ERCIM-98-W003
INESC

European Research Consortium
for Informatics and Mathematics

ERCIM

# Sixth DELOS Workshop

# Preservation of Digital Information

Tomar, Portugal, 17-19 June 1998

DELOS Working Group Reports

# 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.

# "The Digital Library Federation - Goals and Priorities"

Remarks by Hans Rütimann, International Program Officer
Council on Library and Information Resources
Washington,D.C.

## Background

The Council on Library and Information Resources (CLIR) is the outcome of the merger of the Council on Library Resources (CRL) and the Commission on Preservation and Access (CPA). CLIR's mission calls on the organization to identify the critical issues that affect the welfare and prospects of libraries and archives, to convene individuals and organizations in the best position to engage these issues, and to encourage institutions to work collaboratively to achieve and manage change. CLIR pursues its mission out of the conviction that information is a public good and of great social, intellectual, and cultural utility. It has set its sights on a few targeted programs: the Commission on Preservation and Access retains its identity as a program of CLIR, along with programs for Digital Libraries (the Digital Library Federation), the Economics of Information, and Leadership.

As a fundamental principle of all its programs, CLIR will encourage institutions to achieve and manage change through collaboration, and collaborative action is particularly important within the preservation community. The most striking evidence of this is the success of the ongoing effort to rescue through microfilming large portions of the deteriorating print-based collections in the U.S. and abroad. Since its inception, the Commission on Preservation and Access has worked to assure that knowledge produced by the scholarly communities of the world is saved and kept accessible and it will continue that role.

With advice from its standing committees and task forces, CLIR publishes materials that inform and instruct the preservation community, document the economic implications of establishing sound preservation environments for collections, frame the next set of issues to be considered within the changing definition of "preservation and access," and develop new strategies to sharpen the professional skills of individuals with preservation responsibilities. A few examples of recent publications indicate the range of our concerns:

- "Digitizing Historical Pictorial Collections for the Internet"
- "Preservation and Archives in Vietnam"
- "Digitization as a Method of Preservation?"
- "Mass Deacidification: An Update on Possibilities and Limitations"

## The International Program

Because few of the critical issues of preservation and access today can be addressed without an international focus, CLIR maintains an International Program to help promote preservation awareness throughout the world. Through training seminars, workshops, translation projects, publications, and a policy of generous response to requests for counsel and advice from colleagues abroad, it continues to promote the long-term preservation and accessibility of information.

The International Program has focused its efforts to date on Eastern and Western Europe, the former Soviet Union, China, and Latin America. Thanks to support from The Andrew W. Mellon Foundation, the work in Latin America will be extended, and CLIR will initiate new activity in Southern Europe (for example, in Greece) and in South Africa. The Program also hopes to expand in Asia and is seeking funds to develop new projects there. Examples of activities that will be undertaken in the next several years include providing expertise for preservation-needs assessments in libraries and archives abroad; the development of cooperative filming and digitizing projects for specific collections; and designing strategies to increase the production and use of permanent paper.

Two specific examples:

The International Register of Microform Masters: An international register of microform masters is essential if

scholars and librarians are to know what has been filmed at locations throughout the world and to avoid duplication of effort. The Program encourages libraries and archives to contribute records to regional nodes for the collection, organization, and distribution of information about reformatted collections. It supports efforts to link these records to an emerging international register and to reach international consensus on the elements and the record structure for listings of digitized materials. A successful example is the European Register of Microform Masters (EROMM) where through ten partners in nine countries, some 40 European libraries have so far contributed more than 400,000 records of microfilmed items to the database. An exchange arrangement with the U.S. bibliographic network RLIN allowed the addition of another 1.9 million records to EROMM's database..

Another example is the recently concluded project "Translation and Dissemination of Preservation Knowledge in Brazil." Access to information often means translation of professional literature into other languages. An inter-institutional alliance of interested organizations in Brazil guided the project, which included the translation into Portuguese of 52 titles of preservation literature, from environmental control to digital conversion. The translations (plus videos) formed the basis for workshops throughout the country. In the process, the project's coordinators collected valuable information about the state of collections in more than 1,400 libraries and archives.

During a recent meeting (Aveiro, Portugal) with librarians and archivists from Lusophone countries, we offered to make all these materials available in Portugal, Portuguese-speaking countries in Africa, and Macau. Since access to professional literature on basic preservation concerns for print, sound, and images is a top priority in most countries, the offer was received with enthusiasm.

I mention support for traditional preservation efforts for print, images, and sound at some length since there is, particularly in developing countries, a danger of putting too much faith in digital solutions for the preservation of a country's heritage. Librarians and archivists in developing countries are often unaware that digital storage for long-term archiving of information requires careful planning, that many organizational, technical, and organizational issues have not yet been resolved, and that digitization is not a means for preservation unless a long-term plan assures the survival of digitally stored information. What is really disturbing is that in the general excitement about all things digital, many institutions have put on hold traditional and basic preservation activities.

Also, preservation concerns did not start with digital information. There is a long tradition of preserving cultural heritage. The recognition that much of the printed documentary record is in jeopardy because of the introduction of acidic paper around 1850 led to a widespread preservation movement, similar to the one we're beginning to witness for digital information. The preservation community, especially preservation managers and administrators, have much experience in selecting and preparing collections, and coordinating massive collaborative reformatting projects, mostly microfilming. Their experience is useful even though we're now dealing with an entirely new medium.

## The Task Force on Archiving of Digital Information

An important contribution to the debate over long-term archiving of digital information came from the Task Force on Archiving of Digital Information, organized jointly by the Commission on Preservation and Access and the Research Libraries Group (RLG). The Task Force, in effect, was asked to report on ways that society should work with respect to the cultural record it is now creating in digital form -- not only through conversion from print to digital but, perhaps more important -- through a record that is "born digital." Consider that, by the year 2000, an estimated 75% of all U.S. Federal transactions will be handled electronically. When President Clinton leaves office, it is estimated that his administration will hand over eight million electronic files to the National Archives; and these administration files are but a minuscule portion of the electronic record generated by the corporate world. Most of this information is born digital and there is no printed record to fall back on.

The Task Force's 1996 report argues that "the problem of preserving digital information for the future is not only -- or even primarily -- a problem of fine tuning a narrow set of technical variables." Rather, "it is a problem of organizing ourselves over time and as a society to maneuver effectively in the digital landscape. It is a problem of building -- almost from scratch -- the various systematic supports, or deep infrastructure, that will enable us to tame our anxieties and move our cultural records naturally and confidently into the future" (Task

Force 1996:6)

The Task Force's report is available in full at the website of the Research Libraries Group (http://www.rlg.org). The Task Force was co-chaired by Donald J. Waters, now the Director of our Digital Library Federation.

Among the report's conclusions and recommendations are:

- The Task Force recognizes that most of the challenges associated with digital preservation are organizational, not technical.

- The first line of defense against loss of valuable digital information rests with the creators, providers, and owners of digital information.

- Certified digital archives must have the right and duty to exercise an aggressive rescue function as a fail-safe mechanism for preserving digital information that is in jeopardy of destruction, neglect, or abandonment by its current custodian.

The final report focuses on three essential questions: What does digital preservation entail? How do we organize ourselves to do it? What steps should we take to move forward? Time will not permit covering these questions in detail, but one major conclusion of the report cannot be emphasized enough: Our greatest challenges in the digital age are organizational rather than technical. Because we currently lack the infrastructure of practices, standards, and organizations needed to support preservation of digital information, the following elements of infrastructure must be considered:

- Legal bases for deposit and rescue. Nationally and internationally, legislation and agreements are needed to encourage legal deposit of electronic resources in archival repositories, to enable rescue of abandoned resources, and to facilitate access and use of archival files.

- Standards for description. Current library cataloging standards are not sufficient to describe access and contextual information about digital resources. Several efforts to address this issue are underway internationally. For example, existing registers of microform masters are examined for expansion to include digital items. In this context, see also the final report of the *RLG Working Group on Preservation Issues of Metadata* (http://www.rlg.org/preserv/presmeta.html).

There is more, and the need for sharing information about best practices across the wide spectrum of communities is overwhelming. Preservation of digital materials has emerged as a new, critically important field of interdisciplinary and international activity, and much work needs to be done.

## [Introduction to the film "Into the Future"]

[Addressing our digital memory crisis begins with dialogue. CLIR prepared a discussion paper on digital preservation. Copies of the discussion paper are available, but first, if I may, I would like to show the documentary movie "Into the Future," on the preservation of knowledge in the electronic age. The movie was commissioned by the Commission on Preservation and Access and the American Council of Learned Societies. Funding was provided by the Alfred P. Sloan Foundation, the National Endowment for the Humanities, and the Xerox Corporation. It was shown on U.S. national television last January and created much discussion. The mainstream press picked up on the issue and several thoughtful articles appeared on the subject of digital preservation. The discussion paper raises the question "Why should we be concerned?" "Into the Future" provides several answers.]

*Showing of the 1/2 hour version of "Into the Future."*

After the showing: The movie sounds an alarm and raises questions; it does not provide answers. We hope that the answers will eventually be provided by groups such as represented at this conference, the Digital Library Federation (DLF) and its members, and other groups and individuals.

## The Digital Library Federation (DLF)

(For this part, I'm relying extensively on information provided by Donald J. Waters, Director of the Digital Library Federation. For more information about the DLF, see CLIR's website (http://www.clir.org/).

The Council on Library and Information Resources is administrative home to the Digital Library Federation, which includes 19 members -- university research libraries, the Library of Congress, the National Archives, and the New York Public Library.

The primary mission of the DLF is to establish the necessary conditions for creating, maintaining, expanding, and preserving a distributed collection of digital materials accessible to scholars and a wider public. Participants in the Federation are committed to a shared investment in developing the infrastructure needed for libraries of digital works. The infrastructure is intended to enable digital libraries to bring together, or "federate," the works they manage for their readers.

The DLF has set the following program priorities:

- DLF will focus on libraries of materials born digital. It is critical that the library community moves from conversion -- except in specialized and well-justified cases -- to organization, access, and preservation of materials born digital. A high priority is developing the archival mechanisms that preserve the integrity and usability of these digital works over a long term. The Federation is increasingly turning the attention of libraries to the numerous -- and difficult -- issues associated with works born in, rather than converted to, digital form.

- DLF will help integrate digital materials into the fabric of acadmic life. A critical focus for such integration is to define the circumstances under which conversion to digital form is justified. Conversion projects that facilitate the extension of higher education and promise to improve the quality and lower the cost of research and education deserve special attention.

- DLF will help stimulate the development of a core digital library infrastructure. The highest priorities for attention at the present time are the network and systems requirements and means of authentication and authorization, the means of discovery and retrieval, and archiving.

- DLF will help define and develop the organizational support needed for effectively managing digital libraries. Organizational issues requiring early attention include identifying institutional values and strategies for managing intellectual property in digital form, and creating the conditions for the development of the professional skills needed for digital library management.

## CLIR and DLF Plans for Digital Archiving

As one of its primary agenda items, CLIR and the DLF aim to ensure the persistence of digital information, both born digital or created by conversion. Much has been written about the need to document digital information in order for it to serve the role of "record" in an archival sense, as well as about the need to "migrate" digital information to new media in order to prevent loss from media decay and obsolescence. However, relatively little effort has gone toward answering the more fundamental question of how to ensure that digital documents will remain readable and understandable in the future.

At this time, one might usefully distinguish the following strategies:

1) Copying (no changes to the information)
2) Migration (adpation to new hardware and software)
3) Emulation (new platforms mimicking previous platforms)
4) Archeology (doing nothing and trying to salvage neglected information when needed)

CLIR and the DLF commissioned Jeff Rothenberg, author of "Ensuring the Longevity of Digital Documents" (*Scientific American*, January 1995) to investigate approaches currently being considered to ensure the future accessibility and readability (i.e., longevity) of digital material. His preliminary findings challenge the concept of "migration." Since migration involves translation of each different type and format (text, images, sound,

video, animation, etc.), the process is highly labor-intensive. If a document is accepted for long-term storage based on an assumed migration strategy, Rothenberg argues, it is impossible to give even gross estimates of what will have to be done to it in the future, when this will have to be done, how much it will cost, or how long it will be before the document is corrupted by inappropriate translation.

In his further research, Rothenberg will concentrate on emulation as a solution to the problem of digital preservation, one that is predictable and testable and eliminates repeated translation of documents "which must inevitably lead to their corruption and loss." Others argue that "migration" is, at least for the shorter term, the only viable solution. Stay tuned.

Other activities of the DLF that relate to archiving are:

- Digital collections of licensed works. Libraries are pouring substantial resources into developing licenses for access to electronic journals, while still maintaining subscriptions to print copies of the same materials. Not all contracts provide for long-term access to the licensed materials, but terms in some licenses require the publisher to provide the library with a tape or CD-ROM containing a copy of the licensed work. The Copyright Division of the Library of Congress is developing the means to accept archival quality versions of copyrighted works deposited there, but project work is needed to ensure that the tapes and CD-ROMS deposited with libraries and archives as archival copies can and will serve the fail-safe purpose for which they are intended.

- Institutionalizing digital information. Much scholarly information in digital form is currently being managed as a cottage industry by individual faculty on behalf of a particular discipline. The directors of the DLF institutions are considering how DLF can provide leadership in bringing products emerging from this cottage industry into a more stable, institutional environment.

- Cornell has developed a proposal that DLF will support, to explore the requirements and means for preserving materials converted to digital form in a subset of genres.

- Extension of reach: *Making of America* Phase III. Conversion, as the Library of Congress has demonstrated, is a significant means of extending the reach of library collections in the service of general education. One of the founding goals of the DLF is to find ways to aggregate the existing digitized collections. This effort will focus primarily at the level of descriptive metadata as a means of integration and provide a testbed for exploring intersystem searching methods. There is another important dimension to the project: In several institutions (e.g., Yale and Cornell), large quantities of digitized Americana are inaccessible because the digital platforms on which the collections were built have become obsolete. *Making of America* Phase III thus affords the opportunity to develop and demonstrate migration techniques as a means of preservation of digital information.

## Conclusion

Perhaps we should add a fifth strategy to the four already mentioned -- prayers. James Gleick reported in *The New York Times* ("The Digital Attic: An Archive of Everything," 12 April 1998) that "the Daiho Temple of Rinzai Zen Buddhism held a `memorial service for lost information' in Kyoto and online." Gleick adds that "of course, the details are lovingly preserved, in English and Japanese, at its web site. A look at the web site reveals that there is much common ground between this conference and the Rinzai Zen Buddhists (http://www.thezen.or.jp/jomoh/kuyo.html):

"After the effort of transforming all this knowledge into electronic information has been completed, is it enough then to say that we are finished? And from there, can we truly make effective use of that which we have created? Sometimes, the answer is `no.' To provide an example, there are many `living' documents and softwares that are thoughtlessly discarded or erased without even a second thought. It is this thoughtlessness that has drawn the concern and attention of Head Priest Shokyu Ishiko. Head Priest Ishiko hopes that through holding an `Information Service' and by teaching the words of Buddha, that this `information void' will cease to exist."

We will all have to help him.

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...

# Developing a Policy Framework for Digital Preservation

Neil Beagrie
Arts and Humanities Data Service Executive
King's College London
Strand
London WC2R 2LS

**Abstract**

The Arts and Humanities Data Service (AHDS) has been established by the Joint Information Systems Committee of the UK's Higher Education Funding Councils to collect, preserve and promote re-use of digital resources which result from or support research and teaching in the arts and humanities.

Within the UK, the Digital Archiving Working Group (DAWG) has been formed to co-ordinate research into digital archiving between its members: the British Library, the National Preservation Office, the Higher Education sector, the Public Record Office, the Research Libraries Group, and the Publishers Association. DAWG has commissioned a series of studies to examine issues raised in the CPA/RLG Task Force report on Digital Preservation.

As part of this research programme the Executive of the AHDS has recently completed a study into developing a strategic policy framework and implementation guidance for the creation and preservation of digital resources. The framework is based on the life-cycle of a resource from its creation, management and preservation through to use, and examines the dependencies and issues at each stage, how these are interlinked, and the influence of the legal and business environment. The framework and the implementation guidance have been based on interviews with representatives of those organisations which have a major stake in the creation or long-term maintenance of digital research and on an extensive literature survey. Institutions interviewed included a selection of international libraries, museums, archives, scientific and other academic data archives. Together they represent a synthesis of work to date on digital preservation and point directions toward the development of robust preservation strategies.

This paper presents a summary of the major findings and the policy framework from the study report.

## 1. Introduction

### 1.1 The Importance of Preservation and Access to Digital Information

Computerisation is changing forever the way information is being created, managed and accessed. The ability to generate, easily amend and copy information in digital form; to search text and databases; and transmit information rapidly via networks world-wide, has led to a dramatic growth in the application of digital technologies to all areas of life. Increasingly the term "Information Age" is being used to describe an era where it has been estimated we have created and stored one hundred times as much information in the period since 1945 as in the whole of human history up to that time.

Digital information forms an increasingly large part of our cultural and intellectual heritage and offers significant benefits to users. At the same time preservation and access to this information is dependent on impermanent media and technologies; retaining metadata on the provenance and context; and retaining the authenticity and content of the resource. Although experience in creating and managing specific forms of digital data has been built up over a number of decades in the sciences and social sciences, in many areas it is a relatively new medium where much of the future life-cycle, activities and cost models are currently unknown. These factors have led to increasing concern about the potential loss of our "collective memory" in the Digital Age and have prompted further research into the long-term preservation of digital information and maintaining future access to it.

Substantial digital preservation initiatives are currently underway in Britain, for example at the British Library, the Public Record Office, the Data Archive, the Natural Environmental Research Council, and the Arts and Humanities Data Service. Further initiatives are contemplated by the Joint Information Systems Committee, by the British Library, and by individual heritage and educational agencies which find themselves increasingly concerned with long-term preservation of the digital information resources which they are helping to create or archive. Growing British interest in digital preservation is complemented and shared internationally for example by the work of the Commission on Preservation and Access, the Research Libraries Group, and the National

Archives and Records Administration in the US;by the National Library and National Archives of Australia; and by various initiatives in Europe such as the DLM-Forum, and elsewhere.

## 1.2 The Importance of a Policy Framework

The challenges posed by digital information have increasingly led to recognition of the inter-dependence between the stages of creation, use and preservation of digital resources and the importance of the legal and economic environments in which they operate. The potential volume of information which could be acquired or digitised, and the need to make the most cost-effective use of limited resources, have emphasised the need for selection, standards and co-operation between different organisations. Organisations are developing internal policies for the creation, management, and preservation of digital resources and increasingly are sharing their experience in this field.

A key part of this shared experience has been the recognition of the importance of the life-cycle of digital resources and the complex inter-relationships between different practices which may be adopted to create, use or preserve them. Digital preservation is crucial as part of a series of other issues which effect the creation, storage and use of a resource. These issues are all inter-dependent and have suggested the need for an integrated policy framework to develop a cost-effective approach resource creation, preservation and use.

An integrated policy framework may also assist funding agencies in maximising their scholarly and financial investment in the creation of primary and secondary data resources, and data creators in maximising the cost-effectiveness, fitness for purpose, and design, of their digitisation programmes.

## 1.3 The DAWG Programme of Preservation Studies

In 1995 a workshop was held at Warwick University to consider the long-term preservation of electronic materials (Fresko 1996). The workshop was convened to consider issues raised in the draft report of the Task Force on archiving of digital information commissioned by the Commission on Preservation and Access and the Research Libraries Group in the US and published in the following year (Garrett and Waters 1996). The workshop made a number of recommendations for further investigation and research within the UK. The Joint Information Systems Committee subsequently agreed to fund a research programme implementing the recommendations, to be guided by the Digital Archiving Working Group (DAWG) in the UK and administered by the British Library Research and Innovation Centre.

## 1.4 Aims of the AHDS Study

The study undertaken by the AHDS Executive (Beagrie and Greenstein forthcoming) as part of the DAWG Programme of Preservation Studies, aims to identify current practice, strategies and literature relating to the creation and preservation of digital information and to provide the integrated policy framework and guidance, which many believe are crucial to long-term preservation of digital resources.

The study aims to provide a strategic policy framework for the creation and preservation of digital resources, and to develop guidance based on case-studies, further literature and ongoing projects which will facilitate effective implementation of the policy framework. The framework itself is based upon the stages in the life cycle of digital resources from their creation, management and preservation, to use, and the dependencies and inter-relationships between these stages and the legal, business and technical environments in which they exist. The case studies and other guidance incorporated in the report have been developed to illustrate how the framework can be used and applied by different agencies who may have different roles and functions, and in some cases direct interests in only part of the life-cycle of the resource.

The intended audience for the study therefore encompasses all individuals and organisations who have a role in the creation and preservation of digital resources from the funding agencies, researchers and digitisers and publishers, through to the organisations which may assume responsibility for their long-term preservation and use.

## 1.5 Methodology

The study was carried out by the author and Dr Daniel Greenstein (Director, AHDS Executive) between December 1997and March 1998. It was based upon traditional desk-based research methods and on fifteen structured interviews. The former involved extensive and growing literature, much of it available freely on the World Wide Web, and also in subscription-based print and electronic journals, and trade association newsheets. Crucially it also took account of the policies and programmes which large-scale digital preservation and digital collection development initiatives are beginning to provide in some "published" format.

Structured interviews, conducted in person or over the phone or by email, involved senior data managers and specialists working in organisations both in the UK and overseas with experience in digitisation, data management or the long-term preservation of digital information resources. Interviewees were selected to provide a wide cross-section of experience of different media types, and experience in different sectors such as national museums, archives, and libraries; university computer centres and data archives; scientific data centres; and research libraries.

## 2. The Policy Framework

### 2.1 The Development of the Policy Framework

The starting point for the study was a draft policy framework. This represents selected elements of a generic collections policy developed for the Arts and Humanities data Service (AHDS), a distributed national service and collection established by the Joint Information Systems Committee of the Higher Education Funding Councils in the UK.

The AHDS is a multi-disciplinary service with five service providers covering archaeology, history, literary and linguistic texts (the Oxford Text Archive), performing arts, and the visual arts, with a remit to collect, catalogue, manage, preserve, and promote the re-use of scholarly digital resources. Its collections policy was therefore developed to cover a wide-range of subject disciplines and different digital media, and provided a valuable starting point for the study. Further information on the AHDS and the AHDS collections policy is available from the AHDS website at <http://ahds.ac.uk/>.

The AHDS collections policy applies the concept of the life-cycle of a digital resource, which has been widely used in the records management and archival professions (eg European Commission 1997a, 1997b) as part of the framework used for its construction. The policy framework outlined below also employs the concept of the life-cycle of a digital resource. It has extended and enriched the draft framework to reflect the perspectives, experience and roles of other stakeholders who can be involved in the creation and preservation of digital resources, as identified in the study interviews and the literature search.

### 2.2 How to use the Framework

The framework outlines the three main stages (creation, management / preservation, and use) in the life-cycle of a digital resource, the role and functions of different generic stakeholders within this, and the inter-relationships between each stage and the implications for preservation of those resources with long-term cultural and intellectual value.

The inherent properties of digital resources mean that the processes of data creation and long-term preservation will involve a wide range of individuals and institutions which have a short-term or even indirect interest, as well as including institutions with a traditional role in these processes(see 4.2 Applicability and Scope below). The framework therefore identifies the roles and functions of different generic stakeholders so that individuals and institutions can see how they and others fit into the framework. Use of the framework may thus facilitate effective collaboration between different stakeholders over the life-cycle of the resource. The life-cycle of the resource is also heavily influenced by the legal and business environment, so the framework explains the influence of these factors and how they may shape the creation, management, and use of the resource.

To use the framework in drafting strategic policies or implementation guidance the user should "walk through" the framework considering the aims they are trying to achieve, the issues and other players at each stage in the life-cycle of the resource, and how they will be influenced by the legal and business environment in which they operate. The framework therefore effectively provides a high-level checklist which individuals and institutions can use to develop policies and guidance which they will tailor to their specific function or role and environment. In so doing they will also identify the implications across each stage, and the impact on or made by other players involved. The overall effect should be to provide policies and implementation strategies where the cost/benefits have been fully explored and strategic partners or dependencies identified.

The Case Studies included in the study report are intended to illuminate this process further by providing a synthesis of the existing practice, policies and implementation strategies of those interviewed for the study. The Case Studies show how issues have been approached in practice and how different organisational missions shape approaches to creation and preservation of digital resources. This can then be elaborated further by reference to the additional bibliography and references in the report.

### 2.3 Applicability and Scope

The study is concerned with the creation and long-term preservation of our cultural and intellectual heritage in digital form.

For the purposes of digital preservation, long-term can be defined as beginning when the impact of changing technology such as new formats and media needs to be addressed and extending indefinitely thereafter. In a digital environment, the framework an d preservation will therefore include institutions with a traditional interest in long-term preservation but will also extend to a wider range of individuals and institutions which have a short-term or even indirect interest in this process.

The digital information covered by the framework can be the primary form of the data, surrogate versions of primary information held in digital or physical form, or the metadata for collection management of these objects. The framework recognises that digital media are new, distinctive, and require new approaches to their preservation. At the same time it recognises that these approaches may need to be integrated with those for other media and, where relevant, should draw on the existing and extensive professional experience in managing them. It recognises that individuals and organisations may be responsible for hybrid resources consisting of a mixture of digital and other media, or solely focussed on information in a digital form. The framework will therefore be applicable to those seeking to extend and modify existing policies for traditional collections to include digital information and for those developing data policies for purely digital collections.

