Report on the 2nd ECDL Workshop on Web Archiving

Andreas Rauber
Department of Software Technology and Interactive Systems
Vienna University of Technology
http://www.ifs.tuwien.ac.at/~andi

In recent years, we have seen not only an incredible growth in the amount of information available on the Web, but also a shift of the Web from a platform for distributing information among IT-related persons to a general platform for communication and data exchange at all levels of society. The Web is being used as a source of information and entertainment, forms the basis for e-government and e-commerce, and serves as a general platform for meeting and communicating with others via various discussion forums. This situation gave rise to the recognition of the Web's worthiness of being archived and analyzed. We have thus recently witnessed the initiation of numerous projects aiming at the creation of archives of the World Wide Web.

The resulting Web archives are by no means only a collection of historic Web pages. Snapshots of the Web preserve an impression of what hyperspace looked like at a given point in time, what kind of information, issues, and problems people from all kinds of cultural and sociological backgrounds were interested in, the means they used to communicate their interests over the Web, characteristics styles of how Web sites were designed to attract visitors, connectivity, and many other facets of this medium and society in general. They hold a wealth of information that waits to be exploited, information that may be substantial to a variety of disciplines. With the time-line and metadata available in such a Web archive, additional analyses that go beyond mere information exploration become possible. Thus, these archives may well end up forming one of the most fascinating collections of popular digital cultural heritage.

This situation inspired the first Workshop on Web Archiving to be held in association with the European Conference on Digital Libraries (ECDL) in 2001 in Darmstadt, Germany. It served as a first reunion of the numerous players active in this field, showed the diversity of approaches taken, and highlighted the need for further research and a stronger collaboration between the various initiatives.

Thus, 2002 saw the second Workshop on Web Archiving taking place on September 19 in Rome. It was again held in association with the European Conference on Digital Libraries, which itself had a significant focus on Web Archiving related technologies. The main ECDL conference on Sept. 16-18 was opened with a keynote talk by *Hector Garcia Molina* of *Stanford University*, who addressed several issues related to efficient Web crawling, and the WebBase project in general. The subsequent first session of the conference was dedicated to Web Archiving issues, featuring a talk by *Julien Masanes* from the *French National Library (BnF)*. He reported on a first experiences on archiving the French Web, analyzing issues pertaining to the definition of the scope of a national Web archive as well as versioning problems for multiple copies. The second talk in this session, presented by *Andreas Rauber* from the *Vienna University of Technology*, reported on experiments using Data Warehouse technologies for analyzing the Austrian Web archive AOLA. Further sessions of the main conference on collection building, preservation, etc. also contained presentations pertinent to Web archiving.

The workshop, which followed the main conference, was organized jointly by the *French National Library (BnF)* and the *Vienna University of Technology*. It attracted around 50 participants from 20 countries, showing the significant increase of interest in this subject. The program consisted of 12 presentations, which were organized into 4 sessions. The two morning sessions focused on technical issues, while the afternoon sessions provided an update on various national initiatives, and introduced upcoming consortia initiatives, respectively.

Following a welcome and introduction by *Catherine Lupovici* of the *BnF*, the first session started with a presentation by *Raymie Stata* of the *Internet Archive* on their data acquisition, storage and interaction techniques. He described the Internet Archive's current holdings, which amount to more than 120 Terabytes of Web data, and are accessible via a newly released public interface, i.e. the "Wayback Machine". The Web data collection of the Internet Archive is donated by the Web search engine Alexa. It tries to cover as many sites as possible but does not attempt to achieve complete coverage of each site collected.

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* of *Xyleme*, a spin-off from INRIA. A smart crawler is used to collect XML data from the Web, which is then put into a native XML repository, where it can be accessed via the Xyleme Query Language.

The third presentation in this session by *Morgan Cundiff* of the *Library of Congress (LoC)* described an XML schema for cataloging and archiving Web pages based on the Metadata Encoding and Transmission Standard (METS). METS is an open, modular standard for encoding descriptive, administrative, and structural metadata for digital objects. He also demonstrated the encoding of and access to documents using the METS-Viewer software.

