South African Journal
of Information Management
[DoE accredited]

Web sites Vol.6(4) December 2004

Preserving our Internet history – Websites on archiving via the Internet

Melanie Sutton
Independent Information Management Specialist
msutton@i-innovate.co.za

Introduction

The World-Wide Web has developed into an immense international complex of hyperlinked information. While some of the information available today simply mirrors that found in existing print publications, an enormous amount of information cannot be found anywhere but on the Web. Significant historical information and other information, such as that found on medical Web sites may be of long term scientific value. The uniqueness of the information, combined with the ephemeral nature of digital information, has resulted in a growing perception that there is a need for mechanisms to preserve at least some of that immense volume of information for the longer term (Charlesworth 2003).

Our final journey in the footsteps of an information professional delves into the tangibility of archiving Internet content.

Archives and the Internet

If we understand our cultural heritage to mean a generation's established customs in terms of outlook, lifestyle, social conventions, aesthetic and other artistic norms and forms of expression, that is wholly or partially adopted by succeeding generations, then the Internet is undoubtedly one of the most original contributors to that heritage today (Christensen-Dalsgard, Fonss-Jurgensen, Von Hielmcrone, Ole, Brugger, Henriksen and Vejrup 2003).

Information services and archives have traditionally preserved and provided access to society's cultural artifacts. In this era of digital technology, a new breed of information services has sprung up, namely digital archives. These archives, like all services, foster education, scholarship and creativity through access to information. But digital archives excel in providing access to the ephemeral, the 'grey' material, the lost classics, material that is difficult for traditional libraries to locate, acquire and store. To serve their mission of preservation and access, digital archives depend on volunteers, donations and on the public domain.

But even as the Internet, as a medium of communication and as a repository of knowledge is growing every day, a large proportion of material published on the Net is disappearing with disturbing rapidity. Several studies have shown that 40% of the material on the Net disappears within one year, while a further 40% is altered, leaving only 20% in its original form. Other studies indicate that the average lifetime of a Web page is 44 days. As a result, we may soon find ourselves in a position that when it comes to writing the history of our times, a significant part of the source material will be missing (Christensen-Dalsgard *et al.*

2003)

The overall purpose in establishing an Internet archive is therefore to ensure the preservation of this contribution to the cultural heritage and thereby the source materials that will provide the foundation for future research not only into the Internet's own history but also into all the ever more comprehensive cultural, institutional and business activities that take place, especially and in some cases exclusively, on the Internet (Christensen-Dalsgard *et al.* 2003).

Digital technology allows us the opportunity to build a universal library out of these materials. This library will expand our understanding of public access. It will make information accessible. At the same time, digitization will eliminate deterioration caused by the physical handling of cultural artifacts.

Barriers to Internet archiving

The task of preserving Web-based information is not, however, an easy one. Aside from the technical difficulties inherent in preserving transient digital resources, the legal environment in many countries is also often inhospitable to, or unappreciative of the role of the would-be Web archivist. These projects face one significant limitation – copyright. Because of copyright concerns, groups are effectively limited to make available only works no longer under copyright. Hazards also lurk in the form of defamation law, content liability and data protection laws (Charlesworth 2003). Creating an archive of informal and personal information has many difficult legal and social issues, even if the material was intended to be publicly accessible at some point (Kahle 1996).

The Internet Archive

The Internet Archive was founded in 1996 in order to build a digital library with the purpose of offering permanent and free access to researchers, historians, scholars and the general public. The Internet Archive is working to prevent the Internet, a medium with major historical significance, and other digital material from disappearing into the past. Collaborating with the institutions such as the Library of Congress and the Smithsonian, the Internet Archive is working to permanently preserve a record of public material. The Archive holds a collection of archived Web pages, dating from 1996 and, since 1999, it has expanded its collection to include a September 11 television and online catalogue, an Election 2000 online library and archived movies from 1903 to 1973. As well as Web pages, it also archives moving images, texts and audio.

- The Internet Text Archive www.archive.org/texts/texts.php is a free collection of archival texts that consists of the Million Book Project, a searchable digital library, a Children's Library and includes works from the International Children's Digital Library; Project Gutenberg – one of the oldest and largest publishers of public domain literary works on the Internet, Arapanet with its collection of memoranda, interview notes, periodicals, papers and other reference materials that document the development of the Advanced Research Projects Agency Network of the US Department of Defense as well as other Open Source Books contributed by the community.
- The Internet Audio Archive www.archive.org/audio/ includes a Live Music Archive which provides downloadable live concerts for current and future generations to enjoy.
- The Movie Internet Archive www.archive.org/movies/movies.php is involved in digitizing movies for online access, providing a rich and fascinating core collection of archival films.