Digital information can be generated by a number of different processes and for different purposes each of which is considered by the study. The information may exist in a definitive version and be generated by a project or business function with a finite timespan; or it may be dynamic, constantly evolving, and generated by a project or business function with no finite timescale. The purpose for which it is created and preserved may also vary from digitisation of existing information to improve access and/or preservation of existing collections; to the collection of existing digital information and its preservation for future re-use and research.

The chapter of case studies introduces a range of stakeholders and organisational roles in the creation, management and preservation of digital resources encountered during the study. Individual institutions need not be confined to a single role but normally a single role was found to have a greater influence on its approach to data creation, management and preservation, and use. These roles are described in greater detail later in the report and can be summarised as follows:

funding agencies

"digitisers" including research-oriented agencies and individuals, many library and cultural heritage organisations, and publishers

"data banks" archiving digital information at the bit level usually under contract for a third party

institutional archives managing unique electronic records generated by a single organisation

academic data archives maintaining and encouraging re-use electronic resources of interest to specific academic communities

legal deposit or copyright libraries with a statutory obligation to maintain and provide access to non-unique information objects

The information landscape covered by the framework is therefore rich and varied and its implementation will be tailored to the specific needs and responsibilities of individuals and institutions. However whatever the needs and responsibilities, we believe those individuals and institutions will benefit from considering the framework in developing appropriate policies and implementation guidance. In addition it is our belief that the roles of different stakeholders in long-term preservation of the cultural and intellectual heritage cannot be achieved without consideration of the life-cycle of the resource and the co-ordination of the separate interests as embodied in the framework.

2.4 Legal and Economic Environment

This is not a stage in the life cycle of a digital resource but a consideration of the legal and economic environment surrounding the resource and interlinked with the organisational mission of its stakeholders which will also impact on the life-cycle and the application of the framework.

Legal issues may include: intellectual and property rights in the resource or integral software supplied with it; contractual terms attached to a resource or the hardware and software needed to access it; protecting the confidentiality of individuals and institutions; protecting the integrity and reputation of data creators or other stakeholders in the resource; or any legal obligation to select and preserve the authenticity and content of categories of records or individual resources. What rights are vested in a resource will impinge on how and whether it may be represented in machine-readable form; how, by whom, and under what conditions it may be used; how it can and should be documented and even stored (e.g. where 'sensitive' information requires encryption or access restrictions); and how, whether, and by whom it can legally be preserved.

Similarly the business environment(s) in which a resource is created, managed, preserved and used will have a bearing on the application of the framework. Resources created in a commercial environment may have a commercial life-cycle which can impinge on data management, preservation, and use. Some organisations may also be subject to more sudden and abrupt changes in ownership and rights, or location and data management than others.

The returns required on investment in resources may also require physical control of storage and access, and/or systems and procedures for encrypting, marking or locking the resource, user registration and authentication, charging, and rights management. All of these can affect and in some cases can mitigate against long-term preservation unless they are specifically addressed as issues and the requirements of different stakeholders can be met.

The priorities and objectives of funding, and the funding agencies, for the resource through the life-cycle can also vary and impact in a number of different ways. This is particularly important for documentation and metadata on the context and content of the resource which are most easily developed or captured when the resource is created and can only be re-constructed at greater expense, if at all, at a later stage of management and preservation.

The cost-effectiveness over the life-cycle of the resource of completing data documentation and metadata when the resource is created (and often its immediate benefits to the data creator) needs to be recognised and its practice encouraged.

## 2.5 The Life-Cycle of the Resource

### Data creation

Data creation will normally involve a design phase followed by an implementation phase in which the data is actually created. Consideration of the framework will have its greatest benefits during the phase of developing funding, research and project designs, design of information systems, and selection or development of software tools.

The decision to create digital resources can be undertaken for a number of different purposes and involve a range of stakeholders who will have some influence on the process. Data creation may be undertaken by those creating information from its inception in digital form (primary data creators), or by those involved in the creation of digital materials from information in traditional media (digitisers). The timescale for creation of these digital resources can be finite and definitive or dynamic and continuous.
In some cases hybrid resources incorporating both digital and traditional media may be created or the resource hyper-linked to other resources.

Each of these processes and the form of resource entail a range of decisions which will involve selection and determine a data resource's cost, benefits, intellectual content, fixity, structure, format, compression, encoding, the nature and level of descriptive information, copyright and other legal and economic terms of use. Accordingly how data is created and its form will impinge directly upon how it can be managed, used, retained and preserved at any future date. All or most of these criteria will also determine a resource or collections usefulness to the data creator and funding agencies and its fitness for its intended purpose.

The process of data creation by individuals or institutions may be influenced by a number of different stakeholders. Funding agencies, publishers, and software developers can influence or determine different aspects of the decision process. Curators interested in the development of policies and guidance for the creation and long-term preservation of the resource should therefore identify strategic partnerships and dependencies and ensure that these are addressed. This will usually involve developing a dialogue with internal or external data creators, users and other stakeholders, and considering the implications of how a resource has been created and documented for its management, preservation and future use.

Data and collection management and preservation may involve a number of stakeholders who can fulfil different functions and roles. These functions and roles may be for a fixed or indefinite duration and can involve direct or indirect participation in the process. Immediately after creation of the data and usually for a period after this the primary data creators and digitisers will be responsible for the management and short-term preservation of the resource. The resource can also be deposited or w ill be transferred at a subsequent point to institutions or internal departments which will support or assume responsibility for long-term preservation and access.

These functions can be undertaken by internal departments within the digitisers where their organisations' roles extend to long-term preservation. Alternatively these functions will be achieved by offering to deposit with and/or acquisition of the resource by the institutional archives, copyright and deposit libraries, and academic archives.

In addition, digital information may be created as part of the process of collection building or collection management of a resource. This can be seen as an extension or supplement to data creation process and similar criteria will apply. Collections m ay be extended or new aggregations of resources created by licensing, copying or mirroring existing digital information created by others. New digital information can also be created in collection management processes e.g. the computerised cataloguing or digital research materials generated from existing resources in digital or traditional forms.

In some cases the resource or collections may be managed and preserved by administrative processes which we have described as "remote management". For dynamic constantly changing information, a single deposit and acquisition for long-term preservation may be inappropriate. In such cases digital information may remain with the data creator who will assume responsibility for updating and maintaining it. The primary data creator may be legally obliged or voluntarily abide by standards and procedures established by an external organisation with established procedures for deposit. Decisions may be taken to periodically sample or copy the resource which will provide an archive of the resource at particular points in time.

"Active" resources which are still used by their creators in a current project or business process may be managed and preserved by a similar process of remote management in which the data creators abide by standards and procedures agreed with and monitored by an external organisation. In such cases the data may be reviewed and selected for deposit and acquisition when it is no longer in an active phase of use by the data creator. Alternatively a copy of the data may have been deposited during this active phase but access may be denied or restricted for an agreed period.

The organisations we have identified as "data banks", and to a more limited extent other organisational types, may also be involved as contractors in remote management of resources. They frequently manage resources under contract to others who retain legal responsibility for the resource and set terms and standards in the contract for their management.

Data management and preservation involves organisational decisions about whether collections or parts of collections are stored centrally or distributed across several sites, contracted out to a data bank, or the technical decisions about what magnetic media and hardware platforms, physical security, refreshing or replacement of storage media, and contingency procedures, are used. Options are constrained by the resources' structure format, compression, and encoding; by whether the resource is dynamic or fixed in its nature; the need to maintain authenticity and integrity of the resource; and also upon the relative emphasis given to their use and/or preservation. Accordingly data storage decisions together with the available funding and technologies can constrain data creation or acquisition and help to determine how (even whether) and to what extent a data resource once included in a collection can be preserved and/or used.

Long-term preservation is highly contingent on decisions taken when the resource is created and during its subsequent management, and also rests on available funding and technologies. It is also undertaken to maintain future access and use of the resource and is therefore closely linked and potentially contingent upon data use.

Data Use

Data use can occur immediately after its creation and for an indefinite period thereafter. Its use can be to fulfil its primary purpose when created, involve subsequent secondary analysis, or inclusion in a collection developed to fulfil other aims. The primary data creators, digitisers, funding agencies, publishers, institutional archives, copyright and deposit libraries, academic archives and their user communities may all be involved in data use or defining and servicing user requirements. Use of the data will be highly contingent on the decisions made and circumstances surrounding creation, management and preservation of the resource; the rights management and

economic framework which applies, and the approaches taken to identify and reconcile the needs of different stakeholders.

How data is delivered to and used by end users will be contingent upon: how and why it was created or acquired; agreements to co-operate, share or exchange data between different institutions; conditions and procedures required to meet legal and economic requirements; how/where it is stored; and upon what software and hardware is needed to access it. It's use over extended periods of time will also be contingent on decisions made on data management and preservation.

**3. Conclusions**

Digital information forms an increasingly large part of our cultural and intellectual heritage and offers significant benefits to users. The use of computers is changing forever the way information is being created, managed and accessed. The ability to generate, easily amend and copy information in digital form; to search texts and databases; and to transmit information rapidly via networks world-wide has lead to a dramatic growth in the application of digital technologies.

At the same time the great advantages of digital information are coupled with the enormous fragility of this medium over time compared to traditional media such as paper. The experience of addressing the Year 2000 issue in existing software systems, or data losses through poor management of digitaldata are beginning to raise awareness of the issues. Electronic information is fragile and evanescent. It needs careful management from the moment of creation and a pro-active policy and strategic approach to its creation and management to secure its preservation over the longer-term. The cost structure for securing the cultural and intellectual work of the digital age will be notable and has to be built in at the beginning if these costs are to be minimised and that investment effectively applied. There will be many stakeholders and interests in a digital resource over a period of time. A strategic approach is needed to recognise, address, and co-ordinate these interests and secure the future of digital resources.

The framework elaborated by the AHDS study provides strategic guidance to stakeholders involved with digital resources at various stages of their life cycle. Although its aim is to facilitate awareness about practices which may enhance the prospects for and reduce the cost of digital preservation, it is useful for anyone involved in the creation, management, and use of digital resources. Key issues which should be addressed by stakeholders in order to identify and select appropriate and cost-effective practices may be identified for each stage of the digital resource's life cycle and are summarised in the report.

The study suggests that the prospects for and the costs involved in preserving digital resources over the longer term rest heavily upon decisions taken about those resources at different stages of their life cycle. Decisions taken in the design and creation of a digital resource, and those taken when a digital resource is accessioned into a collection, are particularly influential.

The study also suggests that different (and often, differently interested) stakeholders become involved with data resources at different stages. Indeed, few organisations or individuals that become involved with the development and/or management of digital resources have influence over (or even interest in) those resources throughout their entire life cycle. Data creators, for example, have substantial control over how and why digital resources are created. Few as yet extend that interest to how those resources' are managed over the longer term. In some cases they cannot, particularly where resources are not available or allocated for this task. Organisations with a remit for long-term preservation, on the other hand, acquire digitalresources to preserve them and encourage their re-use but often have little direct influence over how they are created.

One consequence, is that decisions which affect the prospects for and the costs involved in data preservation are distributed across different (and often differently interested) stakeholders. Although stakeholders have a clear understanding of their own involvement with and interest in digital resources, they have less understanding of the involvement and interests of others. Further, they may have little or no understanding of how their own involvement influences (or is influenced by) them, or awareness of the current challenges in ensuring the long-term preservation of the cultural and intellectual heritage in digital form.

The use of standards throughout the life-cycle of the digital resource was emphasised by all respondents in the study. Their application variously ensured that data resources fulfilled at minimum cost the objectives for which they were made. They also facilitated and reduced the cost of data resources' interchange across platforms and between individuals. Standards' selection and use, however, was highly contingent upon where in its life course any individual or organisation encountered a digital resource, and on t he role that that individual or organisation played in the creation, management, or distribution and use of that resource.

The study finally suggests that funding and other agencies investing in the creation of digital resources or exercising strategic influence over the financial, business, and legal environments in which they are created can be key stakeholders. Where they recognise the long-term value of resources created under their influence, their perspective facilitates an interested overview of how those data resources are handled through the different stages of their life cycle. At the same time, their strategic influence may enable them to dictate how those resources are handled. Organisations which retain digital information to document their activities and for other purposes, may have the same perspective and the same degree of control.

**Acknowledgements**

I am extremely gratefully to Daniel Greenstein for his input and comments on an earlier draft and to the individuals who were interviewed and contributed substantially to the study report.

**References**

Beagrie, N and Greenstein, D forthcoming, Digital Collections: a strategic policy framework for creating and preserving digital resources.
<URL for consultation draft: http://ahds.ac.uk/manage/framework.htm>

European Commission 1997a, Proceedings of the DLM-Forum on electronic records, Brussels 18 to 20 December 1996, INSAR - European Archives News, Supplement II, 1997, Office for Official Publications of the European Communities Luxembourg.<URL:http://www.echo.lu/dlm/en/proc-index.html>

European Commission 1997b, Guidelines on best practices for using electronic information, INSAR - European Archives News, Supplement III, 1997, Office for Official Publications of the European Communities Luxembourg.
<URL: http://www.echo.lu/dlm/en/gdlines.html>

Garrett, J and Waters, D 1996, Preserving Digital Information. Report of the Task Force on Archiving of Digital Information commissioned by the Commission on Preservation and Access and the Research Libraries Group Inc. Commission on Preservation and Access, Washington DC.
<URL:http:www.rlg.org/ArchTF/>

Fresko, M 1996, Long Term Preservation of Electronic Materials. A JISC/British Library Workshop as part of the Electronic Libraries Programme (eLib). Organised by UKOLN 27th and 28th November 1995 at the University of Warwick, BL R&D Report 6328, The British Library Research and Innovation Centre, London.
<URL:http://www.ukoln.ac.uk/services/papers/bl/rdr6238/>

# PANDORA at the Crossroads-lssues and Future Directions.

**Jasmine Cameron, Judith Pearce**
**National Library of Australia**

## Introduction

PANDORA stands for Preserving and Accessing Networked Documentary Resources of Australia. This is the project name given to the work that has been undertaken by the National Library of Australia to develop an electronic archive of Australian publications on the Internet. Work on the project, which commenced in earnest at the beginning of 1997, has concentrated on two strands of activity. These are the development of a working 'proof-of-concept' archive and the development of a series of documents that provide a conceptual framework for a permanent electronic archive.

The National Library of Australia has envisaged from the beginning of the PANDORA Project that the knowledge gained from the project would form the basis of a much broader strategy for the creation of a National Collection of Electronic Publications. Australia has a long history of national co-operation in the areas of collecting and provision of access to information, and the National Collection of Electronic Publications will involve co-operation with other major libraries and national collecting bodies, with a view to sharing the responsibility for preserving and providing future access to Australian electronic publications.

## Background

The National Library of Australia has a statutory obligation to collect and preserve Australia's printed heritage and it regards the care of electronic publications as a logical extension of this mandate. The PANDORA Project is the first step in developing a strategy for the collection and preservation of Australia's documentary heritage as it is represented through publication on the Internet. The project's two key objectives are: to develop and test policy and procedures for the acquisition, preservation and provision of long term access to Australian information published on the Internet *and* to test the feasibility and determine the cost of establishing a National Collection of Electronic Publications.

Work towards achieving these two objectives has proceeded on two levels, one a purely theoretical level and the other a very practical level. This approach has yielded benefits to the project because both of these streams of work have informed and shaped each other. On the practical level the National Library of Australia has developed a 'proof-of-concept' archive that contains over 200 titles and is growing at the rate of approximately 10 titles a month. Policy and procedures have been developed for each step in the process including

- scanning the Internet and selecting titles for the archive
- liaising with creators for permission to archive their titles and for additional information about the frequency of update and format of their title
- cataloguing the title onto the National Bibliographic Database[1] and providing a hotlink from the PURL in the catalogue record to the entry screen in the archive
- capturing the title on a regular basis using a modified version of Harvest software and creating entry screens for each title in the archive with access to the individual issue within the archive.

[1] The National Bibliographic Database is a shared cataloguing database and a union catalogue of the holdings of Australian libraries.

Policy has also been developed for the management of commercial pay-per-view or subscription titles within the archive.

Work has also proceeded on the development of a conceptual framework for a permanent electronic archive and is described in two key documents; the PANDORA Business Process Model andthe PANDORA Logical Data Model. The Business Process Model outlines the business directions and principles on which the development of the PANDORA 'proof-of concept' archive has been based. The Logical Data Model defines the data elements in the archive, the relationship of these elements to each other and to external data. Extensive work has been undertaken on the definition of metadata needed to describe and manage titles within the archive and this work forms part of the Logical Data Model. These two documents are available on the PANDORA Home Page at 'http//www.nla.gov.au/pandora'. Work is currently progressing on a much broader document that will be completed in August. This document, which may be released as a Request For Information, describes the National Library of Australia's requirements for a digital object management system which will meet the needs not only of PANDORA but the Library's other digital collections. This work is being carried out as part of the Library's Digital Services Project.

**PANDORA Business Principles**

Several key business principles have been incorporated into the design and management of the PANDORA 'proof-of-concept archive' It is important to stress that most of these business principles are based on practical decisions and as the project has evolved so has the Library's thinking begun to broaden in relation to many of these principles.

***Selectivity***

From the beginning of the project, the principle of selectivity has formed the basis of the PANDORA selection guidelines. The National Library of Australia is also selective to an extent in its acquisition under legal deposit of Australian print publications and relies on the State libraries to collect material at a local level. For example, the Library does not collect publications such as school magazines and local club newsletters. The Library's policy in relation to the collection of electronic publications is intended to be the same as that for print, and although it is currently more restricted than our print collecting it is proposed that in the future the Library's collecting of on-line publications broaden to match the print policy.

Unlike our colleagues at the Royal Swedish Library, no attempt has been made to capture the entire Australian domain. Selection guidelines determine a range of publications to be archived, from scholarly titles to those representing popular culture and the use of the Internet by Australians in general. This approach has its merits including the ability to exercise a degree of control over the quality of what you have archived and to provide access to what you have archived. On the other hand to regularly scan the Internet to select individual titles for archiving is resource intensive and valuable information may be missed and therefore lost. The National Library of Australia believes there is merit in both approaches and while continuing to use a selective model, the Library may also experiment in the future with 'snapshots' of defined segments of the Australian domain.

A decision was also made very early on that titles with print equivalents would not be selected for the archive. This decision was made because it reduced the large amount of information that would be eligible for selection to a manageable amount and it was reasoned that Australian titles in print were already being collected and preserved as part of the Library's legal deposit role. However, this decision is also under review because it is readily acknowledged that electronic titles with print

equivalents can often vary in nature and content from their print counterparts. Future ease of access to electronic versus printed information is also an issue.

***Access***

Following on from the Library's selective approach to archiving titles in PANDORA it was considered important to let other Australian libraries, and indeed libraries internationally, know which titles the National Library of Australia has undertaken responsibility for archiving. This has been done by creating a catalogue record on the National Bibliographic Database for each title in the archive. Catalogue records created on the National Bibliographic Database for titles in the PANDORA archive are also downloaded into the Library's Online Public Access Catalogue. This is considered the best mechanism currently available for providing integrated access to information in any format held in the Library's collection.

The issue of resource discovery relating to the actual content in the archive below the title level has not been addressed in the development of the 'proof-of-concept' archive. However, the facility for capturing and/or generating Dublin Core compliant metadata which is indexing the content of publications captured for the archive, and posting this metadata to a designated metadata repository, has been included in the Request For Information referred to earlier. It is recognised that in the future the PANDORA archive will contain a large number of titles that exist nowhere else and that access to the content of the archive will be an important issue. The Library's Australian Public Affairs Information Service which indexes selected Australian printed journals in the Library's collection, is indexing selected Australian electronic journals. This indexing service could also be expanded in the future to routinely index Australian on-line publications.

***Management of commercial publications***

The Commonwealth of Australia Copyright Act, which covers legal deposit, does not currently include the legal deposit of electronic publications, either in physical format or on-line. However, the Copyright Act is under review and it is anticipated that legal deposit will be extended to cover electronic publications for preservation purposes. The Library is lobbying to have the concept of legal deposit extended to electronic publications and has made formal submissions to this effect to the Copyright Law Review Committee. In the event that electronic publications are covered by legal deposit, the Library will not, at this stage, be negotiating to provide access for remote users to current issues of online commercial titles through the PANDORA archive. PANDORA is first and foremost an archive and it is expected that users will visit the publisher's site for all currently available material.

The difficulty for the PANDORA project at the moment is that the use of the Internet for publishing in the Australian domain is still very much restricted to gratis and nonprofit publications so that the Library's thinking on the issues surrounding the management of commercial on-line titles has not been tested. In view of the fact that Internet publications in the Australian domain are not yet subject to legal deposit, the PANDORA project has developed a 'Voluntary Deposit Deed' for on-line publications. This deed closely mirrors the Voluntary Deposit Deed used by the Library when seeking physical format electronic publications. Publishers will be asked to nominate from a standard list of timeframes the period for which they wish their publication to be suppressed from the public domain. Publishers will also be asked to agree to allow gratis access to current information to on-site users. On-site access to electronic publications is seen as a parallel to the on-site use of printed material received on legal deposit.

To date the National Library of Australia has negotiated successfully with only one Australian Internet publisher for the voluntary deposit of their commercial literary and reference 'monographs' in the PANDORA archive. Although a voluntary deposit deed has not yet been signed, the publisher has agreed to allow the Library to provide on-site access to their titles in the archive. A timeframe for future access to the titles by remote users has not yet been agreed. The Library does not see a role for itself as a middle-man for commercial Internet publishers by levying, on behalf of these publishers, a fee-for-use of commercial publications held in the PANDORA archive. One of the first business principles established was that the archive is a secondary resource, for use when issues are no longer available on the publisher's site. The Library is approaching its management of commercial titles in the archive on this basis.

The Library's thinking on this issue may well have to be modified in the future as major Australian publishers move into the Internet publishing market. We are looking closely at arrangements such as those made by the Royal Dutch Library with major commercial publishers like Kluwer and Elsevier. The Royal Dutch Library pays a license fee to these publishers in return for being able to provide access to their on-line publications to both the Royal Dutch Library's registered on-site and remote users. The Royal Dutch Library is taking responsibility for archiving these electronic journals in the same manner as the PANDORA archive.

## Legal Deposit and Copyright

### *Legal deposit*

As mentioned above, electronic publications are not yet covered by legal deposit although the Library anticipates that this will be included as part of the current revision of the Copyright Act. The broader issue of how to best filter and select titles on legal deposit for the archive is yet to be resolved. One method is to request that Australian Internet publishers register their titles with the Library so that these titles can be scanned and selections made from the registry. This registry could then form a valuable resource, doubling as a national bibliographic listing of Australia on-line titles. The value of maintaining such a listing, and the value in monitoring what could be a large number of publications registered for legal deposit, has yet to be fully debated in the Library. What is certain, however, is that coverage of electronic titles by legal deposit will require the Library to establish a new set of relationships with Australian publishers on the Internet and to review some of its present PANDORA principles and procedures.

### *Copyright*

It is somewhat ironic that the electronic age with all its vast potential has brought with it a set of restrictions that often make the provision of access more limited than that for printed material. Copyright is a complex issue in the online environment, particularly where multi-media web sites are concerned. There may often be many more creators involved with an on-line publication than is the case with print. There may be different authors of text, images, software and so on. In the case of electronic journals it is not unusual for the copyright to be held by authors of individual articles.

The PANDORA project has approached the issue of copyright by including a general copyright warning, plus a link to the publisher's own copyright statement, on the entry screen for each title in the archive. Under the current copyright reform agenda, the Australian government has announced recently that a new, broad-based, technology-neutral right of communication to the general public will be introduced. This right will apply to information made available on the Internet and other on-

line services. This right will be subject to exceptions for fair dealing, libraries and educational institutions.

## Standards

The National Library of Australia has a leadership role within the Australian library community in the development of standards. This involves the Library in a wide range of activities including representation on key Australian standards bodies and international working groups such as the Z39.50 Implementors' Group and the Dublin Core group. The PANDORA project has a particular interest in the development of metadata standards for resource discovery, permanent naming conventions, standards for preservation metadata and Internet publishing conventions.

Although not directly involved in standards work, through its contact with Internet publishers PANDORA plays a role in creating awareness of the benefits of generating metadata for resource discovery and of the role of permanent naming for stability of links to documents on the Internet. The ability in the future to provide adequate access to documents in the PANDORA archive depends to a large extent on the development of standards and publishing conventions in many areas. The Library is currently a partner in a project known as Metaweb which is developing metadata element sets, user tools and indexing services to promote the use of metadata. The project has also looked at the concept of a national metadata repository. The Library has also set up a PURL Resolver Service based on OCLC software, although it is recognised that PURLs are only an interim solution for permanent naming. PANDORA has encouraged publishers to assign PURLs to their documents and by way of example routinely assigns PURLS to titles in the archive.

