server.
- Using the ownership of the Web site: it may be difficult to establish the ownership.
- Using the ownership of the intellectual content: it will be difficult to establish the ownership of the intellectual content.

As an example, consider the **english-parliament-news** YahooGroups mailing list and Web site (<http://groups.yahoo.com/group/english-parliament-news/>). The Web site contains the mailing list archives for campaigners for an English parliament. However, the Web site is based in US, and has a .com domain. In some respects this should be treated as a US Web site. However the intellectual content is likely to be created primarily by British citizens and the content is of particular relevance to the UK.

Note that as well as defining "UK Web space" we also need to define what we mean by a "Web site". Do we mean a physical Web server? This is a simple definition, but does not take into account the widespread usage of multiple virtual Web servers from one physical Web

server. In addition, it should be noted that end users may regard areas of a Web site which are maintained by different individuals, departments, etc. as separate Web sites.

## Approaches To Measurement

How might we approach the measurement of UK Web space, based on the definitions given above?

The simplest approach, based on the .uk domain name definition, is to make use of the Internet's DNS (Domain Name Service). However this will tell us about domain names which have been allocated, but not necessarily used.

Another approach is to make use of data collected by automated robot software, such as search engine and auditing robots. This will provide information on existing Web sites which are publicly available.

## Findings – Numbers of Web Sites

Whilst acknowledging the difficulties of measuring UK Web site, an initial attempt has been made using a number of approaches.

**Netcraft** (<http://www.netcraft.com>) is a Web server monitoring company based in Bath, UK. Its robot software continually analyses Web sites in order to report on Web server software usage. Netcraft has a Web-based interface to their data, which provides a search facility with wild-card capabilities, making it possible to search for Web sites that have a .uk domain. Results are usually grouped in batches of 2,000, making it difficult to obtain accurate answers. However, in response to an email request, Netcraft provided a summary of statistics for the .uk domain, which are listed in Table D.1.

**Table D.1: Numbers of Web sites in .uk domain, March 2002**

| Domain       | Total            |
|--------------|------------------|
| .co.uk       | 2,750,706        |
| .org.uk      | 170,172          |
| .sch.uk      | 16,852           |
| .ac.uk       | 14,124           |
| .ltd.uk      | 8,527            |
| .gov.uk      | 2,157            |
| .net.uk      | 580              |
| .plc.uk      | 570              |
| .nhs.uk      | 215              |
| ...          |                  |
| <b>Total</b> | <b>2,964,056</b> |

*Source: Based on figures supplied by Netcraft*

**OCLC's Web Characterization Project (WCP)** makes use of an annual Web sample to analyse trends in the size and content of the Web. Analysis based on the sample is publicly available.

In 2001 the figures published on OCLC's WCP Web site indicated that UK Web sites consisted of 3% of the overall total of 8,443,000 Web sites. This gives over 253,000 unique Web sites.

It will be noticed that these figures are less than one tenth of the figures given by Netcraft. This is because OCLC measure only physical Web servers, which are identified by the IP address of the server. The OCLC approach ignores virtual Web servers, which now form a significant proportion of Web servers.

## Findings - Numbers of Pages

Another approach to measuring the extent of UK Web space is to measure the number of pages on UK Web sites. One approach to this is to record the numbers of pages in the .uk domain indexed by search engines such as AltaVista and Google. It should be noted, however, that the coverage of search engines is highly variable and they tend to be poorly documented. (Snyder & Rosenbaum, 1999; Borgman & Furner, 2002, p. 35)

The search interface for AltaVista and Google enables wild-card searches based on domain names. In AltaVista the search term `url:* .uk` can be used to search for pages in the .uk domain. In Google the format is `site:.ac.uk uk`.

A summary of the findings, from March 2002, is given in Table D.2.

