Introduction

José Borbinha
INESC - Instituto de Engenharia de Sistemas e Computadores (Portugal)

Lex Sijtsma
KB - Koninklijke Bibliotheek (The Netherlands)

The DELOS Working Group is an action of the ERCIM Digital Library Initiative funded by the ESPRIT Long Term Research Programme, and one of its main targets is the periodic organisation of workshops addressing issues related with digital libraries.

The Sixth DELOS Workshop, dedicated to the subject of the "Preservation of Digital Information", was held in Tomar, Portugal, from the 17 to the 19 June 1998. The local organisation was a co-operation between INESC, which is a member of the ERCIM Consortium, and the National Library of Portugal. Both the organisations are also members of the NEDLIB consortium (http://www.konbib.nl/nedlib), a TELEMATICS project addressing the problem of the legal deposit and preservation of digital publications at national libraries and which was also involved in the workshop.

The event counted with 15 presentations and more then 40 attendees, coming from the Europe, USA and Australia. The presentations covered a broad range of issues, which created the required environment for an alive a very interesting debate of the problem.

Motivation

The workshop started with a motivating presentation coming from the USA, provided by Hans Rutimann from CLIR (http://www.clir.org). Hans presented the goals and priorities of "The Digital Library Federation", and he illustrated is talk with the very interesting video "Into the Future". This is a 30 minutes movie produced by the Commission on Preservation and Access and the American Council of Learned Societies, and it was shown last January on the national television in the USA with a great impact. "The movie sounds an alarm and raises questions" related with the problem of the preservation of digital information and the risks of the loss of memory for the organisations and the society in general, and it is really convincing in that purpose.

Strategies and technology

Two presentations reporting initial steps to address the problem come from the UK, supported by the eLib programme. Neil Beagrie presented the Arts and Humanities Data Service (AHDS), and the first results of their initiative for "Developing a Policy Framework for Digital Preservation" based on the life cycle of the resources. Michael Day, from the UKOLN, presented the CEDARS project, a complimentary perspective focused on the requirements for metadata to support digital preservation.

Alan Heminger, from the Air Force Institute of Technology (USA), raised the problem of the preservation of the knowledge and technology necessary to access actual digital documents in the future. In fact, currently available digital documents, although perfectly preserved, may not be readable by future systems. Even if the bits of the document can be preserved, the semantics can be lost. Solutions are the use of standardised, open system environments (e.g. SGML), conversion to other formats, keeping antiquated hardware operational and emulation. The last option involves the use of next-generation intelligent software to create viewers that enable you to get access at the document. However, to be able to do this you need the specifications of all those (old) formats and hardware/software environments. Alan suggests "The Rosetta Stone Model, a concept of a "metaknowledge" archive that collects such information. Such a task could be done by the National Libraries, National Standard Agencies or completely new organisations...

With a similar perspective, Dave MacCarn from the WGBH Educational Foundation (related with the public TV in the USA) presented "The Universal Preservation Format". Dave suggests that an alternative to the centralised storage model of the Rosetta Stone is to store a document jointly with all the information about its coding and logical structure. This is the approach of the UPF: Universal Preservation Format.

UPF has originated in the motion pictures and television business, and it is related with the need o to preserve and transfer images and their metadata from one system to another. For small documents the amount of metadata required by the UPF approach vastly exceeds the size of the data itself, thus requiring a lot of storage. A clue to solve that problem is the storage system of Norsam Technologies. They have built a disk-based system that uses nickel-coated disks with a diameter of 2 inches, which have a lifetime of thousands of years (they claim) and can hold the equivalent of one terabyte of data (a pile of approx. 210 *km* high of typed A4-paper). The information is recorded in the disks in an etching process, using a charged particles bean, and it can consist in any digital and/or human readable formats. To read it the only we need is an optical microscope. At the moment tests are being conducted together with the Library of Congress in the USA.

The role of the national libraries

PANDORA and EVA are two projects involving national libraries.

Judith Pearce, from the National Library of Australia, presented "PANDORA at the crossroads". The aim of PANDORA is to create an electronic archive of Australian publications on the Internet. Started in 1997, they now have a working prototype (*proof of concept*) with some 200 titles in it and growing with 10 titles/month. They used an approach that was both theoretical and practical, which is always a very effective way of working. PANDORA has put much effort in the definition of needed metadata. There is also a logical data model of the system and policies and procedures for each step in the archiving process.