Within the PANDORA project a large amount of work has been done on identifying the key data elements of the PANDORA archival management system. It is anticipated that this work will be compatible with future preservation metadata standards. PANDORA is also interested in, but has not been directly involved with, the development of Internet publishing conventions. A wide range of actions, for example non-standard file structures and file names, undertaken by Internet publishers affects the ability of the PANDORA project to satisfactorily capture and maintain the look and feel of titles in the archive.

## Technical requirements

The PANDORA project has reached the stage where the 'proof-of-concept' archive needs to be underpinned by a robust software and hardware platform suitable for the establishment of a permanent electronic archive. PANDORA's technical infrastructure requirements form part of the Library's new Digital Services Project referred to earlier. It is hoped that solutions will emerge following the completion of a Request For Information later in the year. The two key elements which will assist the progress of the PANDORA project are the identification of a suitable 'gathering' or document capture mechanism, and the identification of a suitable digital object management system which will facilitate version control, rights management and authentication within the archive.

### *Gathering*

At the commencement of the PANDORA project it was decided that it would be better to pro-actively capture documents for preservation in the archive than to rely on publishers to 'push' information to the archive. This decision was partly influenced by the fact that

the majority of Internet publishers in Australia were new to publishing and did not have established relationships with the Library. It seemed a lot easier to say "we will come and get it" rather than to rely on publishers to send the information to us, and particularly in situations where a regular capture schedule is necessary. The PANDORA project currently uses a modified version of Harvest software to capture publications for the archive. However, Harvest is essentially designed for indexing and is not suitable for use in the medium to long-term as a capture mechanism.

Finding an alternative to Harvest is a top priority for the PANDORA project. The difficulty in this endeavour is that few other institutions require software for this exact purpose and it may be that we will have to settle for a modified version of a more sophisticated indexing software. Another option would be to develop, preferably in collaboration with another national library (or libraries), software specifically for this purpose. Towards the end of 1997 the PANDORA project began to investigate alternative capture strategies. This was spurred on by the emergence of publications structured as databases which operate on creating data for the user on the fly. It had become obvious that 'pull' technology would not cope with this new publishing format and as a result experimentation with publishers 'pushing' their publications to the PANDORA archive has commenced. While PANDORA still intends to use gathering software to capture many of its publications, a mixture of both push and pull technology will be used in the future.

***Management of documents in the archive***

The PANDORA 'proof-of-concept' archive has grown to the stage where it is critical that a digital object management system be implemented to facilitate the wide range of functions associated with the selection and management of the titles in the archive including the negotiation status of the publication, procedures for the capture of copies for the archive including an automated capturing schedule, updating the archive, restrictions on access, version control, retaining the 'look and feel' of a publication and authentication.

The Library is relying on developments in the commercial sector to provide solutions to some of the broader business issues facing the PANDORA project. For example, authentication is an issue for most people doing business on the Internet and solutions are beginning to emerge for general use such as encryption, time stamping, watermarking and digital signatures. For PANDORA, the aim is to provide a record of Australia's documentary heritage and that must be an accurate record. The documents that are archived must remain unchanged and true to the original. It is anticipated that the current work on encryption and other authentication methods which will facilitate the verification of the content of a publication will provide an answer to the issue of authentication that is suitable for PANDORA.

***Look and feel***

One of the important PANDORA business principles is to retain, as far as possible, the look and feel of a document in the archive. Preserving the integrity of a document selected for archiving is considered important for the accuracy and completeness of Australia's future historical record. However, it is recognised that in some cases this may simply be impossible. The technical issues surrounding the retention of the look and feel of a publication are numerous. It requires the PANDORA archive to maintain a software repository and to capture copies of the software supporting the viewing of publications selected for the archive, if not already in the software repository. In the future, it means that this supporting software will need to be migrated to a new format. And all this activity will have to be tracked and managed within a digital object management system.

## The Australian National Collection of Electronic Publications

From the beginning of the PANDORA project the National Library of Australia has envisaged that the task of managing and preserving electronic publications would be taken up nationally once the Library has completed its developmental work and a viable set of policies and procedures are in place. The Library is looking to establish the National Collection of Electronic Publications in co-operation with other collecting institutions such as the other Australian deposit libraries, all of which have responsibility for collecting and preserving Australia's electronic documentary heritage, whether originally published on-line, or in digitised or physical format. It is anticipated that in the long term the National Collection of Electronic Publications will operate within a distributed framework for the selection, description, acquisition and provision of long term access to all information in any electronic format.

In order to develop the Library's thinking in this area and to test the procedures and policies developed to date within the PANDORA framework, two of the Australian deposit libraries, the State Library of Victoria and the State Library of New South Wales, have agreed to participate in the PANDORA project. The State Library of Victoria will take responsibility for all the procedures from selecting and cataloguing, to liaising with the publisher for permission to archive, and creation of entry screens in the archive. The National Library of Australia will continue to capture and store the publications within the PANDORA archive until such time as the State Library of Victoria has the capacity to undertake its own capture and storage of publications. The critical issue here for the Library is the identification, or development, of a robust archive management system which will minimize as far as possible the work entailed in supporting a National Collection of Electronic Publications. The current thinking on the national co-operative model envisages that the technical framework will consist of distributed digital object servers separately maintained by each collecting body, with unified access provided by the catalogue entries in the National Bibliographic Database. This will be the primary strategy for providing access to the information in the PANDORA archive at the whole item level.

## Resource and costing model

Producing a resource and costing model has proven to be a difficult exercise and has not been completed satisfactorily to date. This is largely due to the fact that the PANDORA project is still very much in a research and development phase, with policies and procedures evolving and maturing. We know how we would like things to work, in the sense that we would like to 'automate' as much of the work being done manually as we can. However, the extent to which this will possible over the next few years depends on advances in software development in areas like the capture of digital objects, and digital object management systems. Procedures currently being followed in the PANDORA project are interim until something more sophisticated is available.

As procedures are developed for the PANDORA project they are taken up and implemented within the Library's 'Electronic Unit', which currently consists of four librarians who do all the tasks from regular scanning and selection of titles from the Internet, liaising with the publishers, cataloguing, devising a capture schedule, version comparisons and creation of title entry screens for each new title selected for the archive. It is currently estimated that it takes one librarian a full working day of seven and a half hours to undertake all of the procedures linked to selecting and capturing a title for the archive, including all the steps from regularly scanning the Internet to the creation of an individual listing of issues attached to each title in the archive. On-going maintenance per title is obviously less time consuming. In addition to these four staff, there is a unit manager who assists

with the research and development of PANDORA and a full-time Information Technology staff member who works with Harvest and solves technical problems associated with the capture and storage of publications in PANDORA.

Those closely involved with the PANDORA project feel that they have virtually replicated a whole new acquisition and control system, equivalent in some ways to the work originally required to devise and implement the Library's legal deposit acquisition program. The Library has yet to clearly quantify the likely on-going cost of routinely selecting, acquiring and providing access to on-line Australian publications. The likely cost of a robust software and hardware platform to underpin the archive is still unknown, as is the amount of in-house or co-operative software development the Library may have to undertake on this project.

**International collaboration**

The National Library of Australia sees great benefit in collaborating with other national libraries or institutions working in the field of digital management or archiving. The Library believes that other national libraries are natural partners and is actively seeking to share information and undertake co-operative research and development with other national libraries. International collaboration is in fact the next logical step in the development of the PANDORA project. Although each national library may approach the actual management of electronic publications in a different way there are a range of key issues where the simple exchange of ideas and information, or the development of agreed principles, would provide mutual benefits for all national libraries.

In order to make progress towards collaboration, the National Library of Australia intends to write an issues paper for discussion amongst national libraries with an active interest in electronic archiving. The paper will attempt to describe the status of work currently being undertaken by national libraries in this area, and to distill those issues where some sort of common approach or agreed principles may help influence key developments such as permanent naming, or even the development of software and hardware platforms.

# EVA

# The Acquisition and Archiving of Electronic Network Publications In Finland

Kirsti Lounamaa
CSC - Center for Scientific Computing
FUNET network services
Kirsti.Lounamaa@csc.fi

Inkeri Salonharju
Helsinki University Library - The National Library of Finland
Inkeri.Salonharju@helsinki.fi

## Abstract

This paper presents the current state of the effort to capture and preserve electronic documents published in the Finnish Internet. We describe the policy used to select the material to be included in the collection, the capturing process, the storage technology and accessing method. Some advanced tools to extend the access to the documents as well as the results and some statistics from the first collection are also presented. Electronic documents are not yet as a part of the Legal Deposit in Finland, but the revision of the Act is going on and the aim is to preserve also electronic publications for the coming generations.

## Background

Helsinki University Library - The National Library - is co-ordinating the project EVA, which is a joint activity by libraries, publishers and expert organisations, being a part of the Information Society Strategy Program by the Finnish Ministry of Education. The main aim of the project is to test methods of capturing, registration, preserving and providing access to the on-line documents published by the established publishers or freely in the Finnish Internet. In this presentation we mainly concentrate on the latter type.

EVA is a cross road of several other development projects on the Nordic basis and serves as a test bed for new tools. The Dublin Core metadata template and converter, the URN generator and the harvesting and indexing application NWA (Nordic Web Archive) are tested in EVA. The design and specifications of NWA are done in co-operation with the Royal Library of Sweden and the CSC/FUNET - Center for Scientific Computing/Finnish University Network - is responsible for the implementation of NWA in Finland.

## The advanced tools tested in EVA

*Metadata template.* The Dublin Core metadata template was originally created by Lund university library NetLab unit with the Nordic Metadata project. The template builds all the required HTML syntax automatically and allows the user to concentrate on creation of the content. The Perl script that creates the template is available for free and has become very frequently used world wide. There is already a significant number of documents in the Web which contain their own description in the Dublin Core. We realised early on, that we needed to offer an easy to use and fast template to encourage writers to produce metadata themselves. The template is an electronic form which, after filled in, immediately returns data parsed as a Dublin Core record.
(http://linnea.helsinki.fi/cgi-bin/dc.pl)

*The URN generator* which can build Uniform Resource Names based on National Bibliography Numbers, NBNs. The syntax of the produced URNs is authorised by the IETF URN WG and the numbering schemes are assigned by the Finnish and Swedish National Libraries. URN identifiers are persistent and unique: the URN given to a document will never change, if the intellectual content remains the same. The browsers are at the moment not capable to find the document when a user puts the URN into the Location-box. In the future the Nordic Web Index system's national full-

text and metadata databases will offer the full and correct searchability of all URNs used in WWW documents published in Nordic countries.
(http://linnea.helsinki.fi/cgi-bin/urn.pl)

*The metadata harvesting and indexing application*. This is in practice an enhanced version of the Nordic Web Index which can extract metadata in Dublin Core format and other formats from HTML documents and make this information searchable via metadata databases.
(http://nwi2.funet.fi/)

*The Dublin Core to MARC converter*, which can extract Dublin Core data from a document and convert it into a MARC record. From libraries point of view it is important to be able to utilise the Dublin Core records in the maintenance of the National Bibliography database and in the library OPACs.
(http://www.bibsys.no/meta/d2m/)

## On-line documents with limited access

Not all the electronic documents can be acquired using automatic methods. The library must use other methods for collecting restricted documents from commercial publishers. In the future deposit of on-line documents is defined on national level either in the Act of Legal Deposit or in voluntary agreements between the National Library and publishers. According to the proposal for the new Finnish deposit law publishers will be responsible for sending the on-line deposit documents to the library.

The National Library is experimenting the transfer of the documents with different methods. At present, document delivery using ftp-protocol seems to be the most effective way to accomplish this. Situation may change rapidly when new Internet-tools come to the market. To prevent document spoofing during transmission, digital signatures are used. In addition to this purpose, a digital signature can be used for two other purposes, firstly for authentication of the sender, and secondly for authentication of the content.

In the future the National Library as a Legal Deposit will maintain a list of approved file formats. The organisation responsible for depositing the document (generally the publisher) should also be responsible for converting the document to an acceptable format. If it is not possible to convert the document to a suitable format with reasonable effort, it is not necessary to deposit it.

It is important for National Libraries to retain the original look and feel of the documents. Therefore the list of approved formats should be quite exhaustive and updated frequently. Long-time preservation is in these cases secured by in-house conversion of the document to "better" format, or by emulation of the original usage environment.

## Selection criteria for harvesting freely available documents

The general design is that we don't try to make any specific selection but the collection will include all the freely available, published, static HTML-documents with their inline material like pictures, video and audio clips, applets etc. In our approach, 'published and freely' accessible means that the document is accessible by standard HTTP protocol, it is referenced in some other document and there is no fee or password required. Only static documents are captured leaving out programs, database searches and so on. Since we are interested only in documents published in Finland, we have limited the collection, so that we have included only Web sites with network address ending in '.fi'. We are well aware that there are a large number of Web servers which are located in Finland, but whose address ends in '.com', '.net' or something else. These are becoming more and more popular and we have to find a method to recognise them. In the Kulturarw3 project in Sweden this has been solved by asking InterNIC for the list of DNS entries whose owners have given an address in Sweden. This is probably a satisfactory solution for us too.

The Web consists of documents which are linked together with hyperlinks (addresses) embedded in the documents. The documents are identified by their location, so called URL (Uniform Resource Locator). The problem is that the only thing we know about a document is its URL. There is no relation between the URL and the contents of the document, which means that one document may have several locations over time and also that in one location there may be different documents at different times. We see that this problem can't be solved with the current technology before there is a way to identify the content of the document, not the document itself. A URN (Uniform Resource Name) is a unique identifier given by some authorities and it can be embedded in the document, so that it follows the document all its lifetime. Before the usage of URN is more popular, we have to capture the documents by URLs. The limitation is that there is no concept of document version: a document in a location in one snapshot has possibly nothing to do with a document in the same location in the next snapshot.

## Capturing

Regardless of the limitations, we try to create a series of snapshots of the contents of the Finnish Web that exists over time.

Technically the system is quite simple. We have written a harvester robot program whose task is to fetch Web pages. Once a page is captured, our software analyses its content for inline material and cross-references, i.e. hyperlinks. In our model all the inline material, for example pictures and video and audio clips, is considered as essential part of the document's content and has to be fetched regardless of its location. This means that this material is captured also outside of Finland. Where as cross-references are not part of the current document and they are only fetched if they reside on the Finnish Web.

The system consists of four parts: capturing, analysing, indexing and archiving. Every day a program starts which analyses all the documents that were captured the day before. As an output a list of URLs is generated. This list is sorted according to the capturing policy. The policy ensures that Web sites are checked often enough for new and modified documents. At the same it must ensure that the Web server's normal activity is not interfered and the network traffic is kept in minimum. The capturing process then starts reading and processing the list.

The capturing process, i.e. harvester robot, fetches the document and stores it in the local disk to an ordinary unix file system. To maintain the integrity of the snapshot, only the documents with successful HTTP return code (i.e. 200) are stored. The documents are stored together with HTTP headers *as is*, no modifications to their contents are made.

The documents are given a document-id. At the moment we are using the md5-checksum of the document as an identifier because it's guaranteed to be unique. In the implementation the document-id is also the name of a file in the local file system. This is also a way to prevent storing duplicates: if the document already exists, it is not stored again.

## Storage

The captured documents are packed and compressed daily using TAR- and ZIP-utilities. This compressed TAR-file is then sent to the archive server using FTP. The file contains 10 000 - 50 000 documents and its size is usually 0.5 - 1 gigabytes. The reason why we have to pack the documents in big files is because at the moment the long-term, low-cost storage is based on tape technology that can't handle small files efficiently.

A concept of hierarchical storage is proven useful in several applications which store large amounts of data for long period of time. In hierarchical storage, there are several layers of storage, usually fast, expensive, small-capacity disk storage at the top and slow, inexpensive, large-capacity tape-drives at the bottom. There may be several layers in between. A software, HSM - hierarchical storage manager - migrates unused data downwards. When the data is accessed the software migrates the data upwards to the fastest device. The system is transparent to the user, the software decides the correct placement of the resource on basis of its usage.

Our HSM software is UniTree running on a HP machine and the tape technology is StorageTek Timberline cartridge robot. The archive technology was selected simply because we already had it installed and it's been used for many years. No investments were needed. The tape technology is very conventional and we are quite confident that the data will be preserved for 15 years, what is guaranteed by the manufacturer. This gives us enough time to decide what to do with the archive and to which storage technology it will be migrated next. For safety the UniTree software writes two copies of files to separate tape pools of which the other is to be kept in the vault as a backup copy.

In this first version of the model we decided to store the document and its parts, i.e. inline material separately. This is probably not a good idea. When the document will be presented, it is difficult to maintain the integrity if all its parts are stored in different times in different files. For example, how to guarantee that the pictures are the right ones? A new approach is to change the capturing process so that all the inline material is captured immediately after the document. We also have to select an appropriate file format to store the document and its inline material in an aggregate. The idea is that everything that is needed to show the document in the browser as it was at the moment of capturing, is collected together and stored sequentially. A similar approach can be used as in RFC 2112, which describes how to encapsulate multimedia documents to be transferred in a single e-mail message. This is anyway something we a working on in the project in the near future. The drawback of this design is that it will make the archive much larger than today because for example all the pictures will be stored in several copies. But since the data is stored in the low-cost, high-capacity tapes it is not a problem.

## Statistics

We finished the first snapshot at the end of March 1998. Now we are taking the second snapshot. The first snapshot contains about 1.8 million documents from about 7500 Web sites in domain '.fi' . The majority of documents is text,

86 % of documents are of types html and plain text. About 10 % of documents are images, whose main types are giff, jpeg, tiff and x-bitmap. About 4 % of documents are applications, whose main types are tar, zip, octet-stream and pdf . The rest (less than 1 %) include audio and video and about 180 other types of document of which most are unknown to us. Main multimedia types are realaudio, wav, midi, mpeg, quicktime and ms-video.

About 50 000 documents are collected every day. If there are duplicates of documents already in the archive, they are dismissed. (About 12 % of documents are duplicates.)

Even if there are a lot of different document types found, we will be able to show most of them in the future because the most common file types represent over 90 % of all types. Even if we had to throw away the rest, the snapshot is representative.

Other thing that justifies taking of snapshots is that the documents are not modified too often. The average time between modifications is 200-300 days. Of course this reflects our selection policy, only static documents are collected. But it can also be seen that it is normal practice that documents stay untouched for long period of time after they are published.

The size of documents on the other hand is growing rapidly. On March 1998 we calculated that the average size of a Web document is 15 kilobytes, now it is over 20 kilobytes. This obviously has to do with the increasing usage of multimedia features. This might cause problem in the future. At the moment the estimated growth of the archive is 0.5 terabyte per year. If it grows much faster we might have to change the selection policy towards more strict rules.

## Access

Since FUNET runs a Finnish service point of NWI - Nordic Web Index, we have combined the archiving and indexing processes. All the captured documents are indexed using the NWI-profile. The NWI database is accessible by Z39.50 (host: nwi.funet.fi, port:2100) and it contains records of all the online Web-documents. The NWI-profile contains:

Date the resource was checked
Date the resource was last modified
Content type
Content size
MD5 checksum
Availability: URL to the on-line document
Title
Headers
Sample text (about 20 % of the full text)
Cross-references: linkage and title

We have also started to collect a database FinMeta, which contains all the Dublin Core descriptions that are found in the Finnish Web. There are currently about 1000 records in the database.

The methods for providing access to the archived material will be studied, specified and tested later this year within the NEDLIB project.

## References

EVA - Electronic Virtual Archive: http://linnea.helsinki.fi/eva/
The Kulturarw3 Heritage Project: http://kulturarw3.kb.se/html/projectdescription.html
NEDLIB- Networked European Deposit Library: http://www.konbib.nl/nedlib/
NWI - Nordic Web Index: http://nwi.funet.fi/

# Digital Rosetta Stone: A Conceptual Model for Maintaining Long-term Access to Digital Documents

Alan R. Heminger, Ph.D.
*Air Force Institute of Technology*
**aheminge@afit.af.mil**

Steven B. Robertson, Captain
*United States Air Force*
**steve.robertson@worldnet.att.net**

*The views expressed in this article are those of the authors and do not reflect the official policy or position of the United States Air Force, Department of Defense, or the US Government*

## Abstract

*Due to the rapid evolution of technology, future digital systems may not be able to read and/or interpret the digital recordings made by older systems, even if those recordings are still in good condition. This paper addresses the problem of maintaining long-term access to digital documents and provides a methodology for overcoming access difficulties due to technological obsolescence. The result of this effort led to the creation of a model, which we call the Digital Rosetta Stone that provides a methodology for maintaining long-term access to digital documents. The underlying principle of the model is that knowledge preserved about different storage devices and file formats can be used to recover data from obsolete media and to reconstruct the digital documents. The Digital Rosetta Stone model describes three processes that are necessary for maintaining long-term access to digital documents in their native formats--knowledge preservation, data recovery, and document reconstruction.*

## 1. Background

Any large organization has the need to retain and, on occasion, refer to various stored documents. Until recently, documents were generally paper or microfilm based. However, modern data storage methods have evolved to include digital storage as well. Due to the rapid evolution of storage technologies, future digital systems may not be able to read and/or interpret the digital recordings made by older systems, even if those recordings are still in good condition [24].

Within many organizations today, digital documents that are official records must be categorized and managed in accordance with approved records schedules. This means that these records must be retained and accessible throughout their life cycle in accordance with the same laws and standards that govern paper records. In the case of government documents for instance, the law dictates that an official Government record must be classified into one of 26 retention periods set forth by the Archivist of the United States. These retention periods range from 30 days to Permanent storage and include time periods of 30 years, 50 years, and 75 years.

Digital documents that require long retention periods face accessibility problems due to the technology obsolescence of hardware and software. As time goes on, and more documents are stored in digital format. The National Research Council [20] best describes this problem in the following statement:

> The fact that most electronic hardware is expected to function for no more than 10 to 20 years raises very serious problems for long-term (more than 20 years) archival preservation. Even if the operating systems and documentation problems somehow are dealt with, what is the archivist to do when the machine manufacturer declares the hardware obsolete or simply goes out of business? Will there be an IBM or Sony in the year 2200? If they still exist, will they maintain a 1980-1990 vintage machine? Moreover, it must be realized that no archival organization can hope realistically to maintain such hardware itself. Integrated circuits, thin film heads, and laser diodes cannot be repaired today, nor can they be readily fabricated, except in multimillion-dollar factories.

As digital technology continues to evolve at a rapid pace, superseded technologies are quickly discarded and new technologies are embraced in the hopes of gaining improved efficiency, effectiveness, or a competitive advantage. It is crucial that, in the haste to adopt newer and generally better technologies, we don't lose the ability to access historical digital documents.

The purpose of this paper is to respond to this call for action by developing a model with which we can assure ongoing access to the ever-increasing repository of digitally stored information. We will concentrate on maintaining access to digital documents in their native formats without converting them to emerging digital format standards. The resource and financial burdens of converting an entire archival collection every 10 to 20 years is "likely to be out of the question except for relatively small collections that have great historical importance, sustain heavy use, or require rapid access [20]."

This study is largely exploratory and prescriptive in nature because of the relative newness of this subject area and the lack of previous studies. The use of secondary data analysis techniques will be used to develop and explain a model that can be used to recover and reproduce digital documents from their native file formats.

As information systems are upgraded the ability to

problem due to the technological obsolescence of the hardware and software systems needed to access them. This is because most digital documents contain information that is only meaningful to the software and hardware systems that were used to create, edit, and access them. Therefore, because of these non-standard digital document formats, organizations that archive digital documents must develop a method that will allow them to maintain continual access to digital documents in their native formats.

## 2. Current Strategies for Preserving Digital Documents

A number of strategies for preserving digital documents have been discussed in the literature. Charles Dollar [9] suggested that customers should demand that vendors provide cost-effective migration paths to advancing hardware and software systems. Many vendors do provide a limited form of this capability in that an advanced version of their hardware or software system will provide the ability to migrate a customer's operations from the superseded system to the advanced. However, this type of conversion is generally limited to the previous generation of the hardware or software system and therefore, it is a short-term fix that must be repeated with each successive system upgrade. Additionally, the translation of a digital document into successive short-term standards over its life cycle may result in the loss of the document's original content. Without the original document and the original software to accurately interpret the document, then the format and content of the document may be compromised and the original meaning lost.

Dollar and Rothenberg also suggested promoting a "trend toward non-proprietary standardized open systems environments, which are designed to overcome compatibility between computer systems and applications and are reflected in international standards" [9]. While these open system standards would make digital documents accessible through any software system that conform to the standards, there is still the problem that even the open system standards will change as information systems technologies continue to advance. Thus, over time, as hardware and software systems continue to evolve, it will still be necessary to either migrate digital documents to an updated standardized format or to provide some other method to maintain continual access to these documents.

Rothenberg [27] suggested a means of maintaining long-term access to the information contained within digital documents by extending the life of the original computer hardware and software systems on which the digital documents were created. These life-cycle extensions involve the operation and maintenance of antiquated hardware systems and the archiving of the software needed to access digital documents in their native formats.

While maintaining a depository of antiquated hardware might be achievable in principle, it is also plagued with problems. The main drawbacks being the cost of operating multiple information systems and the difficulty in acquiring antiquated hardware system components [20]. These problems make it unrealistic to expect that any organization could effectively and efficiently maintain multiple, aging information technologies in order to maintain access to superseded digital documents.