**Table D.2: Numbers of pages indexed by major search engines, March 2002**

| Domain  | AltaVista  | Google    |
|---------|------------|-----------|
| .ac.uk  | 5,598,905  | 2,080,000 |
| .co.uk  | 15,040,793 | 3,570,000 |
| .org.uk | 1,644,322  | 898,000   |
| .gov.uk | 975,506    | 343,000   |
| .uk     | 24,862,369 | 4,760,000 |

Source: AltaVista (<http://www.altavista.com/>), Google (<http://www.google.com/>)

Of course, as with measuring the numbers of Web sites, these figures must be treated with caution. In particular note that:

- The search engines only index publicly available Web sites.
- The search engines only index the "visible Web" and will not index dynamic pages, many proprietary file formats, etc.
- The search engines may duplicate findings, by not be able to differentiate between identical pages.
- Results from search engines can fluctuate due to the load on the search engines.

## Conclusions

We have looked at approaches to counting the numbers of Web sites and Web pages within the UK Web space. We have found that this is an ill-defined concept. Measurements based on the .uk domain are easiest to address, as this can be carried out using automated tools. However even automated tools give inconsistent findings.

## References

Borgman, C.L. & Furner, J. (2002). "Scholarly communication and bibliometrics." *Annual Review of Information Science and Technology*, 36, 3-72.

Snyder, H.W. & Rosenbaum, H. (1999). "Can search engines be used as tools for Web-link analysis? A critical review." *Journal of Documentation*, 55 (4), 375-384.

## Appendix E: Questionnaire sent to Web archiving initiatives

This was sent to representatives of selected Web archiving initiatives. Replies were received from: Allan Arvidson (Royal Library, Sweden), Andreas Aschenbrenner and Andreas Rauber (Austrian On-Line Archive), Stephen Bury and Deborah Woodyard (British Library), Leonard DiFranza (Internet Archive), Julien Masanès (Bibliothèque nationale de France), Margaret Phillips (National Library of Australia) and Dave Thompson (National Library of New Zealand).

Date: Wed, 9 Oct 2002 11:47:44 +0100 (BST)  
From: Michael Day <lismd@ukoln.ac.uk>  
To:  
Subject: JISC/Wellcome Trust Web archiving study

UKOLN is currently undertaking a feasibility study on Web archiving for the Joint Information Systems Committee of the UK funding bodies for further and higher education in England, Scotland, Wales and Northern Ireland and the Library of the Wellcome Trust.

As part of this study, we would like to ask you some questions about your Web archiving initiative, especially about collection policies, technical approaches to preservation and costs. The information given will help support a review of existing Web archiving initiatives that will appear in the report and will help inform any recommendations for the JISC and Wellcome Trust.

If you are not the most appropriate person to answer these questions, I would be grateful if could you pass this on to the relevant person(s) in your organisation. All questions are optional, but please answer as many as you can. If possible, I would be grateful if you could send a reply by the 23rd October.

If there is any information that should not be made publicly available, please indicate this and it will only be made available to the research team led by UKOLN and representatives of the two funding organisations.

Michael Day  
Research Officer, UKOLN, University of Bath

### Questionnaire

#### 1. General

1.1 What is the name of your Web archiving initiative?

1.2 Which organisations are responsible for it?

1.3 How is the initiative funded?

#### 2. Organisational issues

In common with other repositories, Web-archiving initiatives deal with issues of selection, organisation and access. The

questions in this section attempt to shed some light on selection criteria, frequency, indexing and end-users.

- 2.1 What criteria do you use to select sites for inclusion in the archive? Are these criteria publicly available?
- 2.2 How frequently do you capture Web sites?
- 2.3 Do you maintain a list of URLs?
- 2.4 Do you attempt to collect all levels of each Web site?
- 2.5 Do you attempt to index your Web archive in any way? If so, please could you give details?
- 2.6 Do you collect and maintain any other metadata, e.g. about rights management, provenance or context?
- 2.7 Is there any overlap with national legal deposit arrangements, e.g. with newspapers or periodicals?
- 2.8 Do you receive Web resources directly from publishers? If so, how?
- 2.9 Do you do the archiving within your own institution or is it undertaken in co-operation with external organisations?
- 2.10 How does your initiative deal with public access to the archive? For example, is access limited to certain locations or individuals?

