

# ECDL 2002 Workshop Report: Web Archiving

*Contributed by:*

*Andreas Rauber*

*Vienna University of Technology*

*Vienna, Austria*

<[rauber@ifs.tuwien.ac.at](mailto:rauber@ifs.tuwien.ac.at)>

*Julien Masanès*

*Curator*

*Bibliothèque Nationale de France*

*Paris, France*

<[julien.masanes@bnf.fr](mailto:julien.masanes@bnf.fr)>

Following a session on Web Archiving on the first day of the ECDL conference, the Workshop on Web Archiving, organized jointly by the French National Library and the Vienna University of Technology, attracted almost 50 participants from 20 countries. It featured a program of 12 presentations, organized into 3 sessions, covering the wide range of issues in this field.

The first session focused on technical issues, starting with a presentation by Raymie Stata on the data acquisition, storage and interaction techniques, used and provided by the Internet Archive. This was followed by the presentation of a specialized, adaptive crawler as well as an associated XML repository allowing flexible querying of XML data by Patrick Ferran from Xyleme. The third presentation in this session, by Morgan Cundiff from the Library of Congress, described the METS XML schema for cataloging and archiving web pages, was followed by a brief demonstration of the METS-Viewer Software. Gregory Cobena from INRIA presented a high-performance crawler that uses a variety of techniques, such as page-rank and update frequency, to identify important pages on the web. This was followed by a presentation by Julien Masanès from the French National Library who presented a general framework and results from a pilot study for web-archiving with a two-tracks approach, one for the surface web, based on automatic tools, the other for the 'hidden' or 'deep web' based on deposit. Donna Bergmark from Cornell University presented a paper on automated collection building based on focused crawls using tunneling.

The second session was devoted to presentations of various web-archiving projects and showed the diversity of requirements and approaches in this field.

Deborah Woodyard from the British Library then provided an update on the "Britain on the Web" project, formerly known as "Domain uk", a pilot study on archiving a selection of 100 sites from the United Kingdom. Woodyard reported results in terms of efforts required and challenges encountered. In the next talk Hans Liegmann from the

German National Library gave an overview of three prototype applications for archiving electronic documents based on submission, i.e., the "On-Line Thesis" project covering 80 universities in Germany, the "Springer Link" archive housing a copy of about 500 journals by Springer, and a generic submission and delivery interface to allow document submission by publishers. Birgit Henriksen from the Royal Danish Library reported on results from their "Netarchive.dk" project, analyzing different archival approaches and the usefulness of the archived material for research. Henriksen pointed out the importance of re-thinking the concept of a web-site and the limits of site- or domain-restricted collections. Neal Beagrie from the Joint Information Systems Committee concluded the session with a presentation of JISC initiatives with respect to building and archiving community collections. He stressed the need for distributed archives, as well as the necessity of links between subject gateways and internet archives.

The last session was devoted to two international consortia initiatives in the field of Web Archiving. It started with a presentation by Andreas Rauber from the Vienna University of Technology on the European Web Archive initiative, which is currently being proposed by a consortium of more than 27 partners from national and university libraries, research centers, and companies as an Expression of Interest (EoI) for an IST 6th Framework Integrated Project. The focus of this initiative lies with the creation of a European Web Archive, covering strategies and tools for data selection and acquisition, archive maintenance and preservation, as well as access provision and exploitation. The second consortium proposed by Internet Archive and several national libraries was presented by Michele Kimpton from the Internet Archive. It aims at defining collection policy and common practice for web archiving in an international collaboration with national libraries. The consortium would specify and develop a new web crawler targeted towards the archival needs of libraries, crawl each participant's national web space, and provide national data collections crawled by Internet Archive since 1996.

Further information and links to the presentations are available via the Workshop Homepage at <[http://listes.cru.fr/wws/d\\_read/web-archive/pgr\\_ECDL2002.html](http://listes.cru.fr/wws/d_read/web-archive/pgr_ECDL2002.html)>.