To overcome the problems associated with maintaining aging hardware, Rothenberg [27] suggested the creation and use of system emulators that can imitate the behavior of antiquated hardware systems. This method would allow the operation of superseded software on advanced systems as a way to view digital documents in their native formats. However, in order to emulate an antiquated information system this method requires exhaustive specifications on the original system's hardware. Therefore, this method may require extensive participation by hardware manufacturers. Many manufacturers may be reluctant to supply all of the specifications to software developers because some of the technology may still be in use in advanced systems they have developed.

## 3. Which Strategy to Use?

None of the strategies discussed above is entirely satisfactory by itself. Therefore, as information systems and their operating environments continue to evolve it may be necessary to use some combination of one or more of these strategies in order to maintain access to digital documents in superseded formats. The strategies chosen will need to evolve from organizational requirements and conform to the limits of its financial, physical, and human resources [26].

Because a long-term strategic plan may call for a conglomerate of the methods mentioned here, it is conceivable that no existing organization can afford the financial, physical, and human resources necessary to carry out such a tremendous task. Therefore, it may be necessary to establish organizations or processing centers that specialize in maintaining long-term access to digital documents [20]. To recapture the information in the myriad digital documents that will be an increasingly large proportion of our information storage may require something comparable to the Rosetta Stone that opened up the writings of ancient Egypt to scholars of today.

## 4. The Rosetta Stone

At some point during the fourth century, all knowledge of ancient Egyptian scripts was lost, leaving no method available to decipher the language of hieroglyphics which had been richly preserved on ancient Egyptian monuments, stone tablets, and sheets of papyrus. Fortunately, while on an expedition to Egypt in 1799, Napoleon's army discovered an artifact which has become known as the Rosetta Stone. This stone contained the inscription of a decree issued in 196 BC by Ptolemy V Epiphanes. The decree was repeated three times in two languages, Greek and Egyptian, with the Egyptian version appearing twice, once in hieroglyphics and once in demotic, a cursive form of the hieroglyphic script. Fortunately, there is an abundance

of information on ancient Greek dialects and therefore, the stone's Greek version of the decree contained the key to decipher the meaning of the ancient Egyptian texts. Today, because of the Rosetta Stone, we can interpret many ancient texts and inscriptions of Egyptian hieroglyphic and demotic scripts found on sheets of papyrus and monuments throughout Egypt.

## 5. A Digital Rosetta Stone (DRS)

We draw on the strategies discussed above and add others to create a model for maintaining long-term access to digital documents. We call this model the Digital Rosetta Stone (DRS) because it offers a way for those in the future to be able to gain access to the information stored in the digital documents that we have stored, and will continue to store, in increasing numbers. The DRS will contain multiple levels of knowledge about specifications and processes by which information is stored on various types of storage media. It will also contain archives of knowledge about how to meaningfully interpret that information so that the original meaning can be recovered.

Rothenberg [27] stated that if the behavior of an information system could be sufficiently described, then future generations could re-create that behavior and reproduce digital documents without the need for the original systems. However, he also said that currently, information science cannot sufficiently describe this type of behavior in a way that will allow this strategy to succeed. One way to describe and preserve the behavior of information systems for our posterity is to create a DRS that can be used to reconstruct digital documents.

The processes and metadata maintained by the DRS will catalogue the many different aspects of digital technologies. After all, "in the digital world, preservation must be concerned with entire technology systems, not one or another component, such as a film or a storage disk" [5]. In digital equipment each component is dependent upon other components of the digital systems in order to perform a specific task. The process of viewing a file created by a word processor can demonstrate a simplified example of this interdependence. The file must be interpreted by the application program which is dependent upon the operating system which is further dependent upon the system's hardware. Each layer of digital technology involved in this process contributes some form of information necessary to view the digital document.

## 6. DRS Components

Unfortunately, creating a DRS is not as simple as the creation of the original Rosetta Stone that held the key to Egyptian hieroglyphics. Instead, a DRS is composed of three major processes that are necessary to preserve and access our digital history-- knowledge preservation, data recovery, and document reconstruction. The knowledge preservation process supports the data recovery and document reconstruction processes.

### 6.1 Knowledge preservation

Knowledge preservation is the process of gathering and preserving the vast amounts of knowledge needed to recover digital data from a superseded media and to reconstruct digital documents from their original formats. In a DRS, the preservation of knowledge of media storage techniques and file formats will be maintained in a metaknowledge archive. Metaknowledge is the knowledge or awareness of facts, heuristics, and rules, and the context in which they are used and manipulated. The creation of standardized data dictionaries will be the tools used to store the metaknowledge necessary to aid document recovery personnel. The data dictionaries will contain the names and descriptions of the data items and processes necessary to recover a digital document [14]. The metaknowledge archive (MKA) is the foundation upon which the DRS is dependent and it must extensively preserve the knowledge in two key areas--*media storage techniques and file formats.*

The knowledge of media storage techniques is a collection of the way data are defined and stored on specific media. While it is expected that some data will be migrated to new storage devices for archival purposes, it is likely that some data will not be migrated. Therefore, it is necessary to maintain a record of the methods in which bit patterns are used to represent data on storage devices. The knowledge of the location and meaning of these bit patterns will be necessary to recover data if equipment to access a storage medium is not available or no longer exists. This is not to say that all specifications for storage devices must be accurately preserved so engineers can manufacture them in the future. Instead, it requires only that the techniques in which the bit patterns are stored and accessed on the media needs to be preserved.

Just as the knowledge of the techniques used to store data on a digital media must be preserved, so must the information on file formats be collected on data files created using different software applications. The knowledge of file formats is a collection of the techniques used by specific software applications to define formatting operations within digital documents. Software applications that create digital documents use data located in specific positions and predefined character sequences to define the digital document's appearance. Interpretation software is necessary to view a digital document whether it is simply stored in an ASCII text format or in a complex database format. Software products, commercial-off-the-shelf (COTS) and Non-COTS, store digital data using a variety of techniques. Therefore, every data file is dependent upon some form of software to properly interpret and display the data file's contents. Character sequences embedded within a digital document inform the interpretation software how the document's data is to be interpreted. For example, in order to bold a section of text using the Hypertext Markup Language (HTML), all characters following the character sequence "<B>" are bolded until the character sequence "</B>" is encountered. Any software capable of interpreting an HTML document must recognize these

character sequences and all other format character sequences that are characteristic of HTML documents. Likewise, any software capable of interpreting a digital document must recognize the formatting character sequences unique to the application that was used to create that digital document. File formats knowledge will also be stored in a metaknowledge archive. This will be further discussed in the knowledge preservation section.

### 6.2 Data recovery

Data recovery is the process of extracting digital data from an obsolete medium and migrating it to a medium that is accessible to current information systems. The recovery will, of course, depend on the cost effectiveness of recovering the data. That is, if the need for the knowledge in the digital document(s) is greater than the cost of recovery, then the cost of the recovery method(s) may be justified.

### 6.3 Document reconstruction

Document reconstruction is the process of interpreting digital documents from their original data files by using file format information gathered during the knowledge preservation process. Interpreting digital documents by describing how the original software interpreted the documents is a strategy that was suggested by [15]. The file format information describes the formatting information used by specific software applications. In other words it is a template that can be used to describe the way data is formatted and displayed by word processing, graphics, and other applications that create digital documents. This does not mean that the algorithms used to produce the documents are preserved so programmers can replicate them in the future. Instead, it means that the bit or character sequences and other formatting information are preserved as a template for document interpreters to use to reconstruct and view documents in their original forms. When the reconstruction process is complete the document should appear in its original form. As in the data recovery process, the methods used during document reconstruction are dependent upon the cost effectiveness of reconstructing the document.

## 7. Knowledge Preservation

The metaknowledge archive is the foundation upon which the DRS is built. It contains templates, which can be used to extract, and display data in the form prescribed by the information systems used to create digital documents. To insure the success of the DRS the metaknowledge archive must develop a standardized format to preserve media storage techniques so engineers can extract data from the many different types of media. Likewise, it must also develop a standardized format to preserve digital document formatting information for the different types of digital documents that may need to be recovered.

As long as there is a template that can be used to interpret a document, then a document can be displayed in its original form. After the creation of a digital document, its interpretation is dependent upon the hardware and software systems that were used to create it. However, most modern computer systems have the ability to process and display the multitude of objects that appear in digital documents. Therefore, on any given hardware system routines can be designed to interpret and present the contents of digital documents that were created on another system (even if the systems themselves are incompatible).

### 7.1 Media metadata

Media metadata is probably the easiest type of data to gather for the DRS. This is because the standards for most storage media are rigidly defined before a media is brought to market. For example, ISO9660 is the standard that specifies how data are stored on a CD-ROM. This standard defines the volume structures, file structures, and all other attributes associated with a CD-ROM. This type of data must be gathered for each type of media to be included in the metaknowledge archive.

When trying to recover data, recovery personnel must know where to look in order to find it. Media storage geometry defines where on a medium data are stored. In order for data recovery personnel to find the data they must know the geometric shape of the data's path and the locations of those paths. For example, on a CD-ROM data are stored on a spiraling track with adjacent tracks 1.6 micrometers apart for a track density of 16,000 tracks per inch [23]. Furthermore, the tracks are divided into sectors containing 2048 bytes of data and each sector has an address that is used during the file allocation. This type of geometric storage information must be collected for each type of medium.

After the medium's storage geometry has been identified, data recovery personnel must know the method used to store the data. The data storage method refers to how data are physically recorded on a medium and this information must be known so a device can be engineered to read the digital patterns. In the past, data have been stored on media using a variety of methods. Early storage media stored data as a series of holes punched into lengths of paper tape or punched cards. Hard and floppy disks store data as a series of magnetic patterns stored on a layer of magnetic particles. More recent optical technologies, such as the CD-ROM, store data as a series of lands and pits (0.12 micrometers deep and 0.6 micrometers in diameter) burned into a plastic platter. There are many other storage methods that have been used, that are in use, and that will be used in the future. Knowing these storage methods tells data recovery personnel what to look for to identify the digital data stored on the media.

After data recovery personnel have identified where the data are stored, and the data storage method, they must determine how the data are encoded. Encoding techniques define how the data's bit patterns are stored on the media. The encoding information will be used to decode the data and restore the data bit stream to its

original form. Encoding schemes may be fairly simple with one setting identifying a 0 bit and another setting defining a 1 bit. Or encoding schemes may implement coding algorithms to encrypt and compress recurring bit patterns. Two popular encoding schemes used today are multiple frequency modulation (MFM) and run length limited (RLL). Multiple frequency modulation is a method of encoding analog signals into magnetic pulses or bits. Run length limited is another method of encoding data into magnetic pulses but its encoding scheme allows 50 percent more data to be stored on a disk than MFM.

During the next step it is necessary to determine the file allocation method used on a media. File allocation is how storage space is assigned to files so that storage space is effectively utilized and files can be accessed [29]. Once data recovery personnel can locate, read, and decode the information on a media, they must know the file allocation method in order to properly reassemble the files. Descriptions on items such as volume and file structures are identified in media standards, such as ISO9660 for the CD-ROM. The operating system also controls a media's file allocation method and therefore, it is necessary to access operating system specifications to gather data on file allocation methods. There are several file allocation methods in use and each operating system and media combination uses a specific allocation method. Examples of some popular allocation schemes are the contiguous, linked, and indexed allocation methods. The contiguous allocation method requires each file to occupy a set of contiguous addresses on a disk. With linked allocation each file is a linked list of sectors and the sectors may be scattered anywhere on the disk, and with the indexed allocation method each file has its own index block which is an array of disk block addresses [29]. The allocation method may also provide other valuable information such as distinguishing between the locations of data bytes and error detection/correction bytes.

Collecting and maintaining metadata on these four entities, data storage geometry, storage methods, encoding schemes, and file allocation methods will provide the keys to recover data once an access system is no longer available to access that media type. As hardware and software systems become obsolete this metadata is used to develop hardware and software systems to recover data and migrate it to currently accessible storage media

### 7.2 File format metadata

The first step in gathering file format information is to identify all of the applications used to create the digital documents which may need to be reconstructed in the future. This includes both commercial-off-the-shelf (COTS) and non-COTS applications. Gathering and cataloging metadata to reconstruct digital documents created with COTS and non-COTS applications is going to be a time intensive and difficult task. However, it is necessary because many organizations use these applications to create and store digital documents.

The second step is to identify and catalog the objects that are supported by these applications. An object in a digital document can be text, graphics, audio, video, and any number of other structures that have been included by the document's creator. It is necessary for an interpreter to have the ability to identify the objects embedded in a digital document before the interpretation process begins. If an object is not properly identified then the document is uninterpretable.

Once the objects are identified, interpretation routines are created to present these objects in their original form on the current information system. Since objects are utilized over and over again by different applications, it is only necessary to create a routine to interpret and display that object once. A routine can be used to display an object regardless of the application used to create the digital document. For instance, most digital documents support the use of text objects. Since text is used in multiple applications, it is only necessary to create a routine to handle a text object once. That routine can then be used to interpret and display text on the current system regardless of the software and hardware systems that were used to create the original document.

The final step is to identify and catalog the formatting structures implemented within each application. These formatting structures describe how objects are identified, formatted, and arranged within a digital document. Additionally, this information describes how to determine such things as page size, margins, line spacing, tabs, fonts, footnotes, and a multitude of other page layout information. This information must be maintained in a standardized form so that an interpreter can easily access it and switch between digital documents that were created by different applications. The formatting process may be made more difficult because there is no standardized way in which applications store formatting information. Applications disperse formatting information (1) throughout the document, (2) in designated locations within the document, or (3) in combinations of 1 and 2. Additionally, some applications store document files in an ASCII format while others opt for a binary format. Defining a standardized method to describe these currently non-standardized procedures is one of the goals of the DRS metaknowledge archive.

## 8. Data Recovery

Once it is no longer economically feasible to maintain antiquated hardware systems, it is necessary to implement an alternate method to maintain the ability to recover data from superseded media. If data are stored on an obsolete medium that is not accessible by current systems then the data must be migrated to a currently accessible media before document reconstruction can begin. That is, the data must be recovered.

Data recovery involves the use of the storage technique information gathered during the knowledge preservation process to recover data from an obsolete media. This information is used to modify or construct

the equipment needed to migrate digital data from an obsolete medium to one that is currently accessible.

An example, of this usage can be depicted by data stored on punched cards. Punched cards pass through a punched card reader at the rate of approximately 1,000 cards per minute. As the cards pass between a light source and a row of photo-electric cells the location of the holes are detected and the pattern is transformed into electric signals which are sent to the computer and translated into machine [10]. Because of advances in storage technologies, punched cards are seldom used as a storage medium today because they are slow, bulky, and cumbersome compared to modern storage media. Because of this, few organizations maintain punched card today. So, if stacks of punched cards were to be found and there were no punched card readers available to read the data, how could the data be read? First, the punched card storage technique information that was gathered during the knowledge preservation process is retrieved. Once the information is analyzed and engineers understand the way information is stored on a punched card, they may find that it is a simple task to reprogram a modern scanning device, such as those used in supermarkets or on assembly lines, to read the patterns of holes on a punched card. Therefore, a device can be modified to read, translate, and migrate the data on punched cards to a modern storage device without the need for an original punched card reader. There is no need to engineer a device to write punch cards because there is no desire to change the data. The need is only to read the data and migrate it to a currently accessible storage medium.

While this is a relatively simple example of how the storage technique information can be used, it demonstrates how easily yesterday's digital technologies can be more easily reproduced using today's digital technologies. Likewise, this same method could be used to manufacture readers for paper tapes, CD-ROMs, and other storage devices. If someone finds a CD-ROM disk in the year 2222, perhaps he or she will be able to take it to a DRS processing center to recover the data. Instead of building a CD-ROM drive, the processing center may simply use a high-tech scanner to scan the disk and identify the patterns of lands and pits burned into the disk's surface. Using the data gathered about CD-ROM storage techniques during the knowledge preservation process, an information system analyzes the location and patterns of lands and pits, identifies the file allocation system, processes the data, and then writes the files to a twenty-third century storage device.

## 9. Document Reconstruction

If digital documents are stored in superseded formats then they must go through an interpretation process in order to restore them to their original forms. That is, the documents must be reconstructed. Document interpreters, which accomplish reconstruction, are either (1) trained technicians or (2) software applications that use file-formatting information to reconstruct digital documents.

The DRS relies upon the file format descriptions gathered during the knowledge preservation stage to describe how the original software interpreted files. These file format descriptions identify the information, such as character sequences (and their locations if they are position sensitive), that identify data objects and specify formatting operations within a digital document.

Table 1. **Example character sequences for bold**

| Software Application | Begin Bold | End Bold |
|---|---|---|
| WordStar® | 02 | 02 |
| Ami Pro® | 3C 2B 21 3E | 3C 2D 21 3E |
| HyperText Markup Language | 3C 42 3E | EC 2F 42 3E |

Table 1 contains examples of the character sequences used by three different applications to perform **bolding** operations on text.

When an interpreter is reconstructing an Ami Pro® 3.1 document, the character sequence (hexadecimal values) "3C 2B 21 3E" specifies to the interpreter that all characters following this sequence need to be bolded. Likewise, the character sequence (hexadecimal values) "3C 2D 21 3E" signals the interpreter to stop the bolding process.

This is a simplified view of how file format information can be used, but it demonstrates the types of information that need to be collected and stored to aid document interpreters in the reconstruction of all types of digital documents. In addition to identifying text-based objects and operations, character sequences are used to identify other objects imbedded within digital documents.

## 10. The Digital Rosetta Stone Model

As described above, the DRS model can be represented in three stages. The first stage of the model represents the knowledge preservation process. This is the foundation upon which the DRS is dependent. During this process the data needed to support the data recovery and document reconstruction processes is gathered and stored in the metaknowledge archive.

The second stage of the model is the data recovery process. The data recovery processes uses the knowledge of storage techniques to extract a digital document's bit stream from an obsolete storage device and then migrates the bit stream to a currently accessible storage device. Once a digital document's bit stream has been recovered the bit stream is advanced to third stage.

The third stage of the model is the file reconstruction process. The document reconstruction process uses the knowledge of file formats to interpret the bit stream and display the document in its original form. Upon completion of the reconstruction process, the final product is a reconstructed digital document that appears

in its original form. The complete DRS model is depicted in Figure 1.

The theory behind the Digital Rosetta Stone (DRS) can be demonstrated using an 8-track punched paper tape (8-TPPT). The 8-TPPT technology was widely used during the 1960s and 1970s. This technology was developed before industry standards were the norm and therefore, this technology is largely proprietary. Finding information on the 8-TPPT coding scheme was very difficult. While doing research for this paper, we contacted the technical support and archive sections of the IBM Corporation to get some information on 8-TPPT equipment. Unfortunately, we were told that IBM no longer supported this technology and does not maintain any information in its archives on it. However, some functional 8-TPPT readers still exist.

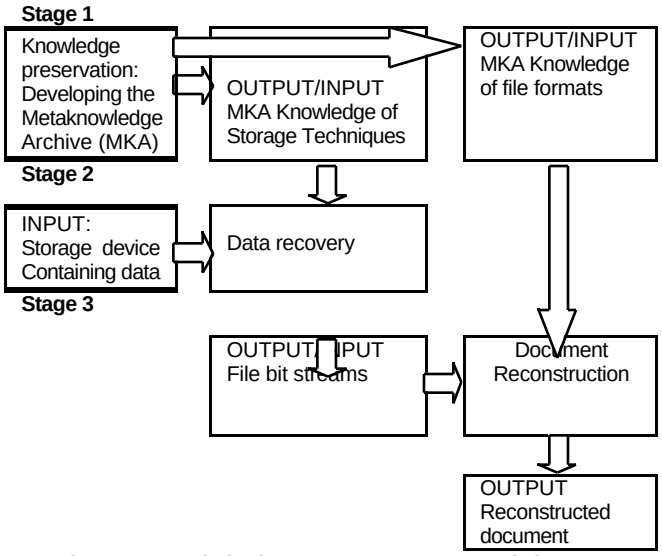



**Figure 1. Digital Rosetta Stone Model**

After being unable to locate a listing of the character-coding scheme, several aging data processing books were consulted to find the information. While much of the coding scheme was obtained from these books, the set is far from complete. The books used to compile this information were written by Awad [2], Nashelsky [18], Langenbach [12], and Williams [32]. All of the information concerning the 8-TPPT used in this example was compiled from these sources.

The 8-TPPT stores data sequentially along the length of the tape. Individual characters are stored vertically on the tape in eight channels. The eight channels represent seven data channels and one check (or parity) channel. From the least significant bit to the most significant bit these channels are identified as 1, 2, 4, 8, Check, "O", "X", and the End of Line (EL). An example of 8-TPPT can be seen in 2. Notice that unlike today, the check bit is not the most significant bit, but instead is in the fifth bit position.

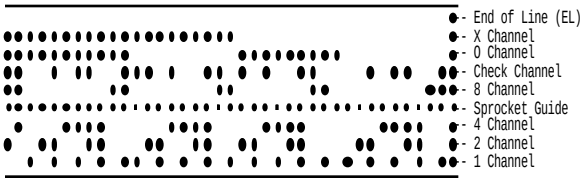



**Figure 2. Example of 8-track paper tape**

Data are stored in the eight channels as follows:

- A punch or combination of punches in channels 1, 2, 4, and 8 represent numeric characters
- A punch in the Check channel is only to be used as a parity check (odd parity is generally used)
- A punch in the "O" and "X" channels are used in combination with channels 1, 2, 4, and 8 to define alphabetic characters, symbols, and other functions such as shift to upper case, shift to lower case, or stop
- A punch in the EL channel represents the end of a line and performs the same function as the return key on a typewriter

The patterns for upper case and lower case alphabetical characters are identical. This is because the equipment used to print documents stored on 8-TPPT operated in a fashion similar to typewriters. That is, shift keys were used to define the difference between upper and lower case characters. Once a shift to upper case symbol was encountered, the type basket was shifted to the upper case position, and all of the characters that followed were typed in the upper case mode, until a symbol was encountered to shift back to lower case. This ability to shift from upper case to lower case mode, and vice versa, provided the ability to use an identical bit pattern for two separate symbols. This along with other meta-information about 8-TPPT is necessary to recover information from 8-track paper tape. [18].

Upon examining the media storage technique information on 8-TPPT in the DRS, engineers find that they can reprogram a modern day scanner to interpret the bit patterns represented by the series of holes and migrate the data to a modern storage device. As the 8-TPPT is scanned, it logically partitions the tapes horizontal tracks and vertical byte regions of the tape. An algorithm analyzes the data regions of the tape and converts the regions with no holes to a 0 and converts the regions containing holes to a 1. The bytes are then assembled into a bit stream, and migrated to a currently accessible storage medium. Once the bit stream is transferred to an accessible medium, it can be interpreted using 8-TPPT file formatting data that has been preserved in the metaknowledge archive. Using information from the MKA, an interpretation algorithm reads the bit stream from the advanced media and breaks the bit stream into 8-bit bytes.

The algorithm performs an error checking routine based on the fifth bit of the 8-bit byte to insure that the integrity of the data has not been compromised. Once error checking is complete, the 7-bit characters are

mapped to the 8-bit character codes that can be displayed by the current system. When mapping the 8-TPPT 's 7-bit characters to the character codes used by the current system it is necessary to use a translation table which maintains two translation schemes--one for upper cased characters and one for lower cased characters. This is because the 8-TPPT character codes receive double use. That is, the same code used for the character "A" (0110001) was also used for the character "a" (0110001). The difference in character case was determined by the position of the type basket. Therefore, the algorithm translating the character set will have to track the position of the type basket and translate the characters appropriately.

Once the character set has been translated the document can be printed. However, this is not as easy as it sounds. Many modern word processing operations, such as bolding, centering, and underlining are transparent to the document creator. However, the keyboarding techniques of the 1960s and 1970s were not as convenient. For example:

- To bold text an individual had to type the text to be bolded, backspace to the beginning of that text, and then retype over the text.
- To center text an individual had to tab to the center of the page, backspace one-half of the total number of characters to be centered, and then type the text.
- To underline text an individual had to type the text to be underlined, backspace to the beginning of that text, and then use the underscore key to underline the text.

Therefore, to accurately reconstruct these documents, algorithms have to identify and translate these types of operations.

After all of this knowledge is brought to bear, a document stored 40 years ago can be recovered and printed. In the future, we will have many more difficult tasks of digital document reconstruction. The DRS can be a significant agent in helping to ensure that we don't lose our ability to read our own history.

## 11. Conclusions

In this paper, we researched the problem of maintaining long-term access to digital documents. We reviewed the methods that have been suggested by others, and combined them with additional ideas to create a model we call the Digital Rosetta Stone. The Digital Rosetta Stone describes a method by which we will be able to maintain long-term access to our increasing repositories of digital documents.

The development of a DRS will be a time intensive and expensive task. Consider the vast number of research projects, books, and museums that have been propagated in order to maintain access to our written history. The mechanics of the written language changes slowly over decades and centuries. However, new technologies for capturing and storing digital documents are evolving faster than ever. This rapid development calls for the preservation of the vast amounts of digital knowledge that has been and is being created. However, unlike written documents, the preservation of digital documents also requires the preservation of the knowledge and technology necessary to access these documents. The Digital Rosetta Stone presents a model for achieving that end.

## References


[1] Adcock, Ken, Marilyn M. Helms, and Wen-Jang Kenny Jih. "Information Technology: Can It Provide a Sustainable Competitive Advantage?," *Information Strategy: The Executive's Journal*, *9*: 10-15 (Spring 1993).