### 3. Technical issues

To date, most Web-archiving initiatives have either been based on bulk collection through harvesting techniques or on the selection and replication of individual Web sites. This section attempts to find out what technical approaches to Web-archiving have been adopted by your initiative, the software and hardware that has been used, the size of the resulting archive, etc.

- 3.1 Is your Web archiving initiative based on bulk collection, the selection of individual sites, or a combination of both of these?
- 3.2 What software do you use?
- 3.3 Is this software proprietary, open-source or developed in-house?
- 3.4 Do you have any comments on this software?
- 3.5 What hardware is in use for this initiative?
- 3.6 Is this equipment used for any other tasks in your institution?
- 3.7 How does your initiative attempt to deal with database-driven, interactive sites, sites that use software plugins, or 'deep-Web' resources?

3.8 What amount of data is currently held in the Web archive, e.g. in Gigabytes or Terabytes?

3.8 How fast is the Web archive growing?

3.10 What data-types appear to be most popular, e.g. HTML, PDF, etc.?

#### 4. Costs

We would be interested in gaining some idea of cost. We realise that this data may be difficult to estimate or may be sensitive, but it would be useful to have some indication of the relative costs of the different approaches to Web-archiving.

4.1 Please could you give an estimate of the archive's cost, e.g. an indication of staff numbers and time, set-up costs, management time, etc.

#### 5. Some information about the respondent(s)

5.1 Your name(s):

5.2 Your position(s) within the organisation:

5.3 Please could you give details of any key policy documents, reports or publications on your Web archiving initiative that have not been referred to already?

Many thanks for finding the time to reply to this questionnaire. If we have any further questions, we may follow this up with an additional e-mail or telephone call. We will attempt to keep you informed about future progress of this feasibility study and any outcomes.

Best wishes,

Michael Day

-----  
Research Officer

UKOLN, University of Bath, Bath BA2 7AY, United Kingdom  
Telephone: +44 (0)1225 383923 Fax: +44 (0)1225 386838  
-----

## Appendix F: Abbreviations used

|        |   |
|--------|---|
| ACM    | Association for Computing Machinery                               |
| AIDS   | Acquired Immune Deficiency Syndrome                               |
| AOLA   | Austrian On-Line Archive  |
| ASP    | Active Server Pages   |
| BBC    | British Broadcasting Corporation                                  |
| BL     | British Library   |
| BMA    | British Medical Association                                       |
| BMJ    | British Medical Journal   |
| BNF    | British National Formulary  |
| BnF    | Bibliothèque nationale de France                                  |
| CBHD   | Center for Bioethics and Human Dignity                            |
| CCSDS  | Consultative Committee on Space Data Systems                      |
| CD     | Compact Disc  |
| CD-ROM | Compact Disc - Read Only Memory                                   |
| CERN   | European Centre for Nuclear Research                              |
| CGI    | Common Gateway Interface  |
| CLIR   | Council on Library and Information Resources                      |
| COIN   | Circulars on the Internet   |
| CORC   | Cooperative Online Resource Catalog                               |
| CPA    | Commission on Preservation and Access                             |
| CSC    | Tieteen tietotekniikan keskus CSC (Finnish IT Center for Science) |
| CSS    | Cascading Style Sheets  |
| DAML   | DARPA Agent Markup Language                                       |
| DARPA  | Defense Advanced Research Projects Agency                         |
| DIPEx  | Database of Individual Patient Experiences                        |
| DLT    | Digital Linear Tape   |
| DPC    | Digital Preservation Coalition                                    |
| DVD    | Digital Versatile Disc  |
| eLib   | Electronic Libraries Programme (JISC)                             |
| EMBL   | European Molecular Biology Laboratory                             |
| EOSDIS | Earth Observing System Data and Information System                |
| ERA    | Electronic Research Archive                                       |
| ERM    | electronic records management                                     |
| ESMRSD | Electronic and Special Media Records Services Division (NARA)     |
| FE     | further education   |
| FP6    | Sixth Framework Programme (EU)                                    |
| FTE    | full-time equivalent  |