The second session focused on collection building, with two presentations on French activities on crawling and deposit and one on a focused crawling setting within the NSDL project. *Gregory Cobena* of *INRIA* presented a high-performance crawler using a variety of techniques such as page-rank and update frequency to identify important pages on the Web. A new algorithm allows the calculation of page importance online without requiring the storage of the complete graph structure, providing highly scalable crawling performance.

This was followed by a presentation by *Julien Masanes* of *BnF* who proposed a general two-tier framework for Web archiving. Within this framework, crawling approaches for the acquisition of the surface Web are combined with deposit strategies for the so-called "deep Web", i.e. Web content that is not reachable by automatic crawling technology. A case study revealed only weak cooperation for the deposit by content providers, pointing at the need for an extension of the legal deposit framework. The study also analyzed ingest, validation, and metadata creation processes involved, as well as the necessity of migration steps for proper ingestion. .

The technical session was closed with a presentation by *Dona Bergmark* of *Cornell University*. She presented a project aiming at automatic collection building, collecting Web material on science and mathematics for the National Science Digital Library (NSDL). For this project, focus crawls were used, which rely on topic centroids for each collection topic to control the best-first crawling strategy. Additionally, tunneling was used to continue the search through a set of off-topic pages. A detailed statistical analysis of experimental results in this context was also presented during the main ECDL conference.

The afternoon session was devoted to presentations of four Web archiving projects, showing the diversity of requirements and approaches in this field.

Deborah Woodyard of the *British Library (BL)* provided an update on the "Britain on the Web" project, formerly known as "Domain uk". During this pilot study, a set of 100 sites from the United Kingdom was manually selected with respect to their expected long-term historical or cultural significance. Of those 100 site, 15% were not contactable, 43%

responded highly in favor of the archiving initiative, and actually none of the sites refused being included in the archival project. The sites were captured automatically at 3-week intervals and subsequently analyzed with respect to size and link characteristics in order to estimate scalability characteristics of the chosen approach.

In the next talk *Hans Liegmann* of the *German National Library (DDB)* gave an overview of the three prototype applications for archiving electronic documents based on submission following the traditional deposit-model. The "On-Line Thesis" project, incorporating 80 universities in Germany, requires participants to specify the metadata for new thesis publications, which are subsequently downloaded into the deposit server. For the "Springer Link" archive, housing a copy of about 500 journals by Springer, documents are provided to DDB on physical media. Furthermore, a generic submission and delivery interface has been developed. It allows publishers to directly submit documents and metadata according to specified submission criteria.

Birgit Henriksen of the *Royal Danish Library* reported on results from their "Netarchive.dk" project, which analyzed different archival approaches and the subsequent usefulness of the material for research. Results point towards a considerable re-thinking of the concept of a Website, as well as the limits imposed by national or domain restrictions in order to obtain useful collections. Based on the results of the study, she proposed a mixed strategy combining regular automatic crawls together with specific event-based collections. She furthermore stressed the necessity of stronger international collaboration on these issues.

Neal Beagrie of the *UK Joint Information Systems Committee (JISC)* concluded the session with a presentation of JISC initiatives with respect to building and archiving community collections, stressing 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* of the *Vienna University of Technology* on the European Web Archive (EWA) initiative. This initiative is currently being proposed by a consortium of about 30 partners from national and university libraries, research centers, and companies for an IST 6th Framework Integrated Project. The focus of this initiative lies with the creation of a distributed Web archive. It covers strategies and tools for a three-tier solution, corresponding to data selection and acquisition; archive maintenance and preservation; as well as access provision, archive exploitation, and data mining.

A second consortium was proposed by *Michele Kimpton* of the *Internet Archive* in cooperation with national libraries. Within this consortium, national libraries would define collection criteria and access requirements, with the Internet Archive providing tools and crawling services. A new Web crawler, targeted towards the archival needs of libraries, would be developed, and each participant's national Webspace would be crawled. The Internet Archive would also provide its holdings of the national data collections since 1996.

The workshop provided a good overview of the existing Web archiving initiatives, and pointed towards several open research issues, especially with respect to the acquisition, preservation, and analysis of the tremendous amounts of data involved. It again highlighted the need for an intensified international collaboration. Further information as well as links to the presentation are available via the Workshop Homepage at

http://bibnum.bnf.fr/ecdl/2002/index.html