[2] Awad, Elias M. *Business Data Processing, Third Edition*. Prentice-Hall, Inc., 1971

[3] Beatty, Jeff. "State Office Streamlines Records," *Managing Office Technology*, *40*:58-61 (November 1995).

[4] Boar, Bernard H. "Logic and Information Technology Strategy: Separating Good Sense from Nonsense," *Journal of Systems Management*, *45*: 16-21 (May 1994).

[5] Conway, Paul. *Preservation in the Digital World*. The Commission on Preservation and Access, March 1996.

[6] Cooper, Donald R. and C. William Emory. *Business Research Methods, Fifth Edition*. Richard D. Irwin, Inc, 1995.

[7] Curle, Howard A., Jr. "Supporting Strategic Objectives: Building a Corporate Information Technology Structure," *Information Strategy: The Executives Journal*, *10*: 5-12 (Fall 1993).

[8] Darling, Pamela W. "Creativity vs Despair: The Challenge of Preservation Administration," *Library Trends, 30*: 179-188 (Fall 1981).

[9] Dollar, Charles M. *Archival Theory and Information Technologies: The Impact of Information Technologies on Archival Principles and Methods*. Publications of the University of Macerata, 1992.

[10] Downing, Douglas and Michael Covington. *Dictionary of Computer Terms*. Barron's, 1986.

[11] Gehling, Robert G. and Michael L. Gibson. "Using Imaging to Reengineer Business," *Information Systems Management*, *12*: 55-60 (Spring 1995).

[12] Langenbach, Robert G. *Introduction to Automated Data Processing*. Prentice-Hall, Inc., 1968.

[13] Lynn, M. Stuart. "Digital Imaging Technology for Preservation," *Proceedings from an RLG symposium held March 17 and 18, 1994 Cornell University, Ithaca NY.* 1-10. Research Libraries Group, 1994.

[14] Martin, James. *Information Engineering Book I Introduction*. Prentice Hall, 1989.

[15] Michelson, Avra and Jeff Rothenberg. "Scholarly Communication and Information Technology: Exploring the Impact of Changes in the Research Process on Archives," *American Archivist, 55*: 236-315 (Spring 1992).

[16] Mohlhenrich, Janice, editor. *Preservation of Electronic Formats & Electronic Formats for Preservation*. Highsmith Press, 1993.

[17] Morell, Jonathan A. "The Organizational Consequences of Office Automation: Refining Measurement Techniques," *Data Base, 19*: 16-23 (Fall/Winter 1988).

[18] Nashelsky, Louis. *Introduction to Digital Computer Technology, Second Edition*. John Wiley and Sons, 1972.

[19] National Academy of Public Administration. *The Effects of Electronic Recordkeeping on the Historical Record of the U.S. Government. A Report for the National Archives and Records Administration*. January 1989.

[20] National Research Council. *Study on the Long-term Retention of Selected Scientific and Technical Records of the Federal Government Working Papers*. National Academy Press, 1995.

[21] *Preserving Scientific Data on Our Physical Universe: A New Strategy for Archiving the Nation's Scientific and Technical Data*. National Academy Press, 1995a.

[22] *Preservation of Historical Records*. National Academy Press, 1986.

[23] Norton, Peter, Lewis C. Eggebrecht, and Scott H. A. Clark. *Peter Norton's Inside the PC, Sixth Edition*. SAMS Publishing, 1995.

[24] OASD (Office of the Assistant Secretary of Defense). *Automated Document Conversion Master Plan, Version 1*. April 1995.

[25] Olson, Margrethe H. and Henry C. Lucas Jr. "The Impact of Office Automation on the Organization: Some Implications for Research and Practice," *Communications of the ACM, 25*: 838-847 (November 1982).

[26] Peterson, Del. "Case Study: Improving Customer Service Through New Technology," *Journal of Information Systems Management, 8*: 28-35 (Spring 1991).

[27] Rothenberg, Jeff, "Ensuring the Longevity of Digital Documents", *Scientific American*, 42-47 (Jan 95)

[28] Schnitt, David L. "Reengineering the Organization Using Information Technology," *Journal of Systems Management, 44*: 14-20+ (January, 1993).

[29] Settani, Joseph A. "Making the Jump from Paper to Image," *Managing Office Technology, 40*: 15-28 (April 1995).

[30] Silberschatz, Abraham and James L. Peterson. *Operating System Concepts, Alternate Edition*. Addision-Wesley Publishing Company, 1988.

[31] Smith, Milburn D. III. *Information and Records Management: A Decision-Maker's Guide To Systems Planning and Implementation*. Quorum Books, 1986.

[32] van Nievelt, M.C. Augustus. "Managing With Information Technology--A Decade of Wasted Money?," *Information Strategy: The Executive's Journal, 9*: 5-17 (Summer 1993).

[33] Williams, William F. *Principles of Automated Information Retrieval*. The Business Press, 1965.

# The Universal Preservation Format
## Background and Fundamentals

Thom Shepard, Dave MacCarn
thom_shepard@wgbh.org, dave_maccarn@wgbh.org
WGBH Educational Foundation
Boston, MA, USA

**Sixth DELOS Workshop**
**Preservation of Digital Information**

## I. Background

> *Film, video and sound recordings are vital components of our collective memory... This vast source of information, inspiration and creativity—the most known contemporary archive of our society—is threatened. ...[We] are losing large parts of our recorded past."*[1]

According to a recent Library of Congress report, video materials in the public and private sector are estimated to exceed several hundred thousand recorded hours. The same report judges the amount of feet of news film and other film used to record television programming to total in the several millions.[2] Much of this historical material is in danger of being lost.

> *[The] preservation of television and video materials faces enormous obstacles, in particular, the vulnerability of videotape to adverse storage conditions, abusive handling, and technological obsolescence.*
>
> William T. Murphy, Coordinator
> Report on the State of American
> Television and Video Preservation
> In a letter dated January 16, 1996

With more than one hundred sixty thousand hours (160,000) of video programming, WGBH has first-hand knowledge of these vulnerabilities, which is rapidly being translated into monetary considerations. For example, while income recovered from the first airing of a production used to be 100%, that figure is now 55%, with an additional 44% of the total income generated being made from additional sales of materials from the production. In short, recovering production costs relies heavily on materials from our archives.

Technological obsolescence, in particular, has hindered the preservation of film and video materials by contributing to the enormous expense of accessing stored materials. The standard format for recording television programs has taken several forms over the past fifty years, including kinescope, 2» videotape, 1» videotape and digital tape. As the standard for recorded programming continues to evolve, the equipment used to access materials produced in earlier formats has become increasingly difficult to find and, accordingly, more expensive to use.

Archivists and industry members have addressed the problem by transferring older formats to digital tape, thus attempting to maintain the quality of the original material. However, this costly process simply puts off—rather than solves—the preservation problem. The enormous and rapid changes taking place in digital technology have resulted in a veritable explosion of formats. Fourteen different digital tape formats are available at present (D-1, D-2, D-3, D-5, D-6, Digital Betacam, Betacam SX, Ampex DCT, Consumer DV, DVCAM, DVCPRO (D-7), DVCPRO-50, Digital S and D-VHS) with several more for High Definition Television. With these format wars heating up, many of these formats may soon become obsolete, making them unsuitable for preserving media information. In addition, digital non-linear editing systems have internal proprietary media formats.

From an archivist perspective this is a nightmare. On one side of the room you store the tapes and on the other side the tape machines and spare parts.

As with the videotape materials produced during the last fifty years, technical obsolescence may make digital formats that are common today inaccessible tomorrow. There is a significant need for a Universal Preservation Format (UPF), designed specifically for digital technologies, that can store compound content (not only the media itself but also information about it) so that it can be accessed easily both today and into the indefinite future.

This thinking is reflected in a national plan for redefining film preservation put forth by the Librarian of Congress in August of 1994. As one of his eight recommendations, he suggested that:

> *[The new electronic technologies] are already transforming film access but archives should insist that certain stringent criteria be met before new technologies are adopted as preservation media.*[3]

Paul Messier, Conservator of Photographs and Works on Paper for the Boston Art Conservation, also called for the establishment of criteria for assessing digital video as a preservation medium in a paper that he presented at Playback '96: A Round Table on Video Preservation. His suggested criteria were adapted from those suggested for still images by Basil Manns, Research Scientist at the Library of Congress in his article «The Electronic Document Image Preservation Format.» These «criteria» anticipate the technical specifications necessary for the selection and description of data to be preserved through a UPF. However, no further action has been taken within the field of archives.

Let there be no ambiguity on this key point: we are emphatically not advocating yet another acquisition format, universal or otherwise. Additionally we need to differentiate between digital archives, which are concerned with the timeless storage of digital materials, and digital libraries, which is primarily concerned with timely issues of access. Our proposed solution is to establish a format expressly for the permanent archival storage of digital materials.

## Background

The concept of a universal preservation format was introduced at the Society of Motion Picture and Television Engineers Conference, October 1996. Sponsored by the WGBH Educational Foundation and funded in part by a grant (97-029) from the National Historical Publications and Records Commission of the National Archives, the Universal Preservation Format initiative may be summed up in the following passage:

> *a platform-independent Universal Preservation Format, designed specifically for digital technologies, that will ensure the accessibility of a wide area of data types -- especially video formats -- into the indefinite future. WGBH will work with both technology manufacturers and archivists to determine a UPF that meets the needs of both non-commercial and commercial interests. At the end of the two-year grant period, WGBH will submit a Recommended Practice to the Society of Motion Picture and Television Engineers (SMPTE), a standards-creating organization, and the Association of Moving Image Archivists (AMIA).*
>
> (Project Summary, "Statement of Purpose," p. 1)

## Mission of UPF

Our mission, as outlined in the original UPF grant proposal, includes the following key tasks:

- to analyze the problem of preserving the video digital data contained in electronic records
- to raise awareness of electronic records preservation
- to build support for an effective solution
- to develop a Recommended Practice for a UPF and encourage its adoption as a cross-industry standard
- to present the concept of a universal preservation format to both archivists and technology manufacturers at professional conferences and working groups, and through articles in professional journals.

Working with representatives from standards organizations, hardware and software companies, museums, academic institutions, archives and libraries, this project will submit the final draft of the Recommended Practice

to the Society of Motion Picture and Engineers (SMPTE). It will suggest guidelines for engineers when designing computer applications that involve or interact with digital storage. We expect to make the process of preserving and accessing electronic records (both original and migrated) more efficient, cost-effective, and simpler.

## Reaching out

From the very beginning of this project, we have actively solicited the participation of several organizations. Our ever-growing database of contacts includes members from the Association of Moving Image Archivists, the Society of American Archivists, the Music Library Associations, Boston Art Conservation, and Conservation Online, as well as individuals who may not be members of these groups but who nonetheless are actively involved in preservation issues. In addition to the UPF listserv[4], which currently has close to sixty members, postings go to the AMIA listserv, the Archivist & Archives listserv, and the Electronic Records listserv. Mailings are also send to archivists who do not have email addresses. A task item planned is to set up teleconferences between engineers and archivists.

On September 22, 1997, the Society of Motion Picture and Television Engineers assigned an official Study Group (ST13.14). At this first meeting, the following objectives and tasks were established:

Statement of Objectives

- The study group will document requirements of data formats for the preservation of electronically generated media and related information.
- Extensive input from the archival community will be gathered through the use of surveys and meetings.

Specific Tasks and/or Documents

- Some areas of study: Containers, Objects, Labels, Metadata, Composition, Rosetta stone for future coding and translation, and the coordination of other SMPTE activities that may be useful.
- Explore the possibility of a universal format and guide to the storage of collections.
- Based on the requirements gathered, the study group will investigate available technology and explore configurations in order to provide a basis for a working groups' recommended practice or standards.

(SMPTE Engineering Committee Work Assignment/Work Statement)

## User Survey

What are these needs and concerns? UPF established a user survey.
Though many archivists said that they realized they would have to "migrate" at some point, most could not justify the costs of either migrating to digital or investing in new digital equipment that will only become obsolete in a few years. Running throughout these commentaries was the frustration that archivists had no control over new technologies. And while digital has qualities that are enormously appealing to archivists -- searchability, mobility, longevity -- computer technologies seem disposable, like snakes shedding their skins.

Some archivists reported that they are feeling pressure from administrators to go digital for all the wrong reasons: consolidating their collections, for example.

For those already involved in some form of digital conversion, the strategy has generally been to convert from analog to digital in an ad hoc manner. No one has developed strategies for replacing their analog collections with digital formats.

The published results of the survey are on the web site[5]. In addition there are posted follow-up questions that invite comment.

# II. Technical Specifics of the UPF

The first step is to separate the data format from the storage format. We can look at this separation through the use of essence, metadata, wrappers and identifiers.

## Essence

Media can be thought of as a data object. These objects can be data types, such as video or music. The term often used to label these data types is «essence.» Many software applications are capable of interacting with a range of these essence files through the use of interchange formats. For example, in the transfer of word processing documents across applications and even operating systems, the Rich Text Format (RTF) is often used. For video, a standard for interchange is SMPTE 259M.

## Metadata

A data object can also take the form of information about the data types. In terms of function, this information, for which the term "metadata" has been coined, may be divided into four basic categories: format, description, association, and composition.

## Wrapper

The wrapper -- or container -- is a file format for storing "essence" along with the information that describes it. The wrapper is a file format that has a framework structure. Anyone familiar with the Dublin Core metadata initiative, specifically the Warwick Framework Architecture, may have some understanding of frameworks as a method for managing data. Warwick describes a metadata structure in which material describing certain objects may either be embedded in the source or be referenced to files or storage areas external to the source. This information might include domain specific descriptions, terms and conditions for document use, pointers to all manifestations of document, archival responsibility, and even structural data.

## Unique Identifier

Identifying digital objects as unique entities is essential to establishing archival integrity, especially when it is so easy to misplace, corrupt or delete digital information. The UPF is looking at initiatives dealing with unique identifiers and expects to include such a system or systems in our Recommended Practice. Basically, each object carries an ID that is unique within its container. As this object undergoes changes, often called "versioning," each new generation is assigned its own identifier, which always references its parent.

## Self Describing Format

The foundations of essence, metadata, wrappers and identifiers can create a «self describing» format. The UPF uses a "digital Rosetta stone" to get at the range of data types held in a digital storage bank. The digital Rosetta stone serves as a key, defining the data types and encapsulating algorithms for deciphering the file. The use of platform-independent algorithms is used to decode file types. The Rosetta stone might also serve as local registry for unique identifiers.

## Use of existing technology

These self-describing technologies are already available. Along with the surge of digital formats are technologies that are designed to handle digital media of all types. Apple Computer's «Bento Specification»[6] , Avid Technology's «Open Media Framework Interchange Specification»[7] and the Society of Motion Picture and Television Engineers/European Broadcast Union's «Harmonized Standards for the Exchange of Television Program Material as Bit Streams.[8]» are media technologies that approach the UPF concept .

## Bento

Apple Computer's «Bento Specification» is the underlying technology of Apple Computer's OpenDoc Standard Interchange Format.[9] Bento is a specification for storage and interchange of compound content. Bento defines a standard format for storing multiple different types of objects and an API to access these objects. An object container is just some form of data storage (such as a file.) This storage is used to hold one or more objects (values) and information about the objects (metadata.) Bento containers are defined by a set of rules for storing multiple objects, so that software that understands the rules can find the objects, figure out what kind of objects they are, and use them correctly.

Bento objects can be simple or complex, small (a few bytes) or large (up to $2^{64}$ bytes, approximately $2^{27}$ hours of D-1 video.) Bento is designed to be platform and content neutral. so that it provides a convenient container

for transporting any type of compound content between multiple platforms. The Bento code currently runs on Macintosh, MS DOS, Microsoft Windows, OS/2 and several varieties of Unix.[10]

## OMF

Avid Technology's «Open Media Framework» (OMF) Interchange, a standard format for the interchange of digital media data among different platforms, adopted the use of Bento containers. Additionally, the OMF format encapsulates all the information required to transport a variety of digital media such as audio, video, graphics, and still images, as well as the rules for combining and presenting the media. The format includes rules for identifying the original sources of the digital media data, and it can encapsulate both compressed and uncompressed digital media data.

OMF Interchange provides for a variety of existing digital media types and the ability to easily support new types in the future. A single OMF Interchange file can encapsulate all the information required to create, edit, and play digital media presentations.

While OMF Interchange is designed primarily for data interchange, it is structured to facilitate playback directly from an interchanged file when being used on platforms with characteristics and hardware similar to those of the source platform, without the need for expensive translation or duplication of the sample data. OMF Interchange provides for the development and integration of new media and composition types.[11]

## SMPTE/EBU

Based on Bento and OMF with the addition of unique identifiers.

## Bento, OMF and SMPTE/EBU as a Preservation format

Preservation requires the handling of many different recording formats—such as 2» videotape, 1» videotape, D-1, D-2, D-3, and others—which can be thought of as having data types (the way the video is encoded, e.g. 4:2:2, $4f_{SC}$.) Although Bento allows for any data type, the OMF Interchange only defines a <u>minimum</u> number of data types (e.g. TIFF, RGBA and AIFF). By adding <u>additional</u> standard data types to the these technologies would result in a storage container format that will be able to encompass all present recording forms and allow for all future forms. In moving from the raw recording format (e.g. videotape) to a data tape (or other media) format that incorporates the UPF, the number of formats that archivists need to preserve will be substantially reduced. The UPF breaks the bond between the recording format and the machine through which the format is accessed.

## Hybrid Technology

Although these technologies bring us a self-describing format there is still the hurtle of reading them with out the native machine that recorded then.

*«Reading and understanding information in digital form requires equipment and software, which is changing constantly and may not be available within a decade of its introduction. ... We cannot save the machines if there are no spare parts available, and we cannot save the software if no one is left who knows how to use it.»*

«Preserving Digital Information» - 1996 Report of the Task Force on Archiving of Digital Information, Commission on Preservation and Access and The Research Libraries Group.

The answer is in our old friend analog. We still use microfiche. We can apply this to digital storage in the form of a hybrid solution. An out of the world example in the 1977 Voyager Interstellar Outreach Program[12]. The Voyager spacecraft included a gold plated copper disk with recording from the planet earth. The playing instructions are in a symbolic language. By taking this idea to the UPF one can create the ultimate self-describing format. The analog portion would contain information about the contents along of the disk with the instructions for the creation of a machine to read the disk (blue prints). There are currently technologies to accomplish this from Norsam Technologies[13]. Norsam focuses on the need to greatly increase the storage densities of micro fiche as well as digital recording systems. Using these two technologies the UPF becomes a preservation format that is whole, unto itself.

## Media Compiler

A media compiler would perform the actual moving of data. It would remove the baggage of the acquisition format as it imported into the archive. It would optionally export whatever metadata you needed from the

archive. Specifically, you could pre-select which set of relationships or media formats you wish to transport for a given need, such as Internet access. And because the relationships among your data objects would be built-in, you could very easily "package" information. For example, you could extract certain media objects, along their associative text files, based on a scholar's search patterns. These materials could then be burned into a CD-ROM or transferred onto some other portable storage.

## Summary

• Working with standards organizations, hardware and software companies, museums, academic institutions, archives and libraries.
• Self-Describing Format, Immune to Technological Obsolescence.

A worthy standard for long-term digital storage will carry forth the traditional practices of analog collections. Specifically, a recommended practice must enable provenance and original order. Its framework must be robust, allowing for certain types of metadata to be embedded with the media, while others to be referenced externally. By concentrating on elemental concepts of how data and information about that data might be stored through time, the Universal Preservation Format initiative is attempting to construct a bridge between engineers and information scientists, between those who make and market technical specifications and those who must learn to use the tools of technology to preserve the rapidly decaying fruits of our cultural heritage.

## References


[1] From Fading Away: Strategic Options to Ensure the Protection of and Access to Our Audio-visual Memory. Task Force on the Preservation and Enhanced Use of Canada's Audio-Visual Heritage, National Archives of Canada, June 1995, pi.

[2] Library of Congress, «Redefining Film Preservation: A National Plan» August 1994.

[3] Redefining film preservation: a national plan; recommendations of the Librarian of Congress in consultation with the National Film Preservation Board [coordinated by Annette Melville and Scott Simmon]. August 1994. Recommendation 3.7.

[4] UPF Listserv: UPF@info.wgbh.org: «Subscribe UPF your name» to listserv@info.wgbh.org

[5] UPF [http://info.wgbh.org/upf]

[6] Open Doc underlying technologies, now maintained by the Sunrise project at the Advanced Computing Laboratory at Los Alamos National Laboratory. [http://www.acl.lanl.gov/sunrise/DistComp/OpenDoc/overview.html]

[7] Open Media Framework Interchange Specification, Copyright © 1995 Avid Technology, Inc. [http://www.avid.com/omf/]

[8] SMPTE/EBU Task Force for Harmonized Standards for the Exchange of Program Material as Bit Streams, Copyright (c) 1997 European Broadcasting Union and the Society of Motion Picture and Television Engineers, Inc. [http://www.smpte.org/engr/tfhs_out.pdf]

[9] Open Doc underlying technologies [see 6]

[10] Bento Specification. [see 6]

[11] Open Media Framework Interchange Specification.[see 7]

[12] Voyager Interstellar Outreach Program [http://vraptor.jpl.nasa.gov/voyager/record.html]

[13] Norsam Technologies, Inc. [http://www.norsam.com/]

# CEDARS: Digital Preservation and Metadata

Michael Day
UKOLN: The UK Office for Library and Information Networking,
University of Bath, Bath, BA2 7AY, United Kingdom
http://www.ukoln.ac.uk/
m.day@ukoln.ac.uk

**Abstract**

CEDARS (CURL Exemplars in Digital ARchiveS) is a UK digital preservation project funded by JISC through eLib. Lead sites in the project are the Universities of Cambridge, Leeds and Oxford. The project aims to promote awareness of the importance of digital preservation, to produce strategic frameworks for digital collection management policies and to promote methods appropriate for long-term preservation. An important strand of CEDARS will concern metadata. Metadata could be used as a means of recording migration and emulation strategies, ensuring the authenticity of digital objects, noting rights management and collection management issues and will also be used for resource description and discovery.

## 1. CURL Exemplars in Digital ARchiveS (CEDARS)

### 1.1. Background

University and research libraries have, in recent years, given their users increased access to digital information resources. Some of these form part of their physical collection, e.g. databases on CD-ROM, while others are provided via computer networks and are made available on different commercial terms [1]. At the present time there is no formal mechanism for ensuring that digital resources are preserved for long term use. Indeed in many countries, including the United Kingdom, there is still no formal legal deposit for digital publications [2].

The Joint Information Systems Committee (JISC) of the UK higher education funding councils funds the Electronic Libraries (eLib) Programme. The eLib Programme was set-up in response to a report published in 1993 by a Libraries Review Group appointed by the funding councils [3]. JISC were aware that digital preservation would have an important role in the eventual success (or otherwise) of the eLib Programme. Accordingly, JISC and the British Library co-sponsored a workshop on the "Long Term Preservation of Electronic Materials" which was held at Warwick University in November 1995 [4]. One outcome of this workshop was that JISC, in conjunction with the National Preservation Office (NPO), agreed to fund a programme of studies which would be administered by the British Library Research and Innovation Centre (BLRIC). These JISC/NPO studies covered several distinct areas:

- An analysis of the US Task Force on the Archiving of Digital Information report [5]
- A framework of data types and formats [6]
- Who should be responsible for preservation and access? [7]
- The *post hoc* rescue (data archaeology) of high-value digital material [8]
- The preservation requirements of universities and research funding bodies [9]
- Guidelines for digital preservation [10]
- Comparison of methods of digital preservation [11]

As part of eLib Phase 3, JISC decided to fund a project that would be able to investigate some of the practical issues of digital preservation. The Consortium of University Research Libraries (CURL) is a consortium of research libraries in the British Isles whose mission is "to promote, maintain and improve library resources for research in universities." Digital preservation is a key issue for all CURL members. CURL accordingly submitted a research proposal as part of eLib Phase 3. The result of this is the CURL Exemplars in Digital ARchiveS (CEDARS) project, funded by JISC through the CURL libraries. The project started in April 1998, and will run for three years. The lead sites in the project are the Universities of Cambridge, Leeds and Oxford. UKOLN has some involvement with the parts of the project relating to metadata. Other collaborating institutions include the Arts and Humanities Data Service (AHDS), the British Library, the Data Archive, the NPO and the Research Libraries Group (RLG).

### 1.2. Objectives

The project aims to investigate strategies which will ensure that the digital information resources typically included in library collections may, with other non-digital objects, be preserved over the longer term. It order to achieve this aim the project plans to:

- Promote awareness about the importance of digital preservation, both amongst research libraries and their users, and amongst the data creating and data supplying communities upon which they depend.

- Identify, document and disseminate strategic frameworks within which individual libraries can develop collection management policies which are appropriate to their needs and which can guide the necessary decision-making to safeguard the long-term viability of any digital resources that are included in their collections.

- Investigate, document and promote methods appropriate to the long-term preservation of different classes of digital resources typically included in library collections, and to develop costed and scaleable models.

### 1.3. Scope of demonstrators

Several different types of digital resources will be included within the CEDARS project scope.