In addition to developing their own collections, they are working to promote the formation of other Internet libraries in the United States and elsewhere.

The Wayback Machine (www.archive.org) provides users of the Internet Archive (the largest known database in the world) the facility to search over 10 billion pages and see what a particular page looked like at various periods in Internet time. A search yields a list of pages that are available as far back as 1996. The Wayback Machine is a repository of more data than that contained in the world's largest libraries including the Library of Congress.

Web archiving strategies and methodologies

Several different strategies and methods of acquiring materials from the Internet are practised. Harvesting strategies include the following:

- Snapshot archiving aims to save a reasonable quantity of publicly available material at well defined intervals. The Internet Archive (www.archive.org) uses this approach.
- Selective archiving evaluates the value of the posterity of particular objects, works or total Web sites, and those that are considered valuable are preserved. An example of this approach is the Pandora initiative in Australia (http://pandora.nla.gov.au).
- Event-oriented archiving is used when a given event is judged significant and materials relating to it are preserved so that both the event and the reactions to it can be followed for posterity. An example of this approach is the September 11 archive on the Internet Archive.
- Statutory archiving refers to categories of material that must be delivered to certain libraries defined by legislation. The difference between this approach and the selective approach is that in the case of legal deposit, types of material are defined as worthy of harvesting, rather than individual documents. Government publications are an example of one type of material that is currently harvested in this manner.

The method of harvesting refers to the method whereby the material is acquired. The various methods are as follows:

- Pull, that is, the material is gathered manually or automatically either by using a harvester or by some other method. Manual harvesting is when individual files are gathered and put into an archiving system. Automatic harvesting refers to materials that are harvested automatically on the basis of criteria such as domains or a series of manually or automatically generated URLs that are automatically included in an archive. Examples of this include the archiving project Kulturarw3 in Sweden (www.kb.se/kw3/) and the Internet Archive in the USA (www.archive.org).
- Push refers to material that is delivered or donated to the archive. Delivery is through a publisher, where the manner of delivery of material is agreed in advance. The materials delivered to the State Archives are an example of this. The National Archives of South Africa can be located at (www.national.archives.gov.za). Donations refer to a collection of material that is donated and, in many cases, may have to be structured for Internet archiving.

Two important parameters in connection with the strategy for the acquiring of Web materials is the frequency of updating and the complexity of the interactivity of the Web site. Depending on these factors, different strategies of harvesting should be chosen (Christensen-Dalsgard *et al.* 2003).

The Danish archiving project

Netarchive.dk (www.netarchive.dk) is the project that plans for the preservation of Denmark's cultural heritage on the Internet for future generations. In 2002, a pilot project was launched in Denmark in collaboration with The Royal Library, The State and University Library and Centre for Internet Research with the objective to develop a fundamental

strategy for archiving Danish Internet material.

In continuation of the pilot project, the Royal Library and The State University Library started working on the second phase of the project with the objective to refine the strategy from the pilot project and to investigate the possible need for any amendments to the legal deposit law.

The Royal Library's special focus is snapshot harvesting, whereas The State and University's Library will concentrate on selective and event-based harvesting and delivered material. The two libraries work closely together on the establishment of the whole archive and develop in collaboration strategies and software for the collection, archiving, preservation and access of material.

Conclusion

Digital archives allow us to preserve our cultural heritage, preserve copyrighted works and prevent their permanent loss, promotes full public access to our cultural heritage, support rich and diverse use of our cultural heritage, extend our cultural heritage and make preservation and access more economical.

References

Christensen-Dalsgaard, B., Fonss-Jurgensen, E., von Hielmcrone, H., Ole Finneman, N., Brugger, N., Henriksen, B. and Vejrup Carlsen, S. 2004. Experiences and conclusions from a pilot study: Web archiving of the district and county elections 2001. [Online]. Available: http:// www.netarkivet.dk/rap/webark-final-rapport-2003.pdf.

Charlesworth, A. 2003. Legal issues relating to the archiving of Internet resources in the UK, EU, USA and Australia. [Online]. Available http:// www.jisc.ac.uk/uploaded_documents/archiving_legal.pdf.

Kahle, B. 1996. Archiving the Internet. [Online]. Available: http://uibk.ac.at/sci-org/voeb/texte/kahle.htm

About the author

Melanie Sutton (BA, Postgraduate Diploma in Information Management) is an information management specialist assisting South African companies implementing practical information management solutions. She is currently contracted to FNB Homeloans as an information specialist in a training design and development team.

ISSN 1560-683X

Published by InterWord Communications for the Centre for Research in Web-based Applications,
Rand Afrikaans University