---

|       |   |
|-------|---|
| FTP   | File Transport Protocol   |
| GIF   | Graphics Information File   |
| HE    | higher education  |
| HEBS  | Health Education Board for Scotland   |
| HIV   | Human Immunodeficiency Virus  |
| HMSO  | Her Majesty's Stationery Office   |
| HSM   | hierarchical storage management   |
| HTML  | Hypertext Markup Language   |
| HTTP  | Hypertext Transport Protocol  |
| IBM   | International Business Machines Corporation   |
| ICT   | information and communication technology  |
| IDE   | Integrated Drive Electronics  |
| IEEE  | Institute of Electrical and Electronics Engineers, Inc.   |
| IETF  | Internet Engineering Task Force   |
| INRIA | Institut national de recherche en informatique et en automatique (French National Institute for Research in Computer Science and Automatic Control) |
| ISO   | International Organization for Standardization  |
| IST   | Information Society Technologies  |
| IVF   | In Vitro Fertilisation  |
| JCALT | JISC Committee on Awareness, Liaison and Training   |
| JISC  | Joint Information Systems Committee   |
| JPEG  | Joint Photographic Experts Group  |
| JTAP  | JISC Technology Applications Programme  |
| KB    | Kungliga biblioteket, Sveriges nationalbibliotek (The Royal Library, National Library of Sweden)  |
| LANL  | Los Alamos National Laboratory  |
| LoC   | Library of Congress   |
| MARC  | Machine Readable Cataloguing  |
| MeSH  | Medical Subject Headings  |
| MRC   | Medical Research Council  |
| MVS   | Manchester Visualization Centre   |
| NARA  | National Archives and Records Administration  |
| NASA  | National Aeronautics and Space Administration   |
| NBII  | National Biological Information Infrastructure  |
| NCDC  | National Climatic Data Center   |
| NCSA  | National Center for Supercomputing Applications   |
| NEC   | NEC Corporation (formerly Nippon Electric Company)  |
| NeLH  | National electronic Library for Health  |
| NHS   | National Health Service   |

---

|            |   |
|------------|---|
| NIH        | National Institutes of Health   |
| NLA        | National Library of Australia   |
| NLM        | National Library of Medicine  |
| NOAA       | National Oceanic and Atmospheric Administration                       |
| NSDL       | National Science Digital Library                                      |
| NWA        | Nordic Web Archive  |
| OAI-PMH    | Open Archives Initiative Protocol for Metadata Harvesting             |
| OAIS       | Open Archival Information System                                      |
| OCLC       | OCLC Online Computer Library Center, Inc.                             |
| OIL        | Ontology Inference Layer  |
| ONB        | Österreichische Nationalbibliothek (Austrian National Library)        |
| PANDAS     | PANDORA Digital Archiving System                                      |
| PANDORA    | Preserving and Accessing Networked Documentary Resources of Australia |
| PDF        | Portable Document Format  |
| PRO        | Public Record Office  |
| RDF        | Resource Description Framework  |
| RDMS       | relational database management system                                 |
| RDN        | Resource Discovery Network  |
| RLG        | Research Libraries Group, Inc.  |
| RTF        | Rich Text Format  |
| SGML       | Standard Generalized Markup Language                                  |
| SHOE       | Simple HTML Ontology Extensions                                       |
| SLAC       | Stanford Linear Accelerator Center                                    |
| SPIRES-HEP | Standard Public Information Retrieval System - High-Energy Physics    |
| TSO        | The Stationery Office   |
| TU Wien    | Technische Universität Wien (Vienna University of Technology)         |
| UKERNA     | United Kingdom Education & Research Networking Association            |
| ULCC       | University of London Computer Centre                                  |
| URI        | Universal Resource Identifier   |
| URL        | Uniform Resource Locator  |
| USGS       | United States Geological Survey                                       |
| VRML       | Virtual Reality Modeling Language                                     |
| W3C        | World Wide Web Consortium   |
| WAIS       | Wide Area Information Servers   |
| WHO        | World Health Organization   |
| XML        | Extensible Markup Language  |