What Should We Preserve? The Question for Heritage Libraries in a Digital World

Margaret E. Phillips

Abstract

A primary role of national libraries is to document the published output of their respective countries. Traditionally, this has meant collecting, describing, and preserving for future generations at

Subsequently the definition of “library materials” was extended to include information stored on other physical carriers such as microfilm, film of various types, audio cassette tapes, video tapes, computer disks, CD-ROMS, be at a crossroads with regard to planning their future directions for digital archiving (Gatenby, 2002). Whether they were engaged in whole domain (comprehensive) harvesting or selective archiving, each was recognizing the at the University of Aarhus to test the viability of the thematic approach (event-based archiving) through the Netarchive.dk project (Royal Library, Denmark, 2003). The Royal Library and the State and University Library together have gone on to incorporate event-based archiving into a three-pronged approach to Web archiving, including automatic snapshot harvesting and selective harvesting (Royal Library, Denmark, 2004).

Archiving Based on Collaborative Agreements with Selected Commercial Publishers

The National Library of the Netherlands has taken a different approach altogether, responding to a particular situation where 30 percent of all scientific publications in the world occur in that country. It has focused on commercial publications and, in association with IBM, has developed technical infrastructure and organizational relationships with a small number of commercial publishers, including Elsevier Science and Kluwer Academic, to archive, preserve, and provide limited access to the whole digital output of the publishers concerned (National Library of the Netherlands, 2004). It takes in large volumes of online publications from a small number of publishers. Collaborative agreements with publishers also work well under the selective model, and the National Library of Australia and the Commonwealth Scientific and Industrial Research Organisation (CSIRO) have recently reached an agreement whereby the library will archive all of CSIRO's online commercial publications.

This enhances our knowledge of preservation requirements and enables risk assessments and preservation strategies to be put in place.

- Sites that are inaccessible to harvesting robots can be identified and archived using other methods, by arrangement with the publisher.

Disadvantages

In selecting titles for the archive, libraries are making subjective judgements about the value of resources and what researchers of the future are likely to fi low staff cost per item gathered. The whole domain is available to future researchers, and resources can be seen in their broader context, with links to other documents retained.

Disadvantages

The only example of a whole domain archive that is readily available for evaluation is the Internet Archive, which attempts to capture the whole Web every two months. Valuable though this resource is, having commenced its work in 1996 and now having amassed a considerable volume of historical data, it does have limitations of concern to agencies looking for completeness and version control of documentary heritage.

Hybrid Approaches

All of the approaches discussed so far have disadvantages—the selective approach misses material that may be of future value, the whole domain

The first task, when we started this work in 1996, was to decide what we would collect, and this resulted in the publication of our first selection guidelines. As we implemented these guidelines and learned more about

> we do our best to solve the problems presented by a publication or class of publications. For instance, our desire to archive Deep Web sites, including databases, has led us to embark on a research project in conjunction with the International Internet Preservation Consortium (IIPC) to find or create tools and methods for collecting and preserving information presented in these dynamic formats.

In 2003, after seven years of selecting and archiving online publications and Web sites, the library conducted a major review of its selection guidelines to ascertain whether they remained relevant and flexible enoug formatshivin7 on

- Publications of tertiary education institutions
- Conference proceedings
- E-journals
- Items referred by indexing and abstracting agencies (which are frequently from the first four categories but also include items with print versions)
- Sites in nominated subject areas (specified in an appendix to the selection guidelines) on a rolling three-year basis, and sites documenting

and its partners to collect online publications at an adequate level. The

Those national libraries that were active in the field have developed a lot of knowledge, expertise, systems, and software to manage the activity. Most of this knowledge and these systems have been developed in isolation, with impact on what is available to researchers of the future. This working group, which consists not only of members but also of invited researchers in the area of Internet studies, is aiming to define a common vision of what needs to be collected (International Internet Preservation Consortium, 2004c).