Archiving Web Site Resources: A Records Management View

Maureen Pennock
Digital Curation Centre (DCC)
UKOLN, University of Bath
BATH, UK
+44 1225 386711
M.Pennock@ukoln.ac.uk

Brian Kelly
UKOLN
University of Bath,
BATH, UK
+44 1225 383943
B.Kelly@ukoln.ac.uk

ABSTRACT
In this paper, we propose the use of records management principles to identify and manage Web site resources with enduring value as records. Current Web archiving activities, collaborative or organisational, whilst extremely valuable in their own right, often do not and cannot incorporate requirements for proper records management. Material collected under such initiatives therefore may not be reliable or authentic from a legal or archival perspective, with insufficient metadata collected about the object during its active life, and valuable materials destroyed whilst ephemeral items are maintained. Education, training, and collaboration between stakeholders are integral to avoiding these risks and successfully preserving valuable Web-based materials.

Categories and Subject Descriptors
K.6.0 Management of Computing and Information Systems (General)

General Terms
Management, Human Factors.

Keywords
Records Management, Archiving Web Sites, Best Practices.

1. INTRODUCTION
Web sites are a particularly transient type of digital resource: the average lifespan of a Web page in 2003 was deemed to be 100 days and it is not unreasonable to suggest that it is even shorter today [1]. Several 'solutions' to preserve Web sites have been developed, many of which are library- or community-based. Whilst extremely valuable in their own right, they cannot and do not affect Web sites at the point of production and in their 'active' life. Proper management of Web-based records during the active phase of their life-cycle is vital if authenticity and integrity is to be assured at a later date. In effect, efforts must be expended to ensure Web sites are 'future-proof' [2]. This paper addresses an alternative and selective approach to preserving Web-based materials based on records management principles, whereby the preservation and record-keeping requirements of institutions, organisations, and archives are fundamental and impact day-to-day management of the live data.

WWW 2006, May 23–26, 2006, Edinburgh, Scotland.
ACM 1-59593-323-9/06/0005.

2. IMPORTANCE OF WEB SITE ARCHIVING & PRESERVATION
Today's Internet reflects the developments and achievements of modern society. The Internet has fostered a change in the social structure of society, enabling a truly global network to develop, and has broadened access to materials in a fashion not experienced since the creation of the printing press. Its cultural and historical value therefore necessitates its preservation. Furthermore, resources published on the Internet constitute a direct communication between originating organisations and their audience. For organisational Web sites, such as government, institutional, and private businesses, such resources may have transactional, evidential, and record-keeping value. The proper archiving and preservation of these resources, as well as the contextual information surrounding them, is particularly important when the Internet is the only medium through which the message is communicated. Web sites and reliable information about Web-based resources thus hold value over time for multiple reasons and must be managed in a manner appropriate to their retention.

3. CURRENT APPROACHES
Most current models for Web site preservation originate from libraries. They often utilise a harvesting approach, whereby robots crawl the Web copying content and metadata about resources they encounter across a particular domain. The Internet Archive follows this harvesting model and attempts to collect and archive the entire Internet. Collections may also be selective or thematic, according to the requirements or remit of the collecting organisation. The Swedish Kulturawr3 project to preserve Web sites of particular relevance to Sweden works in this way. Alternatively, robots may be released on a more subjective or restrictive basis, with instructions only to copy certain sets of Web sites that have already been identified by the collecting organisation as Web sites of interest. This is particularly the case where copyright or permission to archive must first be sought, as is the case with the PANDAS software of the National Library of Australia (NLA) and the UK Web Archiving Consortium (UKWAC) [3]. Unlike other similar initiatives, UKWAC unusually benefits from the participation of non-library organisations, broadening expertise and sharing practical experience across different sectors.

Some countries rely not upon automated harvest but upon deposit of materials, and electronic legal deposit legislation is slowly taking shape in Europe that reflects the ongoing value of online digital resources, especially those published in electronic form only. Such deposit is already compulsory in Sweden, whilst the

National Library of the Netherlands has negotiated voluntary deposit agreements with publishers.

4. RECORDS MANAGEMENT PERSPECTIVE

Records are evidence of business and administrative transactions and are essential resources for accountability and organizational memory. The content of these records can determine the acts of the users who access them. The rise of electronic government, e-commerce, and general use of Web sites to communicate official ordinances has resulted in an increasing number of records existing solely online, and being published only in that form. Traditional paper-based record-keeping infrastructures commonly fail to identify, control, and capture records published in electronic format, particularly when responsibility for creation, control, dissemination and maintenance of records is widely distributed. This is a problem because records must meet certain requirements in order to qualify as evidence.