- Digitised primary resources
- Datasets
- Electronic journals
- Online databases
- Electronic ephemera - pre-prints, Web pages, subject gateways, etc.
- Digital resources where intellectual content is bound to structure, form and behaviour
- Metadata

The CEDARS project is interested in demonstrating the preservation of all material that is the traditional preserve of the research library. It, however, is not concerned with information in the form of sound or video. It is, additionally, only concerned in preserving the intellectual content of resources, not the physical objects upon which they are stored. Each of the lead sites will take responsibility for providing demonstrators for a particular 'flavour' of digital resource. Cambridge will deal with dynamic data, Oxford with primary resources while Leeds will look at digital resources where intellectual content is bound to structure, form and behaviour (e.g. CD-ROMs). In addition, the three lead sites will lead working groups on those related issues that had been identified as most important: Cambridge on rights management, Oxford on metadata, and Leeds on the use of emulation as a preservation strategy.

### 1.4. Deliverables

Key deliverables of the CEDARS project include the production of:

- Guidelines for developing collection management policies which will ensure the long-term viability of any digital resources included in the collection

- Demonstrator projects to test and promote the technical and organisational feasibility of a chosen strategy for digital preservation

- Methodological guidelines developed by the demonstrator projects providing guidance about how to preserve different classes of digital resources

- Clearly articulated preferences about data formats, content models and compression techniques which are most readily and cost-effectively preserved

- Publications of benefit to the whole higher education community

One of these publications will be a study of digital preservation metadata.

## 2. Metadata and digital preservation

Discussions of metadata in the library community have largely centred on issues of resource description and discovery [12]. There is, however, a growing awareness that metadata has an important role in digital resource management, including preservation. Accordingly, in May 1997 the Research Libraries Group constituted a Working Group on the Preservation Issues of Metadata. The aim of this working group is to ensure that information essential to the continued use of digital resources is captured and preserved in an accessible form. A preliminary report has been produced which identifies 16 preservation metadata elements and provides a semantic framework for this [13].

The CEDARS project also recognised from an early stage that metadata issues would be important. A working group has been formed to cover metadata. At this preliminary stage of the project it is difficult to predict what particular recommendations this working group will produce but interest is likely to be shown in the following issues.

### 2.1. Metadata for emulation and migration

The core technical problems of digital preservation relate to inadequate media longevity, rapid hardware obsolescence and dependencies on particular software products. In this context it makes good sense to preserve the data itself, not the physical medium on which it happens to reside. There are several potential technical approaches to this problem. Jeff Rothenberg has suggested, for example, the building of software emulators that would mimic the behaviour of obsolete hardware and software [14]. This would involve encapsulating data together with the application software used to create it and a description of the required hardware environment. To facilitate future use, Rothenberg suggests attaching 'annotation metadata' to the surface of each encapsulation which would both "explain how to decode the obsolete records contained inside the encapsulation and to provide whatever contextual information is desired about these records" [15]. This surface metadata, which could also contain resource discovery information, would be kept in a standard 'bootstrap' format so that it could be converted to new formats as part of the preservation refresh cycle.

Another approach to digital preservation is the periodic migration of digital information from one generation of computer technology to a subsequent one [16]. Using migration, it is important to ensure that preserved documents are what the US National Historical Publications and Records Commission (NHPRC) funded University of Pittsburgh Electronic Records Project describe (in an archives context) as 'inviolate', 'coherent' and 'auditable' [17]. David Bearman defines 'coherent' as follows: "If records are migrated to new software environments, "content, structure and context information must be linked to software functionality that preserves their executable connections or representations of their relations must enable humans to reconstruct the relations that pertained in the original software environment" [18]. Successful migration strategies will, therefore, depend upon metadata being created to record the migration history of a digital object and to record contextual information so that a future user can reconstruct (or understand) the technological environment in which a particular digital object was created.

### 2.2. Metadata for authentication

In addition to the technical problems of digital preservation, there will be a need to address problems of intellectual preservation [19]. For example, how will users know that the digital object that they retrieve is the one that they want? Again, how can one guard against unauthorised changes being made to the information content of digital objects?

A partial solution to this problem would be the general adoption of unique and persistent digital identifiers. This would mean the assignment of a new identifier each time a particular digital object is updated. Current initiatives include the Uniform Resource Name (URN) which is being developed for the Internet community by working groups of the Internet Engineering Task Force [20] and the Digital Object Identifier (DOI), an initiative of the Association of American Publishers [21]. Legacy identifiers will also continue to be used for some of the digital objects within the CEDARS project scope, so - for example - some publishers will assign International Standard Book Numbers (ISBNs) to CD-ROMs or generate Serial Item and Contribution Numbers (SICIs) for online journal articles. On the other hand, other items in the project scope, electronic ephemera for example, are unlikely to have previously assigned persistent and unique identifiers.

An additional approach to ensuring the authenticity of a given digital object would be to use a simple cryptographic technique like the production of a validation key value or checksum for each resource in a digital archive. An authentication checksum could be computed from each resource in a digital archive and stored with the descriptive metadata. When a user, or the archive, wants to retrieve the resource at a later date this checksum could be computed again and compared with the checksum recorded in the metadata. If the two agree there can be confidence that the document retrieved is the one referred to by the descriptive metadata. This general approach has been adopted for use by the European Telematics for Libraries project BIBLINK [22].

Archivists and records managers have similar concerns with authenticity, integrity and preserving 'evidentiality'. The University of Pittsburgh Electronic Records Project, for example, has defined a metadata model for business-acceptable communications [23]. A University of British Columbia project has also worked on defining the requirements for preserving reliable and authentic electronic records [24].

### 2.3. Metadata for resource discovery.

Digital resources that have been physically preserved will also need to be retrievable. For this reason, preservation systems will have to interact with resource discovery systems. Recommendations on resource discovery formats (e.g. Dublin Core) or metadata frameworks (e.g., Resource Description Format) will constitute an important part of CEDARS work on metadata.

### 2.4. Metadata for rights management.

Solving rights management problems in a digital preservation context will be crucial to a practically based project like CEDARS. Within the project, different licensing arrangements will have to be made with relevant stakeholders. This rights management information can be stored as part of the descriptive metadata and this could be used to manage access to digital resources in the demonstrators.

### 2.5. Metadata for resource evaluation

Not all digital resources will be preserved and, indeed, not all digital resources will be worthy of long-term preservation. CEDARS is interested in helping to develop suitable collection management policies for research libraries. This work could build on work carried out on selection criteria for Internet subject gateways produced by the EU funded DESIRE project [25].

### 2.6. Metadata management

Another important issue is how this metadata will be generated and where it will be kept. Metadata could be stored either in a centralised or distributed database and linked to the original resource. Alternatively, metadata could also be embedded in or otherwise directly associated with the original resource. Different solutions might be possible for different types of metadata. Resource discovery and rights management metadata could form part of a searchable database, while metadata specifying the technical formats used, the migration strategies operated and a document's use history could be stored with the document itself. Over a long period of time, this metadata will grow in size and will itself have to be subject to migration and authentication strategies.

## 3. Conclusions

CEDARS is a project that aims to address strategic, methodological and practical issues relating to digital preservation. The project will include the development of demonstrators to check the technical and organisational feasibility of the chosen preservation strategies. One strand of the project will investigate metadata issues. A preliminary report will be made available later this year and a seminar convened.

## 4. References

[1] Day, M.W., Online serials: preservation issues. In *E-serials: publishers, libraries, users and standards*, ed. W. Jones. The Serials Librarian, 33. Binghamton, N.Y.: Haworth Press, 1998, 199-221.

[2] Hoare, P., *Legal deposit of non-print material: an international overview, September-October 1995*. British Library Research and Development Report, 6245. London: British Library Research and Development Department, 1996.

[3] Joint Funding Councils' Libraries Review Group, *A report for the Higher Education Funding Council for England, the Scottish Higher Education Funding Council, the Higher Education Funding Council for Wales and the Department of Education Northern Ireland* [the Follett Report]. Bristol: HEFCE, December 1993.
<URL:http://www.ukoln.ac.uk/services/papers/follett/report/>

[4] Fresco, M., *Long term preservation of electronic materials: a JISC/British Library Workshop as part of the Electronic Libraries Programme (eLib) organised by UKOLN, 27th and 28th November 1995 at the University of Warwick*. British Library Research and Development Report, 6238. London: British Library Research and Development Department, 1996.
<URL:http://www.ukoln.ac.uk/services/papers/bl/rdr6238/>

[5] Matthews, G., Poulter, A. and Blagg, E., *Preservation of digital materials policy and strategy issues for the UK : report of a meeting held at the British Library Research and Innovation Centre, London, 13 December 1996*. British Library Research and Innovation Report, 41. London: British Library Research and Innovation Centre, 1997.
<URL:http://www.ukoln.ac.uk/services/papers/bl/blri041/digpres.html>

[6] Bennett, J.C., *A framework of data types and formats, and issues affecting the long-term preservation of digital material*. British Library Research and Innovation Report, 50. London: British Library Research and Innovation Centre, 1997.
<URL:http://www.ukoln.ac.uk/services/papers/bl/jisc-npo50/bennet.html>

[7] Haynes, D., Streatfield, D., Jowett, T. and Blake, M., *Responsibility for digital archiving and long term access to digital data*. British Library Research and Innovation Report, 67. London: British Library Research and Innovation Centre, 1997.
<URL:http://www.ukoln.ac.uk/services/papers/bl/jisc-npo67/digital-preservation.html>

[8] Ross, S. and Gow, A., *Digital archaeology? Rescuing neglected or damaged digital collections*. British Library Research and Innovation Report, 108.

[9] Data Archive, *An Investigation into the digital preservation needs of universities and research funders*. British Library Research and Innovation Report, 109.

[10] Hendley, T., *Comparison of methods and costs of digital preservation*, British Library Research and Innovation Report, 106.

[11] Beagrie, N. and Greenstein, D., *Strategy for creating and preserving digital collections*. First public consultation and review draft. London: Arts and Humanities Data Service, 24 April 1998.
<URL:http://ahds.ac.uk/manage/framework.htm>

[12] Heery, R., Powell, A. and Day, M., *Metadata*. Library and Information Briefings, 75. London: South Bank University, Library Information Technology Centre, 1997.

[13] RLG Working Group on Preservation Issues of Metadata, *Preliminary report*. Mountain View, Calif.: Research Libraries Group, 7 January 1998.
<URL:http://www.rlg.org/preserv/presmeta.html>

[14] Rothenberg, J., Ensuring the longevity of digital documents. *Scientific American*, 272 (1), 1995, 24-29.

[15] Rothenberg, J., *Metadata to support data quality and longevity*. Proceedings of the 1st IEEE Metadata Conference, NOAA Complex, Silver Spring, Md., 16-18 April 1996.
<URL:http://www.computer.org/conferen/meta96/rothenberg_paper/ieee.data-quality.html>

[16] Task Force on the Archiving of Digital Information, *Preserving digital information: report of the Task Force on Archiving of Digital Information commissioned by the Commission on*

*Preservation and Access and the Research Libraries Group*. Washington, D.C.: Commission on Preservation and Access, 1996.
<URL:http://www.rlg.org/ArchTF/>

[17] Duff, W., Ensuring the preservation of reliable evidence: a research project funded by the NHPRC. *Archivaria*, 42, 1995, 28-45.

[18] Bearman, D., *Electronic evidence: strategies for managing records in contemporary organizations*. Pittsburgh, Penn.: Archives and Museum Informatics, 1994, p. 302.

[19] Graham, P.S., Long-term intellectual preservation. In *Digital imaging technology for preservation*, ed. N.E. Elkington, Mountain View, Calif.: Research Libraries Group, 1994, 41-57.

[20] Sollins, K. and Masinter, L., *Functional Requirements for Uniform Resource Names*. RFC 1737, 1994.
<URL:http://ds.internic.net/rfc/rfc1737.txt>

[21] Bide, M., *In search of the Unicorn: the Digital Object Identifier from a user perspective*, rev. ed. BNBRF Report 89. London: Book Industry Communication, 1998.
<URL:http://www.bic.org.uk/bic/unicorn2.pdf>

[22] BIBLINK project: <UTL:http://www.hosted.ukoln.ac.uk/biblink/>

[23] Bearman, D. and Sochats, K., Metadata requirements for evidence. Pittsburgh, Penn.: University of Pittsburgh, School of Information Science, 1996.
<URL:http://www.lis.pitt.edu/~nhprc/BACartic.html>

[24] Duranti, L. and MacNeil, H., The protection of the integrity of electronic records: an overview of the UBC-MAS Research Project. *Archivaria*, 42, 1995, 46-67.

[25] Hofman, P., Worsfold, E., Hiom, D., Day, M., and Oehler, A., *Specification for resource description methods: 2, Selection criteria for quality controlled information gateways*. DESIRE: Development of a European Service for Information on Research and Education, Deliverable 3.2 (2), May 1997.
<URL:http://www.ukoln.ac.uk/metadata/desire/quality/>

## 5. Acknowledgements

UKOLN is funded by the British Library Research and Innovation Centre, the Joint Information Services Committee of the UK Higher Education Funding councils, as well as by project funding from JISC's eLib Programme and the European Union. UKOLN also receives support from the University of Bath, where it is based.

**For more information on the CEDARS project, please contact:**

CEDARS Project Manager
Edward Boyle Library
University of Leeds
Leeds LS2 9JT

k.l.russell@leeds.ac.uk

<URL:http://www.curl.ac.uk/>

# Project VOCS (Voix de la Culture Suisse - the voice of Swiss culture)

Kurt Deggeller, Director of Memoriav

VOCS is a pilot project by MEMORIAV (Association for the preservation of the audiovisual heritage of Switzerland). The partners are the Radio Suisse Romande (French-speaking Swiss radio network) and the Swiss National Library. The goal of the VOCS project is to ensure the preservation and communication of audio-, written and visual documents concerning key figures in Swiss culture. The final phase will be to make the documents available in digitised form in co-operation with SIRANAU a system which can store, search, consult and handle sound recordings in digitised form as well as other information (text, technical data right images...)

## Memoriav

Preservation of audiovisual heritage (i.e., information in the form of photographs, recordings, films and videos) has developed into a task which ancient mythology would have entrusted to Hercules, if not Sisyphus! The reasons for the difficulty of this task are numerous: The materials on which the information was recorded were not made to last a long time, anymore than the devices enabling access to those sounds and moving pictures. Audiovisual documents are produced in large volumes by institutions and individuals of the most varied origin and goals. What's more, use of audiovisual archives occurs under the most varied guises. These include commercial purposes, scientific research, or pure pleasure. The legal situation is thus anything but clear. Audiovisual documents are, on the whole, barely understandable without some form of written information accompanying them. In the worst cases, serious misunderstandings can result.

Since 1992 a working group has been concerned with the preservation and diffusion of audiovisual cultural heritage in Switzerland. The group was formed by representatives of the most important public production sites and archives for audiovisual documents. An initial reckoning has made it clear that the existing institutions in Switzerland are incapable of doing justice to the tasks confronting them. The first project entrusted with the task of creating a new institution was a failure, not only because of the high costs, but also because of the traditional Swiss dislike for overly centralised structures.

Subsequent considerations led to the concept of a network, a less cost-intensive solution which also took Swiss fears of excessive centralisation into account. At the end of 1995, the association known as Memoriav was founded as the sponsor of this structure.

The group of founding members includes two important federal institutions with archiving tasks: The Federal Archives and the National Library. The producer side is represented by the public radio/TV enterprise SRG with its nine units which possesses an extremely large collection of audio and video documents. The other members include the National Supervisory Body for Radio and Television, the National Office for Communication. Also belonging to the original members were the Swiss Film Archive and the Swiss National Sound Archive, both of which are organised as private foundations.

The projects for which Memoriav plays a role in overseeing or financing the activities can be divided into three categories: "Urgent" measures, projects of the Board of Directors, and external projects which receive financial support.

The urgent measures began already in 1992 upon realising that certain materials would not survive a possibly long process toward a definitive solution to the problem of archiving audiovisual cultural heritage in Switzerland. A special fund provided by the federal government was employed for restoring and copying a large collection of historical radio programs recorded on lacquer discs which are highly susceptible to deterioration. Nitrate films were copied onto security film, and some collections dating from the dawn of photography were saved. By the end of 1999 Memoriav will have spent a total of 6.2 million Swiss Francs for these urgent measures.

The Board of Directors has two goals for its own projects: Firstly, it wants to try out new ways of preserving audiovisual documents and making them available to the public. Second, it wants to venture into those fields whose archiving situation was hardly known up to now. These include private radio and TV organisations or video production other than for television.

## The project VOCS

The project known as VOCS (Voix de la Culture Suisse-Voice of Swiss Culture) is positioned in a field which concentrates on preserving and disseminating culture-oriented information. The goal here is to complete a collection of materials provided by authors while still living or after their deaths to the Swiss Literary Archive in Berne (which is part of the National Library) by adding radio productions from the French-speaking part of Switzerland. Both categories of documents could then be provided to users in a more co-ordinated manner. Selection of audio documents was carried out with the assistance of experts in the field of literary history. Fortunately, this also testified to the extremely high research value of audio documents from radio archives which had been inaccessible up to now. To insure that supply of those documents will be more innovative, the project was connected with development of a modular mass storage system.

## The project SIRANAU

SIRANAU (Système Intégré Radiophonique pour l'Archivage Numérique AUdio) is born from the co-operation of the Radio Suisse Romande with the Swiss Federal Institute of Technology (EPFL Lausanne), the Hewlett Packard Company and the Swiss National Sound Archive (Lugano). The project has the financial backing of the Commission for Technology and Innovation (CTI) of the Swiss Confederation.

The objectives of the SIRANAU are following:

- Develop a prototype of a digital mass storage system by selecting and assembling in a modular view the best hardware and software components for this purpose,
- experiment different technical choices, for example the generation of three types of audio files (original linear WAV - or BWF - format, reduced MPEG L3 at 56 KHz for listening and strongly compressed MPEG 8 KHz for public distribution,
- Verify the feasibility of the integration between SIRANAU and the broadcast production systems, making possible the exchange of sound files,
- communicate with the already existing documentary databases (and not replace them), so as to keep the unicity of search in the whole archives.

The prototype of SIRANAU is now operative. It consists in a Unix HP 900 server, a Informix IUS database, and different storage media. At the day we have tested three categories of media: Fast disks (RAID), Slow disks (MOD) in a HP Surestore 600 FX and Tape system (DLT).

The digital library management software AMASS provides security features for the control of the consistency of audio files.

The integration of the existing databases (one for music and one for spoken word, based on the ID BASIS documentary system in Radio Suisse Romande and VTLS Library System in the National Library) is made by the insertion of an intermediate frontend application, with MS Transaction Server.

SIRANAU is still a prototype but it has yet proved to be a solution for the needs of the radio production as well as for those of the preservation and communication of the audio heritage.

# The First Steps in Creating Cultural Heritage Digital Resources in Bulgaria

Milena P. Dobreva

Institute of Mathematics and Informatics, Bulgarian Academy of Sciences

bl. 8, Acad. G. Bonchev St.

1113 Sofia, Bulgaria

e-mail: dobreva@math.acad.bg

fax: 00359-2-9713649

KEYWORDS cultural heritage, digitisation, national policy, Medievalist's workbench

ABSTRACT

The paper presents the current state of the creation of digitised collections of cultural heritage resources in Bulgaria. Two aspects of this process are discussed in detail:

1. the policy of establishing the first projects in the field,

2. the unsolved problems which require a quick solution.

The experience presented in this paper is characteristic for countries in transition and will be of interest for colleagues, launching projects in the field of digitising cultural heritage collections in similar cultural and/or economic environment.

## 1. Background

Bulgaria has a rich cultural heritage represented by monuments both of local and European importance. The main collections of the cultural heritage belong to the state and their maintenance is totally dependent on the state budget. Collections differ in their nature like everywhere in the world - they are buildings, frescoes, icons, paintings, manuscripts, instruments, vessels, coins, etc. In this paper we will concentrate above all on written monuments, such as epigraphic inscriptions and palaeographic materials (medieval and more recent manuscripts), for example, the Latin and Greek epigraphic inscriptions from Late Antiquity period, kept in Bulgarian repositories which form the third largest national collection in Europe after the collections in Italy and Greece. The repositories of Bulgarian libraries and museums house over 8,500 manuscripts, which are a major historical source casting light on Medieval South-East European literature and history.

One would expect that the development of a national policy for digitisation would be an easy task when most collections of the cultural heritage are state-owned. Unfortunately, there is no national strategy for digitisation of national cultural heritage collections, although there are specialists in Bulgaria who already have valuable experience in this field.

The difficult situation in the country is one of the reasons for the absence of a national strategy in the field of digitisation of the cultural heritage. During the current transition period the issues of preservation of cultural heritage collections have been neglected as the main concern of the state is an economic stabilisation. Since libraries and the museums, the

largest repositories of cultural heritage resources, are almost totally dependent on insufficient Government funding, difficult choices on allocation of resources for support of the current collections and traditional preservation have to be taken. The applications of new information technologies which would contribute to the preservation and study of the collections is considered luxury the budget cannot afford.

## 2. Bulgarian Pilot Projects

### *2.1. Participating organisations*

Five types of organisations are potentially interested in digitisation of cultural heritage: government bodies, repositories, research and/or educational institutions, companies and foundations. These organisations with a different profile have significantly different approaches in the field of digitisation of cultural heritage due to their different aims and needs.

2.1.1. *Government bodies* (the Ministry of culture) are entrusted with the supervision of such activities. A project on networking of the museums is currently underway, however it does not contain any official statement or plan for digitisation of cultural heritage collections in the wide sense.

2.1.2. *Repositories* (libraries and museums), which seem the most natural initiators of digitisation projects because of the close relationship between digitisation and preservation, are currently in the position of observers due to lack of funding on the one hand, and copyright issues for digital collections, on the other hand. The Union of Librarians and Information Services Officers produced in 1997 a National Program for the preservation of Library Collections, which was adopted by the Library Council at the Ministry of Culture. Unfortunately, this interesting program is adopted only formally, without any real work on its implementation into the practice.

2.1.3. *Research and/or educational institutions* are the most active initiators of digitisation projects in Bulgaria as centres of study of the cultural heritage and the impact which digitisation could have on:

- routine work
- potential for large-scale comparative studies
- application of new research methods.

2.1.4. *Companies* are interested in presenting sections of cultural heritage to the world which they believe will be easily realised on the market. Today it is rather difficult to establish customer interest. The Bulgarian market of such products is unsatisfactory. This is why their main market is abroad.

2.1.5. *Funding bodies* (foundations) supported practically all projects undertaken in the field of digitisation. However, the scale of their support cannot meet the real costs of serious digitisation projects.

### *2.2. Current work in the field*

The first initiatives in this field were launched by research institutions and companies under the conditions of an absence of a national strategy and funding for digitisation programmes. Libraries and museums basically provided access to their collections instead of launching their own programmes.

The work of specialists from research institutions is basically directed towards entering data on available resources. Actual work on digitisation has not yet been done on large-scale basis, because of the high costs of such projects. Amongst the projects describing available resources we could mention:

- The Repertorium of Old Bulgarian Literature (co-ordinated by the Institute of literature at the Bulgarian Academy of Sciences);
- The Corpus of Epigraphic Inscriptions in Greek and Latin (co-ordinated by the Institute of Mathematics and Informatics of the Bulgarian Academy of Sciences),
- The 'St Cyril and Methodius: Byzantium and Slavs in the 9$^{th}$ century AD' project (co-ordinated by the American University in Blagoevgrad) [Dobreva, Ivanov 98].

The basic work done by companies is oriented towards creating CD-ROMs (four CD-ROMs already exist, two of manuscripts from the National Library 'St. Cyril and Methodius', one of Macedonian coins and one of Bulgarian Iconography).

A most important problem in the field of digitisation is connected with the copyright on materials for digitisation.

Copyright issues in such a complex field have to be clearly defined for 2 different situations: a. when primarily sources are being digitised, and b. when publications of research of different specialists are being incorporated into the final product. The second case is very important and even more complicated than the first one, because scholarly annotations and commentaries are important components of any digitised collections. Since this work is done in teams, the contribution of each member of the team has to be clearly defined and protected.

## 3. Unresolved Problems

This situation has led to several important issues:

- The absence of a national strategy has resulted in the lack of co-ordination between separate local initiatives which usually do not contribute to each other.
- The work of separate teams in the same field has lead to the application of many different *ad hoc* solutions, instead of a search for a general *ad modum* strategy.
- There is a clear need for international co-operation in the fields of Slavic and Balkan Studies, because of the wide spread of primarily sources throughout the whole of Europe. The lack of a national strategy does not support such co-operation in spite of its importance for real large-scale comparative studies.

- Ambiguity of legal copyright issues has lead to serious problems in persuading researchers to share their knowledge in digitisation projects affecting the level of presentation of materials, and restricting depth of presentation.

