International Approaches to Web-Archiving Panel Discussion

Thorsteinn Hallgrimsson

National and University Library of Iceland, Arngrímsgata 3, 107 Reykjavik

thh@bok.hi.is

Preserving the World Wide Web

The Web is a separate medium just like books, newspaper, periodicals, CD‘s, movies et cetera. It is totally digital but the contents are both born digital material and digitized versions of other media. It contains an enormous amount of data, measured in billions of documents. It is also very volatile and it was soon discovered that a lot of its contents is short-lived and disappears. Preserving this medium for the future therefore presents many new problems and challenges to those who attempt it. The first efforts to preserve the Web by archiving web pages were made in 1996 by Australia, Internet Archive and Sweden and by 2000 several efforts were being made. In 2003 most of the institutions that were seriously thinking about preserving the Web, i.e. 11 national libraries and IA, established the IIPC (International Internet Preservation Consortium) and now it has 38 members including 28 national libraries.

To those involved in preserving the Web it is obvious that currently, and increasingly in the future, a large and significant part of our culture will only exist on the Web and therefore this medium must be preserved for the same reasons that most countries strive to preserve and provide access to their cultural and intellectual heritage by collecting it and storing in museums, archives and libraries. If this is not addressed now an important part of our culture, together with most documentation of the cultural change involved, will be lost. Considering that IFLA has more than a hundred national libraries it is valid to ask why only minority has actively started to preserve their national Web space? There is not a simple answer to this but the following one or more of the following reasons certainly play a role:

- Archiving and preserving the web is on the borderline between the library and the information technology (IT) professions, and the methods used reflect that. It is a library collection, but it requires substantial involvement of IT resources for implementing technical solutions and because of the huge volume of documents involved. Library systems have to cope with up to a 100 million records, a web archive must cope with several billion records. This difference in scale must be considered for all aspects of web archiving and preservation efforts.

- Legal issues and policies are important while in many countries the legal framework for archiving web pages does not exist or is considered to be an obstacle. In recent survey of IIPC members only five countries responded that they had a Law enacted or passed allowing them to collect web pages and archive them. Another four expect a law to be enacted and five are lobbying for a law. Obviously this complicates the issue while national libraries have traditionally relied on the legal deposit law of each country to economically and comprehensively collect and preserved manuscripts and published printed material, and as publishing technology has progressed the libraries extended their collection activity by including physical electronic media like CD-ROM’s and some electronic publications like electronic journals. If the traditional axiom of the legal deposit laws and other collection activity holds true it is therefore an absolute necessity to extend this concept to the web. Still this situation has not prevented many countries from actively working on preserving at least parts of their national web domain by collecting web pages

- National libraries do not agree on what preserving the national web domain really means and what kind of collection building rules should be applied. The web is very different from other media while everybody can input documents without any editing or screening in almost any format one can think of. The contents reflect almost every aspect of the daily life, concerns and issues in those parts of the world that have easy access to the Web and range from the trivial to very important data about society. It must be noted here that in many countries common people do neither have easy access to PC‘s nor to the Web.
  Some national libraries have decided to use traditional librarian values, where „quality“ web sites are selected, harvested and catalogued by librarians, and access is by structured search.
  Others have decided to endeavour to use bulk harvesting to take periodic snapshots of their countries’ entire web domain trying to preserve everything with the aid of computer technology. There are several reasons for this. One of which is

> the difficulty in establishing what will be of value to future researchers and what will not. To make selections from millions of web sites requires enormous personnel efforts at high costs, whereas costs for data memory storage are decreasing at a rapid rate.
>
> A good solution may be to combine periodic snapshots of the entire national domain with selective collections using thematic/event based criteria, and in some cases select web sites that change very frequently like the news media.

Building and sustaining a web archive incorporates the same main activities as in building a traditional library or archive collection, i.e. selection, collection, registration, access, and preservation. From the outset it is important to realize that because of how volatile the web is, it is practically impossible to collect every object present in the web sites or web domains selected. The data that is published on the Web will not be sent to the libraries for preservation but must be actively and systematically collected (harvested) by the libraries. Therefore preservation of the Web starts with the harvesting activity.

The presentation by the National Library of Australia, Bibliothèque nationale de France, British library, The National and University Library of Iceland and the Danish Netarkivet should give the audience a feeling for what is happening worldwide in trying to preserve the Web.