From Finland, Kirsti Lounamaa (CSC) and Inkeri Salonharju (Helsinki University Library) presented "EVA, - The Acquisition and Archiving of Electronic Network Publications", an effort to harvest and index all the documents found in the WEB under the Finnish domain.

Digitising to preserve

Six projects, coming from the Switzerland, Bulgaria, Portugal, Greece and the UK addressed the issue of the digitisation and the usage of digital contents with purposes of preservation.

Kurt Deggeller, from Memoriav, presented the "Project VOCS - Voix de la Culture Suisse". In this project documents are stored and retrieved in a multimedia environment. VOCS aims to develop a system which "can store, search, consult and handle sound recordings in digitised form as well as other information (text, technical data, rights images…)". They have a prototype with some 300 sounds (100 GB), new sounds are being added and there is a Web interface for end-users. At the moment they have a problem in defining what kind of metadata to enter about the sounds.

"Vidion - An on-line archive for Video" is a Portuguese project involving INESC and RTP, the Portuguese Public Television. It was presented by Paula Viana (INESC), and it is concerned with the conversion, restoring and preservation of an audio-visual library of more than 400 000 documents representing more than 300 000 hours of video.

Related with the digitisation and/for preservation of still images, we had the presentations of Milena Dobreva, from the Bulgarian Academy of Sciences, and Ross MacIntyre, from the University of Manchester. Milena brought us the organisational, political and cultural problems and motivations, presented from the unique perspective of a country situated in the crossroads of the European geography and history. Ross presented the project to digitise actual and former issues of Nature "one of the most widely cited interdisciplinary science journals in the world today", and their plans for exploitation in the future.

ARIADNE, another Portuguese project presented by Nuno Maria (ICAT/FCUL) shown us how "Publico", one of the most important Portuguese newspapers, is addressing the problem of the production, management and long-term storage of their information, complemented with new perspectives and opportunities to extend their business in the information area.

Finally, "Beyond HTML: Web-based Information", presented by Chandrinos Kostas from FORTH (Greece), is a multi-user architecture to store and access to large image databases using the Web. They have been concerned with the metadata for archiving, annotate and retrieval of historical manuscripts.

Organisation and access

To preserve information means also to be able to organise, find and access it, now and in the future.

From this perspective, Abdel Belaid, from France, (LORIA-CNRS) told us about the "Retrospective Conversion of Old Bibliographic Catalogues", a very well know problem faced by almost all the big libraries in their transition from the traditional card based to the computerised catalogues.

With clues about the new ways we can build and use those new "computerised catalogues", we had the presentations referring an "Effective Terminology Support for Distributed Digital Collections" (Martin Doerr - FORTH / Greece), and "TopicMark: A Topic-focused Bookmark Service for Groups" (Hui Guo - GMD / Germany).

A practical case

The last presentation of the workshop was "Preserving the U.S. Government's White House Electronic Mail: Archival Challenges and Policy Implications", and it was given by David Wallace, from the School of Information of the University of Michigan (USA). David presented us with a practical problem, involving the backups of the electronic mail in the White House during the Reagan Administration and their importance in the case Iran-Contras (the case is still in court).

For more than a decade now there are cases into court in the US about the creation, use, management and preservation of electronic mail messages from the government. The big problem seems to be that the US government has a record keeping system based on paper records, not covering electronic contents. The government attitude was that electronic mail is more or less comparable to a telephone message, not used to create official documents. In the case that an email message becomes official the policy is that it should be printed and filed using the usual (paper based) channels. But a coalition consisting of the National Security Archives and others claim the printed and electronic versions of a document are not the same. For instance, we miss all the kinds of metadata and context in the printed version. This is just one example of the problems we can encounter with electronic documents.

Because the case has been in court for quite a while we already see that problems as discussed above in the Rosetta Stone presentation are beginning to appear. At the moment nearly 6000 backup tapes plus some 150 hard disks are stored, under the order of the court! Tapes can be read, but the content is not recognisable anymore because of changes in hardware and software; some hard disks had special security features on them and at a certain point in time there was only 1 PC left in the US government that could read those disks; ...

This was a very interesting presentation. It just shows how new this really still is that apart from just preserving digital information there is still a lot of work to do with respect to selection and the establishment of proper procedures for filing electronic document. By the way: because of this currently 90-95% of the electronic information of the US government gets lost...