- In order not only to present materials, but to support research work in the field of Ancient and Medieval Studies, all data should be properly organised and processed. It is not sufficient to have collections of digital images, or text corpora. The application of information technologies in the Humanities is complicated by the specificity of the models in the subject domain. If we consider the example of Medieval Slavic studies, a commonly accepted model of which elements should be included in a formal model of the subject domain of knowledge can not be decided. For this reason, the creation of a generally accepted model is more a wish than a reality even after having the experience of several projects and organising a wide scientific discussion [Birnbaum et al. 95]. The creation of a specialised workbench for Slavic Medieval studies which will be sufficiently flexible to support those views and materials, which serve the needs of the concrete specialist is one of the possible solutions for this problem. I would like to stress that existing workbenches for the study for example of Latin manuscripts [Calabretto, Rumpler 98], will not match the needs of specialists in Slavic studies because of the impressive variety in Medieval Slavic texts for which computer presentation is still a subject of wide discussion (see the papers presented at the International workshop on Text Variety modelling [Dobreva 98]).

## 4. Conclusions

The paper deals with the initial Bulgarian experience in the field of developing electronic resources in the presentation of cultural heritage. The first projects in the field fall into two categories: research and commercial.

Bulgarian specialists encountered problems related to:

- The lack of a national strategy and co-ordination amongst institutions in the field of digitisation;

- Copyright issues (both for primarily sources and results of their scientific examination);

- Difficulties with the setting up of adequate workbenches for specific research tasks like Medieval Slavic manuscript studies.

The solution of those problems will contribute significantly to the development of real digital resources in the field of cultural heritage which, its turn, will contribute to the processes of European integration. Probably the basic problem for countries in transition is whether they will be able to set-up their own programs and start work on them meeting the quality criteria of the European Union.

There are two strategies which can be followed under these conditions:

1. Waiting for better economic conditions and for guidance of more experienced countries in the field of cultural heritage digitisation. The dangers in this approach come from the poor conditions for the preservation of our cultural heritage.

2. Searching for *ad hoc* solutions, which will not lead to qualitative preservation of the whole cultural heritage, but at least will partially preserve it. The danger in this approach comes from the serious differences in the quality standards in the field of digitisation. Is it worth spending money on digitisation projects with insufficient budgets?

These decisions are very difficult. But they should be taken, and sooner the better.

## Acknowledgement

I would like to thank for the financial support of the Research Support Scheme of the Open Society Institute/International Higher Education Support Programme, which granted the project RSS No.: 1743/481/1997 entitled 'Cyril and Methodius and the Early Medieval Slavic World: Byzantium and the Slavs in the 9th century AD'. The work on the corpus of epigraphic inscriptions dating from the period of Late Antiquity was supported by project MU-O-06/96 of the National Science Fund in Bulgaria.

## References

[Birnbaum et al. 95] D.J. Birnbaum, A.T. Bojadzhiev, M.P. Dobreva, A.L. Miltenova (eds.), Proceedings of the first Int. Conference "Computer Processing of Medieval Slavic Manuscripts", Blagoevgrad, 24-28 July 1995, 336 pp.

[Calabretto, Rumpler 98] S. Calabretto, B. Rumpler, *Distributed Multimedia Workstation for Medieval Manuscripts*, In: F. Rowland and J.Smith (eds.), *Electronic Publishing'98: Towards the Information-Rich Society*, Proceedings of an ICCC/IFIP Conference, Budapest, Hungary, 20-22 April 1998, pp. 166-178.

[Dobreva 98] M. Dobreva (ed.) Text variety in the Witnesses of Medieval Texts: Study from Co-operative Writing Perspective, Int. workshop proceedings, Sofia, 1998.

[Dobreva, Ivanov 98] M. Dobreva, S. Ivanov, *Issues in Electronic Publishing on the Medieval Slavic and Byzantine World*, In: F. Rowland and J.Smith (eds.), *Electronic Publishing'98: Towards the Information-Rich Society*, Proceedings of an ICCC/IFIP Conference, Budapest, Hungary, 20-22 April 1998, pp. 55-64.

[NPPLC 97] National Program for the Preservation of Library Collections, Sofia, 1997.

# VIDION - An On-Line Archive for Video

Paula Viana[1,2] , Ulisses Silva[1]
{pviana,ausilva}@inescn.pt

1 - Instituto de Engenharia de Sistemas e Computadores
Praça da República, 93 R/C
4000 Porto
Portugal

2 - Instituto Superior de Engenharia do Porto
Rua de S. Tomé
4200 Porto
Portugal

## Abstract

One of the problems people face nowadays in most of the jobs is the big amount of information that is produced and should be stored, accessed and re-used later. This problem is even more important when talking about a Television Broadcaster, as this kind of information has strong demands on storage capacity, communication bandwidth, classification data, etc.

This paper presents an experience developed in order to preserve a broadcaster audio-visual archive, the architecture of the system being tested and the goals already achieved.

## Introduction

VIDION (Digital Video On-Line) is a Portuguese R&D project involving RTP (the Portuguese public TV broadcaster), INESC (a research institute with some experience in the area of digital television) and Europarque (responsible for a science centre park).

RTP owns one of the largest audio-visual archives in Portugal with more than 400 000 documents in different formats (analogue, digital and analogue with special characteristics from the historical archive) amounting more than 300 000 hours.

VIDION main goal consists in proposing a strategy for the evolution towards the digital domain of the complete RTP audio-visual archive. The project will develop a small-scale prototype archive to be used by the News Service, consisting basically in two different kinds of servers - a broadcaster and a browser quality servers - workstations for searching, selecting and previewing video sequences, digitising stations and the required communication infrastructure.

Additionally some extra funcionalities will be developed: a restoration module to automatically correct video impairments and a text-based intelligent search engine to assist the indexing and search processes.

One of the main concerns of the project was on choosing standard video formats, long last storage devices and on finding solutions for easy and efficient access to the information. These three aspects will allow the real preservation of this valuable archive.

## System Architecture

The system developed within VIDION is composed by a number of elements as shown in Figure 1.

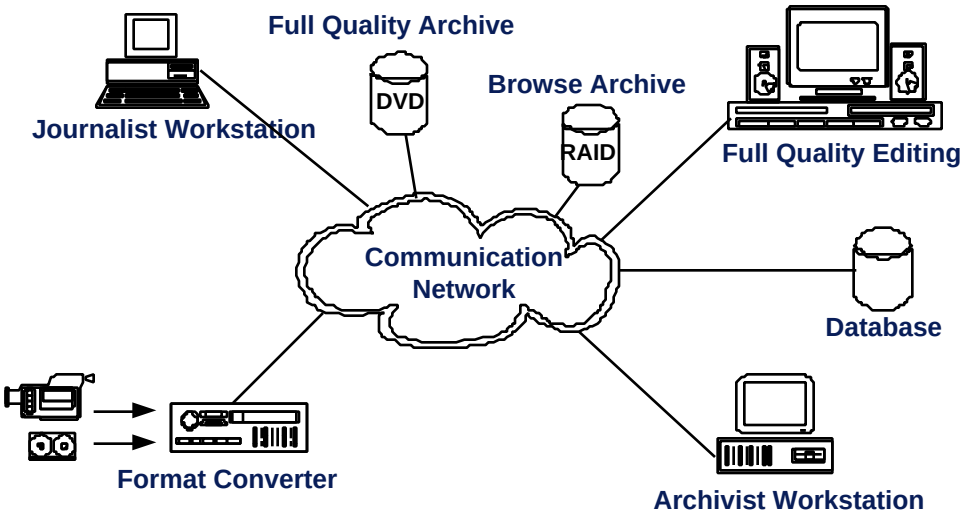



**Figure 1 - VIDION Architecture for Digital Studio Archive**

The first step on building a digital archive is the digitalisation block and format converter which will enable the creation of two versions, with different dimensions and quality, of the same source. The browse version will be stored in a RAID based archive and after being conveniently documented by the archivist, through a set of indexing GUI and auxiliary tools, will be available for searching, previewing, selecting and pre-editing in the journalist workstation. This will enable the journalist to produce a draft version of a video sequence which will later be produced, based on automatically generated information, in studio/broadcast quality format. The full quality archive, due to the big amount of storage space it represents and to the longevity required, will most probably be based on a DVD based server.

One of the points that should be considered in a system like the one described is the format of the information to preserve. Standard solutions are needed in order to be able to use different kind of equipments from different vendors and to interchange information between similar systems. Based on these ideas, MPEG1 at around 1.5Mbps was chosen for the browse server while MPEG2, the format which is expected to be used in the broadcast world, will be available in the full quality server. Due to the special characteristics and cultural importance of the historical RTP material and also due to the increasing demand for video material from this archive special interest was put in an automatic digital restoration processes.

## Digital Video Restoration Module

The major advantage of the digital restoration methods over the traditional ones is that thousands of techniques can be tested without damaging the original copy, therefore allowing the restoration operator to choose the technique that gives the best perceptual and/or objective results.

The main problem with digital video restoration is that it is impossible to find a technique that removes all kind of artefacts from the video. Based on RTP experience, blotches and line scratches (vertical lines) were considered the two most frequent and annoying artefacts in the archive and so the algorithms developed must try to remove them from the original video.

The restoration process can be viewed as a black box which includes the artefact detector, restoration algorithms (interpolator) and a quality measurement process (Figure 2). The input and output of the restoration chain is a CCIR 601 digital video signal, in order to avoid the extra noise resulting from further coding schemes as MPEG2.

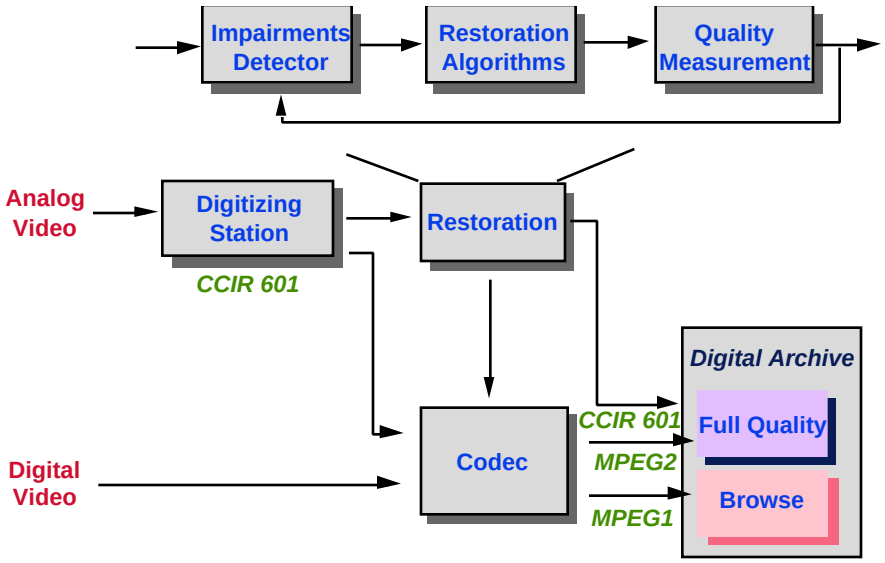



**Figure 2 - Digitising and Restoration Modules**

The impairments detector, which can be manual, automatic or semi-automatic, analyses the original sequence and marks the pixels in the image that are degraded or corrupted. Blotches are, by definition, features that didn't exist in the original image and that affect just one frame of the sequence, meaning that there is no matching feature in the previous or next frame. According to this definition, an heuristic that tries to find regions that, after motion compensation, show no correlation with the previous and next frames was implemented. A modification of this heuristic is used for detecting line scratches because, in some cases, this artefact stays in the same position from frame to frame so we could have degraded pixels in the same position in two consecutive frames. To eliminate this problem instead of marking the degraded pixels in the entire sequence and then using the restoration algorithms, the restoration is made immediately after the detection.

The restoration block looks at the marked pixels and, after processing them, generates a restored sequence. A multilevel spatio-temporal median filtering, which can restore large degraded areas of the image with little processing time, is used for blotches correction while for line scratches a very simple spatial weighted filter to correct the entire degraded column is in use. Figure 3 shows the results obtained in the correction of a video sequence containing blotches.

The quality measurement block judges the final quality of the output sequence using mainly perceptual distortion metrics as a distortion free sequence to compare with the restored one is not available. A simple automatic perceptual quality measurement algorithm was implemented although allowing the human operating the system to have the ultimate decision regarding the good or bad restoration results of a specific video sequence.

The implementation of further algorithms doesn't interfere with the existing ones providing a continuous upgrade of the restoration chain. While the original copies are always preserved in its original support, multiple copies of the restored material can be made without any risk of damaging the original ones.

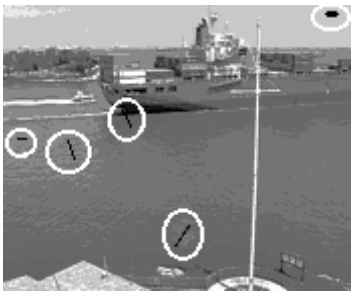


**Figure 3 - a) Degraded image containing blotches b) Restored image**

## Conclusions

The evolution of video archives towards the digital domain are still in the childhood. However the recent developments and the interest put by the scientific community in this area will make greater advances possible allowing the construction of functional systems. The tests and achievements already obtained in VIDION enabled the definition of an architecture which will help on the preservation of an important assets of a TV broadcaster and that can be extended to similar systems in different application areas.

# *Nature* - A Prototype Digital Archive

Ross MacIntyre[1], Simon Tanner[2].

[1]***Manchester Computing, University of Manchester, UK, email:r.macintyre@mcc.ac.uk*** [2] ***Higher Education Digitisation Service (HEDS), University of Hertfordshire, UK, email:s.g.tanner@herts.ac.uk***

## Abstract

In response to the proposal to digitise the journal ***Nature*** (1869->1992), published by Macmillan, a pilot project was commissioned to discover the technical issues and ascertain costs.

The initial conversion and digitisation elements are being provided by the Higher Education Digitisation Service, University of Hertfordshire, (HEDS) whilst Manchester Computing, University of Manchester, is providing all the data management and access elements.

This report details the processes so far undertaken, the results ascertained from these pilot processes and the techniques used. The pilot, though not complete, is now at the stage where certain conclusions can be drawn from the progress made so far. This is also the point at which further progress requires certain decisions regarding format and techniques to be focused and validated.

## 1 The Journal *Nature*

***Nature,*** was first published in 1869 and has been a weekly science journal to the current day. The journal's objectives were stated in the first issue by founding editor Norman Lockyer, who remained at the helm for fifty years: to bring the accomplishments of science to the "general public". ***Nature*** became the international journal of first choice for the presentation of original results and discoveries during the inter-war years, coinciding with the change of editor to Richard Gregory. ***Nature*** remains one of the most widely cited interdisciplinary science journals in the world today.

The establishment of a Higher Education archive would provide the following:

- A valuable research tool
- An aid to teaching - Macmillan advise that they receive many requests for access to papers from ***Nature*** for teaching purposes and making these available electronically would make such access more cost effective
- Provide students and researchers with a vast backstore of source material
- A unique view of scientific history and would therefore also be an aid to historians and sociologists of science.

## 2 Pilot Objectives

The ***Nature*** digitisation pilot project has the following objectives as expressed in the initial proposal:

- To test the digitisation of historical scientific material with textual, graphical, photographic and formulaic content. This content being expressed in many fonts, styles and standards in changing paper formats.

- To test whether the source material can be digitised, accessed and archived to standards sufficient to support teaching and research with in Higher Education in the UK.
- To compare various types of search technology on the source material.
- To test possible delivery mechanisms.
- To ascertain the costs and techniques required to meet the above objectives.

## 3 Method

The basic scan specification used was:

| Page Format | Tone / Colour | Resolutions |
|---|---|---|
| Text only. | Black and White | 400 & 600 DPI |
| Text with line drawing or bi-tonal graphic. | Black and White | 400 & 600 DPI |
| Pages with photographic or any half tone or other greyscale graphical content. | 256 Greyscales | 400 & 600 DPI |
| Pages with colour content | 24 bit colour | 400 & 600 DPI |
| | 256 Greyscales | 400 & 600 DPI |

Each page to be scanned was assessed for its content and the specification above applied. For all the 400 DPI samples an “image and text” PDF plus a text file was created and delivered to Manchester Computing. The 600 DPI samples were processed to these formats to discover the process but not shipped to Manchester Computing. The process was measured, refined and measured again to gain production metrics.

Following discussions HEDS has taken several pages through the processes to create downsampled “image and text” PDF’s.

HEDS has also visited the Macmillans Publisher’s offices to assay the complete collection of ***Nature*** to ascertain an idea of the numbers of pages in total for the project and the proportions of the various pages types identified in the table above.

Why "image and text" PDF? This format gives many of the benefits of PDF conversion with the main difference from full PDF being that it is viewed on the screen as a bitmap image of the original. All the features of text search and retrieval are available, but hidden from immediate review. The TIFF images are also to be retained for archiving and used again if a different technology for viewing/searching is selected at a later date. Other approaches using page image rendition, e.g. JSTOR [1], Internet Library of Early Journals [2], store the OCRd text in separate, searchable files, but there is no indication on the image itself of hit terms.

## 4 Processes and Techniques

A number of processes and techniques were used to produce the samples, this section details these and the reasons for following certain routes. See Appendix A for Production Process Maps.

### *4.1 Scanning*

***Nature*** is complex in the density of its layout and the font sizes used over the lifetime of the publication. In more recent years the content has become more graphical, with heavy use of photographs and colour images. This is especially challenging in the genetic

subject areas where the gradations of tone used in graphic representations are critical to the scientific accuracy of the material. There is also a period of years in the 1970's and early 1980's (not fully assayed yet) where the paper of the publication is thin and there is significant show through.

As ***Nature*** has been presented in bound volumes for scanning, HEDS has utilised its Zeutschel Omniscan 3000 Bookscanner to gain the best possible output from the journal with minimum degradation to the originals. HEDS are constrained by the technology available in the Bookscanner market. HEDS were the first in Europe to receive the Zeutschel greyscale scanning capability and fully expect to be among the first to gain 600 DPI capability in the Bookscanner. Currently, for the 600 DPI or colour scans, HEDS has had to flatbed the ***Nature*** samples provided, with the additional time premium and attendant degradation on the originals.

***Nature*** is also tightly bound and the pages are slightly larger than A4 with rather small margins. This results in non standard scan image file dimensions that have to be taken into account in all other software processes to ensure that the whole page content is retained in the PDF file. It also means that the content of the originals may appear to run very close to the gutter of the bound volume with curvature in the paper. This must be resolved in the scan process to get the best image from the page as possible. This is possible with the handling capabilities available from Bookscanner technology with the use of a book cradle, but not so easy to achieve with flatbed processes.

The time and cost of scanning are distinctly affected by the standard to which the image is being reproduced. There is a marked increase in file size and attendant scan time for increases in resolution. These file sizes are increased again by any increase in tonality from black and white to greyscale and to colour. This is demonstrated in the table below:

***Average TIFF file sizes for various resolutions and tones / colours***

| Tone / Colour | 400 DPI Resolution | 600 DPI Resolution |
|---|---|---|
| Black and White (bi-tonal) | 1,300 Kb | 3,400 Kb |
| Greyscale (256 tones) | 11,000 Kb | 20,000 Kb |
| Colour (24 bit) | 15,000 Kb | 30,000 Kb |

The file size detrimentally affects the speed of writing and retrieval of the file to and from disk for any processing that is required from the scan stage onwards.

### *4.2 Post Processing*

As there is some skew in many of the image files created, the need for post-processing the image increased. This includes deskewing the image, removing some dirt or speckling from the image, and output of the image file in the correct format for further processes. The despeckling or dirt removal has to be managed with care to ensure that the process does not affect the content of the text or graphical elements of the original content.

The deskewing process when automated is very quick and effective but the software can have some occasional problems dealing with certain types of page formats. ***Nature*** pages have a horizontal line across the top of each page with the text arranged in columns below. Where the alignment of the horizontal line with the text is not at a true 90 degree angle then the post processing is quite likely to align the page with the horizontal line and thereby introduce a skew to the page image. Also where there are vertical lines or dirt in vertical lines in the image then this could also introduce skew into the image file.

HEDS are also constrained by the technology available for post processing greyscale and colour image files. These types of files cannot be automatically processed to remove skew etc. by the top products in the market, ScanFix and PixEdit. Both of these software tools are developing greyscale capabilities, but there is no timescale for availability as yet. Any deskewing of greyscale or colour images has to be done manually at the scan

stage and this adds an additional time element to the process. The manual deskew available within the scan engines is quite basic and not as effective as can be achieved with tools such as ScanFix. This means that some colour and greyscale image files will retain some small level of skew in a production process.

### *4.3 "Image and Text" PDF Production*

The conversion of the TIFF files into "image and text" PDF format is being done using the Adobe Capture 2.x conversion tool. The Adobe PDFWriter module is set at 600 DPI with no compression or downsampling selected to achieve the maximum resource representation into the PDF file. It is not possible to independently test with confidence what the exact resolution is within the PDF file, only the means by which it was created. The Capture 2.x engine is slow at converting files into PDF format, sometimes as bad as 10 minutes machine time per page image converted. However, when compared with the quicker production times possible using Capture 1.x engine, the Capture 2.x engine is far more reliable, and the output requires lower levels of QA. Another issue with Capture 2.x is that the engine has a hardware dongle that effectively adds an additional cost per page converted.

The Adobe Capture process also adds complexity to the production process for image files which are 600 DPI greyscale or colour. Whilst Capture can process 600 DPI bi-tonal files, it is not capable of converting greyscale or colour files sourced at above 400 DPI. Therefore, if HEDS creates 600 DPI originals they have to be converted to 400 DPI prior to conversion into PDF format. This does not involve downsampling the original file, but merely changing the TIFF header information through a batch save process using a product such as Paint Shop Pro 4.x. The processing of such large files through Capture requires a very large amount of memory to be available, approximately 5 times the image file size being converted. As the image file sizes average between 20Mb and 30Mb this places a big overhead on the machine processing the image files and also slows the processing of colour and greyscale 600 DPI images relative to the other samples completed to date.

***Average PDF file sizes for various conversion processes:***

| Tone / Colour | Average PDF at 600DPI conversion. | Downsampled size - minimum PDF |
|---|---|---|
| Black and White (bi-tonal) | 340 Kb | 222 Kb |
| Greyscale (256 tones) | 2,225 Kb | 249 Kb |
| Colour (24 bit) | 4,200 Kb | 851 Kb |

The downsampling process in Adobe Capture means setting lower resolutions and maximum compression for the output PDF file. These changes mean that information content is being lost in the conversion. The result is that picture elements still look reasonable, but that the text appears blurred. This may possibly be due to interpolation. To achieve downsampled PDF files would require the Adobe Capture process to be repeated in its entirety for all PDF's, doubling the costs of this portion of the process.

### *4.4 ASCII Text Production*

The ASCII text from each page is required to assist Manchester Computing create search resources to find individual pages from ***Nature*** and validate the results. This file is being created at the same point at which the "image and text" PDF is being written, using the OCR'd content from the PDF file. The results show a higher level of accuracy in the OCR than expected, so that the indexes will be a richer search tool.

### *4.5 Preparation and Quality Checks*

There are a number of preparation functions that need to be completed before the originals can be converted. These include marking all advertising pages to ensure they are not scanned, checking for colour prints and marking them for different processing, setting up a production log and data entry of document pages and data structures. Obviously, there is also the set up time for each of the production processes to ensure the machinery is at the optimum setting for the originals to be converted or data processed.

There are a number of points in the process where there are basic checks made to ensure the quality of the output. These are to ensure that every page has been scanned and that all the pages are in the correct order. There are further checks on the content of every TIFF file output to ensure that the content is representative of the original and that the correct file name has been assigned to the TIFF image. Similar checks to the file name and content are carried out for the PDF and text files. The fact that a single page creates three separate single output files adds to the cost of quality assurance due to the total volume of files to be controlled and checked against every page scanned.

## 5 Assay of *Nature* Collection at Macmillan

HEDS have done a survey of one issue of ***Nature*** per year of the publication from 1869 to 1992 to ascertain the proportions of greyscale and colour format pages in the whole collection. From such a survey it has been possible to gain an estimate of the total number of such pages in the collection and thus estimate with higher accuracy the total costs of digitising the whole collection. The results of the survey are presented in graphical form in Appendix B.

(It is interesting to note that the complete set held in the Editor's office had been written on, in ink, over most text passages for the majority of the publication run, making them unsuitable for scanning due to the obscuring of the text.)

- There are an estimated **298,950 pages** in the total production run of ***Nature*** from 1869 - 1992. This figure was found by taking the number of pages per issue per year sampled and multiplying that by 52 for each year and then adding up the results. This figure has been cross checked by dividing the publication into blocks according to the chronological changes in design and layout of ***Nature*** and then averaging the number of pages per issue per year across each design change. Then by adding up the results, a figure of 298,600 pages is found which cross checks favourably with the above figure.

- That the proportions of black and white, greyscale and colour pages are:

| Tone / Colour | Percentage of total | Total no. of pages |
|---|---|---|
| Black and White (bi-tonal) | 87.6% | 262,028 |
| Greyscale (256 tones) | 11.8% | 35,204 |
| Colour (24 bit) | 0.6% | 1,196 |

## 6 Application Development

### *6.1 Outline Application Specification*

The following documents the prototype's specification drafted by Prof. David Pullinger from Macmillan. The sections that follow subsequently describe the application development undertaken and some planned future directions.