Application of records management principles to Web site maintenance ensures that records published on the site are controlled and can meet the requirements for organisational accountability. The starting point is that not all data needs to be retained: the basic premise of records management is that it is neither necessary nor desirable to retain everything. The first step in implementing proper records management for Web-based records is therefore to identify that which should be preserved and for how long, against that which should be destroyed. Ignoring this principle results in the capture of ephemeral, redundant, or incorrect material alongside (or instead of) permanently valuable and transactional records, which may adversely impact the entire records management strategy.

Retention periods specify the length of time information must be retained. Different categories of information have differing retention periods. Legal ordinances may impact upon retention periods, particularly the Data Protection Act and the Freedom of Information Act. Records management guidelines and retention periods must be applied to Web-based data to ensure that such legal obligations are met, for records posted on both the Internet and Intranets.

Information posted to Web sites is frequently updated on a casual basis, with no permanent records kept of changes that were made - such as error correction, date of creation, date of posting, date of deletion, and author. Whilst frequent updates and corrections are useful features of Web-based materials, they can cause confusion and lead to errors or misunderstanding by users if such changes are not recorded. Measures should therefore be taken to ensure that identified records are a) retained, b) unchanged, and c) authentic and their integrity assured. This can only be reliably achieved if control is exercised over the entire life-cycle of the records, i.e. from creation through to publication, archiving, preservation and possible eventual destruction. Failure to do so impacts upon the perceived value of the records, their reliability, and their legal status. A clear and institutionally embedded policy on Web site management and the relationship between Web-based materials and records management helps ensure conformance and thus high-quality records.

4.1 Use Case Scenarios

A simple example illustrating problems caused by frequent and non-recorded data updates involves an ordinance specifying a submission date which incurs financial penalties if missed, for example, from a Taxation Agency: if the date is incorrect and brought forward, without maintaining a record of the error and several users subsequently incur financial penalties, an enquiry could be launched that potentially damages public perception of the organization and casts doubt on its accountability.

Data Protection scenarios involve inappropriate information posted and available online, such as information that clearly identifies personal circumstances of students. Freedom of Information and Data Protection issues can also arise from, for example, public enquiries and comments submitted via Web-forms that are stored in the Web-management system.

Legal discovery notices apply to records also hosted on Intranets; if unnecessary and potentially damaging documents are stored on the Intranet that are not part of a records retention schedule and should have been destroyed, the organization is liable to a whole range of problems, depending on the context of the legal notice.

5. FURTHER WORK AND CONCLUSIONS

This short paper introduces the value and need for records management principles in Web site management. In order to ensure that the principles can be implemented, education, training, and collaboration between records creators, records managers, and records-hosting bodies is essential. Good practices for Web site accessibility additionally contribute to the persistence of records hosted online: persistent identifiers, use of standards, embedded metadata that offers alternative mechanisms for accessing or understanding content, and log file maintenance. Systems that keep Web pages up-to-date and adhere to records management and record-keeping principles, whilst retaining old and valuable records and metadata offline, limit liability, ensure accountability, and contribute towards future preservation activities. Although this is a challenge to achieve with the increasing complexity of dynamic Web based resources and remains the subject of further research, it is an absolute necessity for a successful and accountable online presence. Within the UK, the Digital Curation Centre (DCC) has been established and is developing a tool for 'archiving databases', i.e. preserving past states in an efficient form, that may prove useful for this purpose [4].

6.ACKNOWLEDGMENTS

The authors acknowledge the contribution of Chris Rusbridge, Director of the Digital Curation Centre.

7.REFERENCES

[1] Weiss, R quoting Brewster Kahle, in *On the Web, Research Work Proves Ephemeral* Washington Post Nov 24 2003.

[2] DCC/Wellcome Library workshop, *Future-Proofing Web sites* http://www.dcc.ac.uk/events/fpw-2006/

[3] PANDAS/PANDORA archive, http://www.kb.se/kw3/ENG/Default.aspx

[4] Archiving Scientific Data, Buneman et al in ACM Transactions on Database Systems, Vol 27, 2004, 2-42.