Home Page :

Paragraph explaining the contents and pilot project.
Invitation to fill in feedback form

Three routes to articles in the archive
Link to ***Nature***'s website

Navigation by introductory pieces and indexes:

Introductory piece explaining what has been in ***Nature*** and its value and whom it might interest
Links to focus area
Each focus area has introductory paragraph explaining the interest of this section and a linked index of articles

Navigation by tables of contents:

When each article is scanned, the whole issue is done at the same time.
Agreement on header information will lead to the automatic construction of table of contents.

Navigation by search:

Search by text
Search by bibliographic citation

### *6.2 Technical Development*

A prototype application has been created, available on the WWW, with access restricted by password. It is intended to replace this with IP address checking should the prototype be more widely available. This has been loaded with the selected issues from ***Nature***, provided by HEDS.

The application infrastructure was developed as part of another research project and was not specifically designed for this purpose. Some tailoring has been necessary, in particular, the use of objectbase management system software added a level of complexity to the archive which is unnecessary. The data is highly ordered, its hierarchical structure being reflected explicitly in the directory structures defined. The files are assigned unique, self-identifying, names. In addition to the PDF files, the header data is held in consistently named files in SGML which is dynamically converted to HTML for display using (OmniMark[3]) scripts.

### *6.3 Data Loading Approach*

The application currently contains two versions of digitised ***Nature***:

1) page-by-page viewing

2) article-by-article viewing

Both offer 'traditional' hierarchical browsing, i.e. Year->Issue->Table of Contents-> Header and/or Article.

The data as it arrives is loaded into the 'page-by-page' viewing interface, remembering that the unit of digitisation is one file per page, so nothing else is known about the contents at this stage.

As metadata is defined and received, and the PDFs are combined (see 6.5 below), the data is then accessible by the second viewing interface. This gets content on-line as soon as possible, following digitisation, removing the existence of metadata from the critical path. It is recognised that the content at that stage is cumbersome to navigate, but is accessible via searching. The creation of the metadata is, however, vital for sensible browsing.

### *6. 4 Metadata*

The creation of metadata has been performed as a 3-pass operation:

1st pass - automatic creation of page-level data

The aim is to serve the digitised content at the earliest opportunity. When individual pages are received, a header record is created in order to load the file into the application.

Initially this contains just 'standing data', e.g. journal name, publisher, ISSN, plus the page number. The header data is loaded and also inserted into the PDF files for consistency.

2nd pass - manual creation of (minimal) article-level metadata.

Article metadata files have been created by Manchester Computing for an initial, small set of issues. However, an external 'keying agency', Saztec Europe Ltd., has been appointed to create and validate further header records. Their prime objective is to identify the editorial contents of each issue at an acceptable level. This could be viewed as little more than the re-keying of the table of contents for each issue, though it should be noted that the data present in the Table of Contents is not actually sufficient. The data is again inserted into the PDF files, for consistency and to improve the presentation of search results.

3rd pass - on-going improvement of archive based on experience and feedback received.

The Archive will present numerous possibilities for subsequent cataloguing. The amount and type of work undertaken could be the subject of separately funded initiatives outside the scope of the pilot. The infrastructure within the application should support such embellishment.

### *6.5 Minimal Metadata Defined*

Manchester Computing have defined a minimal set of metadata for the archive and created an SGML DTD to reflect.

Dublin Core [4] Tagged Items:1) * <TITLE scheme="Internal"> Article Title </TITLE>2) * <CREATOR scheme="Internal">
<FNMS> AuthorForename(s) </FNMS>
<SNM> AuthorSurname </SNM>
<SFX> PostNomial </SFX> <AFF> Affiliation </AFF>
</CREATOR>
3) <SUBJECT> n/a for ***Nature*** </SUBJECT>4) * <DESCRIPTION scheme="Internal"> Description </DESCRIPTION>5) <PUBLISHER scheme="Internal"> Macmillan </PUBLISHER>
6) <CONTRIBUTOR> n/a for ***Nature*** </CONTRIBUTOR>7) * <DATE scheme="ISO 8601"> Cover Date YYYY-MM-DD </DATE>8) <TYPE scheme="DCObjects"> "Article" </TYPE> * <TYPE scheme="Internal"> TypeOfContent </TYPE>
9) <FORMAT scheme="IMT"> "application/pdf" </FORMAT>
10) <IDENTIFIER scheme="SICI"> SICI </IDENTIFIER> <IDENTIFIER scheme="Internal"> PhysicalFileIdentifier </IDENTIFIER>
11) * <SOURCE scheme="Internal"> <JTL> JournalTitle </JTL> <VID> Volume </VID> <IID> Issue </IID> <PPF> StartPage </PPF>
<PPL> EndPage </PPL>
</SOURCE>
12) <LANGUAGE scheme="ISO 639"> "EN" </LANGUAGE>
13) <RELATION scheme="ISSN" relation="IsPartOf"> ISSN </RELATION>14) <COVERAGE> n/a for ***Nature*** </COVERAGE>15) <RIGHTS scheme="Freetext"> Copyright String</RIGHTS>Note that additional labels have been introduced within the SOURCE and CREATOR elements for clarity, though they would not necessarily appear explicitly in future instances of the data, e.g. as HTML meta elements.
* Items 1, 2, 4, 7, 8 and 11 would need to be manually created and passed to Manchester Computing. Item 8, internal scheme, would be the appropriate one from the types of contribution ***Nature*** publishes: Articles, Letters, Review Articles, Progress Articles, Scientific Correspondence, News & Views, and Supplementary Information, etc. The remaining items are either not applicable or can be generated or derived by Manchester Computing.

Example: corresponding HTML v4.0, simple, unstructured metadata elements:

1) <meta name = "DC.Title" content = "The Comet">2) <meta name = "DC.Creator" content = "Hind,J.R., FRS, Greenwich Observatory">

3) n/a
4) <meta name = "DC.Description"
content = "The Comet by Prof.J.Brocklehurst, University College, Dublin. ">5) <meta name = "DC.Publisher" content = "Macmillan Publishers Ltd, Crinan St, London">6) n/a7) <meta name = "DC.Date" scheme = "ISO 8601" content = "1874-06-25">8) <meta name = "DC.Type" scheme = "DCObjects" content = "Article">
<meta name = "DC.Type" scheme = "Internal" content = "Book Review">
9) <meta name = "DC.Format" scheme = "MIME" content = "application/pdf">10) <meta name = "DC.Identifier" scheme = "SICI" content = "0028-0836(18740625)10:243">11) <meta name = "DC.Source" content = "***Nature*** 10-243 pp149-150">
12) <meta name = "DC.Language" scheme = "ISO 639" content = "EN">13) <meta name = "DC.Relation.IsPartOf" scheme = "ISSN" content = "0028-0836">14) n/a15) <meta name = "DC.Rights" content = "Macmillan Publishers Ltd. 1874">

The PDF files are combined, via Acrobat Exchange, based on the metadata created during the '2nd pass'. There is 1 file per distinct start page. So, If article 1 runs from p2-p3, article 2 is only on p3 and article 3 runs from p3-p4, two files are created 2 files, 1 containing p2 & p3 PDFs and the other p3 and p4 PDFs.

The reason for combining the physical files include: likelihood of retrieval of subsequent page(s); availability of byteserving, so pages are downloaded only as required or in the background; logical consistency with metadata; averts problems printing and obtaining 'next page'.

The combination has been done manually so far, but a command line batch process will be developed.

The full classification of ***Nature***'s contents has not been undertaken. Classifications noted include: Letters to the Editor, News & Views, Book Reviews, On Our Bookshelf. Macmillan will help better classify over time.

The metadata being proposed has been discussed with TASI, the Technical Advisory Service for Imaging[5]. They are a recently established, JISC-funded body who help support image-based development projects involving the UK Higher Education community.

### *6.6 Searching*

Both a bibliographic and a free text search capability have been implemented, using Verity' s Search97 Information Server software[6]. It is refreshing to report that the software worked as expected, including 'hit-term highlighting'. This means that even though an image is being displayed on screen, the term(s) searched for are highlighted, due to the presence of the text within the PDF file. The facility requires later versions of web browsers and the Acrobat Reader v3 plug-in. There have been some browser version specific patches applied to the software on the server, but all have worked successfully.

Software Scientific [7] have code that given a set of text documents, will generate a 'topic tree'. This can be used in conjunction with Verity to assist in searching. Effectively the search for 'like' terms is biased based on the meanings of the words in context, i.e. terminology actually used in the documents. This appears to help with problems associated with terminology changing over time. It could possibly be used in time-slices, as appropriate.

### *6.7 Themed Access - 'Nature Trails'*

The Electronic Publishing Research Group at Nottingham University[8] and the Multimedia Research Group at University of Southampton[9], have agreed to a formal collaboration in support of the 'themed' access required. Southampton's DLS software (available commercially from Multicosm[10] ) will be used to establish links within the archive, as defined at the outset and also to support the subsequent definition of linkbases in support of teaching. The dynamic generation and insertion of links into the PDF files

has already been demonstrated at Southampton using ***Nature*** test files. The technology was the subject of a recent paper presented at EP98[11].

Coincidentally, Software Scientific also offer code to assist in the development of linkbases and identify 'themes' across collections of documents. It is intended to explore use of their software in conjunction with DLS.

# 7 Observations

### *7.1 The originals should be stripped and tested to gain production metrics.*

Stripping enables the optimisation of the scanning phase, eases preparation, handling and reduces overall costs substantially. Stripping also allows for a wider choice of scan equipment and supplier of scan services to be considered. An acquisition cost may be incurred, but this can be offset against the savings in the cost of processing.

### *7.2 400 versus 600 DPI*

600 DPI has been established as a standard for full archive scanning of black and white text in the USA on projects such as JSTOR and at Cornell University. The reason for Cornell and JSTOR recommending 600 DPI is that in bi-tonal scanning there is a risk of losing some data at lower resolutions.

The main issue that drives the resolution requirements for ***Nature*** is related to the information content of the resource. The resolution of 400 DPI will represent all the textual information in the journal. 600DPI would add a level of detail that would not add to the readable content but would add a small amount of character edge smoothing. This effect in the TIFF files is negligible from the perspective of the end user of the material. Please also note that the end user will only ever view the PDF's which will not be at a measurable 600 DPI whatever the input source file used.

[Aside: HEDS is using a Bookscanner with optical technology specifically for 400DPI or iterations of 400DPI. Using the HEDS equipment to gain 600DPI images would incur interpolation in the characters which would actually be less faithful to the original than a lower resolution. Unless HEDS can strip the bound volumes of ***Nature,*** the best standard of image will be obtained by using a resolution of 400DPI, to optimise the Bookscanners optical characteristics.]

Therefore, the only remaining reason for 600DPI must be to futureproof the TIFF files for potential uses not defined at this time. It is doubtful that increasing resolution alone, when not needed for the immediate application, will ever futureproof image files. TIFF file standards and the 400 or 600 DPI standards have a maximum shelf life of about 7-10 years and it is likely after this period that two outcomes will have occurred. First, the images will be deemed of too low resolution whatever choice is made for 400 or 600 DPI at this time, creating a potential rescan requirement. Or, the other possibility is that technology will develop such that for OCR or other post scanning processes, resolution becomes a non-issue due to the fuzzy nature of the technology. Thus, the choice of 600 DPI for archive is a belt and braces approach to archiving but with a shelf life of 10 years maximum. Therefore, the decision is whether the additional cost is warranted for the security of the next 10 years.

### *7.3 600 DPI would be unsuitable and unnecessary for the colour and greyscale pages.*

There are no fixed standards for colour or greyscale images, but discussions by HEDS with Anne Kenney, Associate Director of the Department of Preservation and Conservation at Cornell University has derived a recommendation that any higher resolution than 400 DPI for colour or greyscale would not add any further content to the scanned images. Therefore, whatever the resolution chosen for the black and white text pages of ***Nature***, 400 DPI should be used for greyscale and colour type pages.

### *7.4 Further experimentation to reduce PDF file sizes is required.*

The size of the PDF files can be onerous, e.g. 150Kb->3Mb per page, so avenues to reduce should be explored. Though the archive could be digitised to produce 'normal' PDF, the cost associated with OCR correction and the almost absolute certainty of error, have ruled this out as an option. Nevertheless, perhaps certain PDF files could be OCR corrected. Candidates would be all articles included in a '***Nature*** Trail'.

Experiments with downsampling so far have been disappointing, giving a blurred appearance to the text and this would be unsuitable as the main means of access to the information in ***Nature***, but further work is recommended.

## References

[1] JSTOR - http://www.jstor.org

[2] ILEJ - http://www.bodley.ox.ac.uk/ilej/

[3] OmniMark - http://www.omnimark.com

[4] The tag labels are from "Dublin Core Element Set: Reference Description" http://purl.oclc.org/metadata/dublin_core (last updated 2/11/97).

[5] TASI - http://www.tasi.ac.uk

[6] Verity - http://www.verity.com

[7] Software Scientific - http://ourworld.compuserve.com/homepages/swsci

[8] The Electronic Publishing Research Group at Nottingham University - http://www.ep.cs.nott.ac.uk

[9] The Multimedia Research Group at University of Southampton - http://www.mmrg.ecs.soton.ac.uk

[10] Multicosm - http://www.multicosm.com

[11] S.Probets, D.F.Brailsford, L.Carr and W.Hall, ***Dynamic Link Inclusion in Online PDF Journals,*** EP98, International Conference on Electronic Publishing, Document Manipulation and Typography, April 1998, Saint-Malo, France. (http://www.ep.cs.nott.ac.uk/~sgp/ep98.pdf)

## Acknowledgements

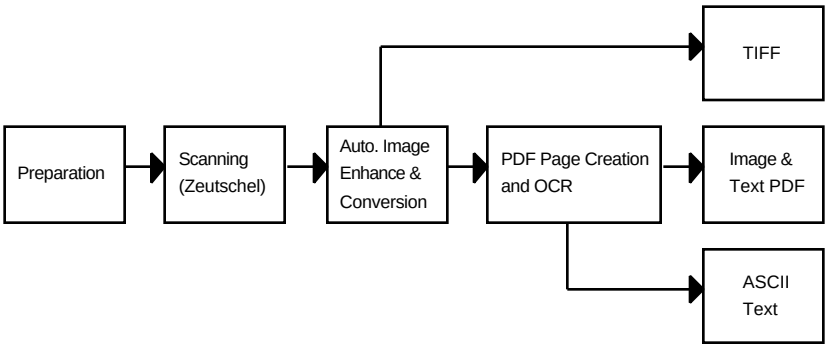
Prof.D.Pullinger, Macmillan Publishers Ltd, provided inspiration and guidance, and continues to do so.

This pilot project is funded by the Joint Information Systems Committee (JISC) of the Higher Education Funding Councils.

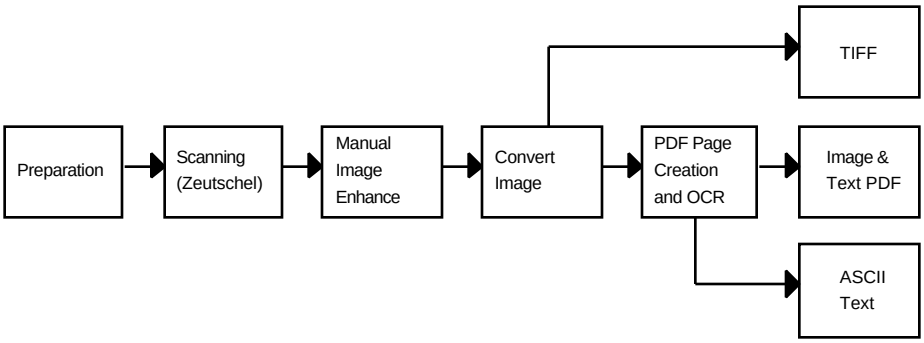
## Appendix A Production Process Maps

The following are graphic descriptions of the production process for each of the types of sample completed. They show functional and output file paths. Please assume quality checks throughout the process and before the final delivery of the output files, not shown here to keep the graphics simple.

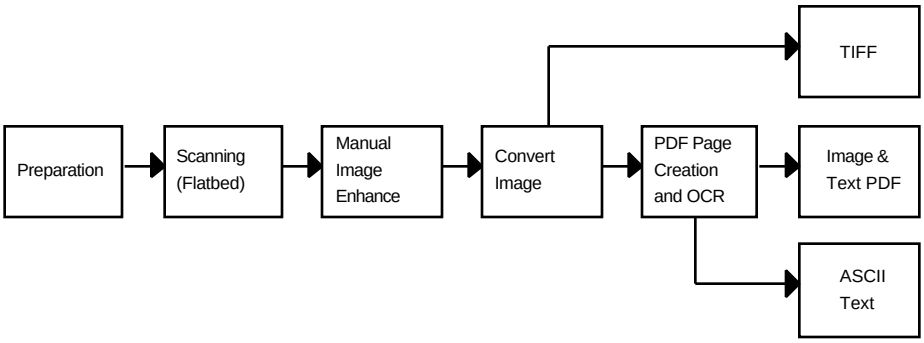
***400 DPI Black and White Pages***



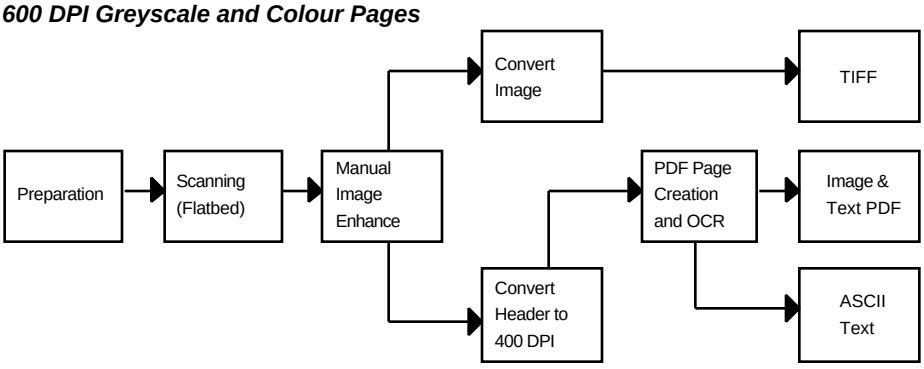

***400 DPI Greyscale Pages***



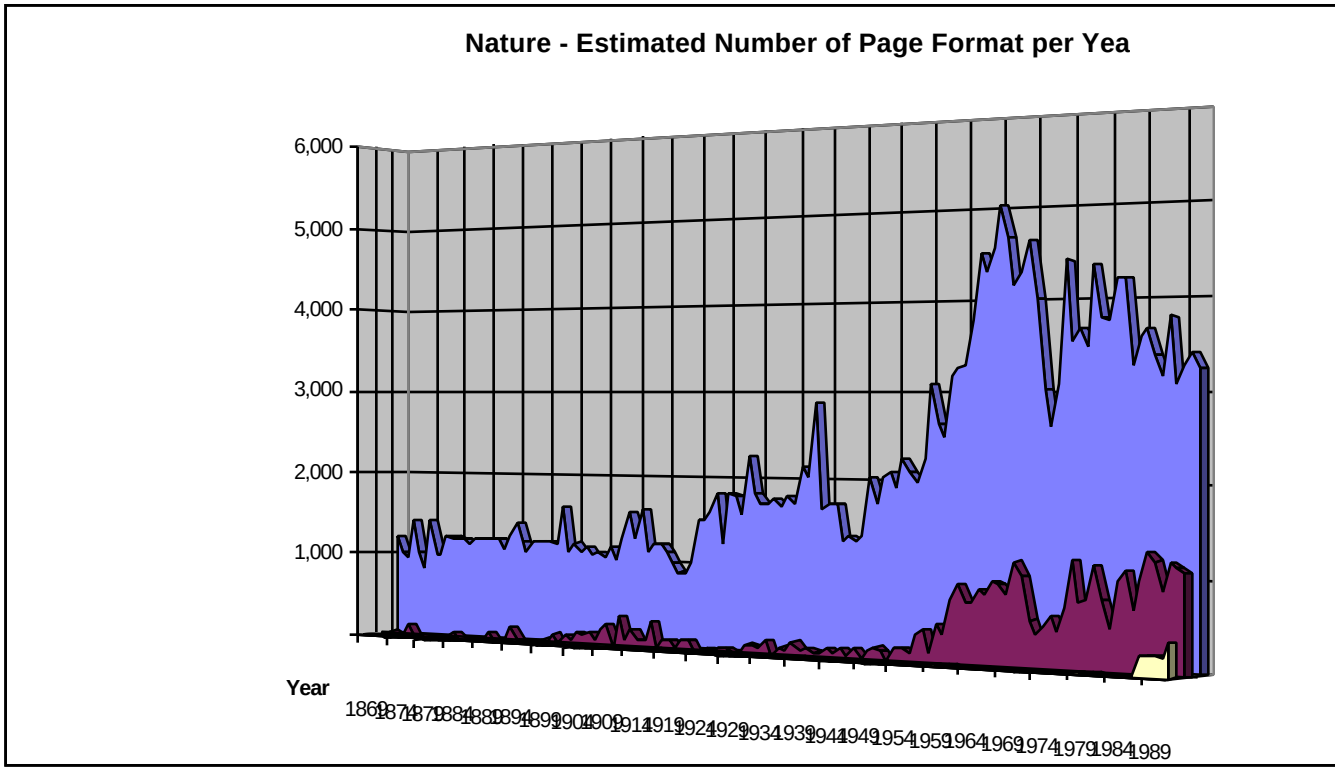

***400 DPI Colour and 600 DPI Black & White Pages***




***600 DPI Greyscale and Colour Pages***



## Appendix B


Nature - Estimated Number of Page Format per Yea
6,000
5,000
4,000
3,000
2,000
1,000
Year
1869
1874
1879
1884
1889
1894
1899
1904
1909
1914
1919
1924
1929
1934
1939
1944
1949
1954
1959
1964
1969
1974
1979
1984
1989

# Effective Terminology Support for Distributed Digital Collections

Martin Doerr
Institute of Computer Science, Foundation for Research and Technology Hellas
Heraklion-Crete, Greece

Paper presented on the Sixth DELOS Workshop, June 17-19, 1998

Abstract

Issues of embedding terminology service into large federations of multilingual digital collections are analyzed. A distributed system architecture is proposed, which preserves on one side a necessary degree of autonomy at the different sources, but allows on the other side to consistently correlate the chain of used terminology from the end-users to the collection maintainers. It should in particular allow for maintaining over the whole federation certain recall and precision properties of collections that use controlled vocabularies.

## *Introduction*

One can roughly separate the problem of heterogeneity in distributed digital collections into a structural one - the differences in the schemata or document structures, and a terminological one - the differences in data values, which may refer to the same real items [Kram97]. There are two kinds of references, those to actual things, "***instances***", as "me, my house, my computer, my publications", and those to groups of things, either ***concepts***, as "researchers, buildings, PCs, essays, roads", or ***non-discrete sets***, as areas on the surface of earth. If one accesses a series of electronic collections, and wishes to retrieve data about certain things, there is the permanent problem of the identity of things referred to by terms. Each social group, be it a scientific discipline or a nation, uses other terms, and even individuals may differ in their use of terms. This problem is tackled by the use of so-called "authorities", which define and standardize terminology of a certain group and domain for consistent use in documentation and retrieval.

This works quite well on isolated databases, but is still insufficient for larger federations of databases. The hope to create a "world-wide" authority can be fairly regarded as an illusion. There are many reasons. First, authorities must be managed and evolved by initially autonomous groups at different paces; second, they often express different, not comparable views even on the same subject; last, the sheer size would be extraordinary, to mention only the most prominent reasons. In this situation, systems of interlinked thesauri or "domain knowledge bases" are proposed (e.g. [Wied94],[Soer96],[Doer97b],[Kram97]).

## *Terminology Support*

Authorities basically try to solve two problems: The identification of a notion, and the definition of a concept (see also [Fosk97],[Soer96],[Doer97b]). For identification, linguistic expressions, so-called "terms", i.e. possible or preferred ***noun phrases*** or ***names*** are associated with a notion according to the practice of a social group. The notion in turn is described by attributes, as life-data of a person, free texts, geo-coordinates etc. The user may

select due to these descriptions fitting and unique terms, which can be used by the retrieval agent to match the notion behind with database records and occurrences in texts. If the database records use unique terms in the same sense (i.e. they use the authority for vocabulary control), the matching is immediate and precise. Else, the authority should list probable expressions for each notion, which can be used by a retrieval agent to calculate approximate matches (free text search). Much sophistication can be put into the context dependency of the probability of an expression and into the related matching algorithms. For "instances", this translation problem between notions and terms can be solved to the best possible, when we have gathered the terms of all groups for each notion. See e.g. the *United List of Artist Names* from the Getty Information Institute.

For concepts and non-discrete sets, however, each group tends to have its own definitions. Moreover, even the same items may be classified or referred to by coarser or narrower concepts. In addition to the pure translation problem, a correlation problem appears. If a data record uses a term for a real item, obviously all queries to broader concepts should include this item. Concepts of one group are therefore organized in subsumption hierarchies as thesauri. Concepts from different thesauri are correlated by equivalence and subsumption expressions in so-called "multilingual thesauri".

This knowledge allows retrieval agents to match data records about related items, but classified with different concepts (e.g.[Doer98]). The precision of this kind of matching is an open research issue. It may return larger or smaller answers than originally requested, e.g. subsumption properties invert in NOT clauses. Users may wish to have control on subsumption properties in the answers they get. If a literature subject is referred, the generality of the concept may or may not relate to the generality of the text. E.g. "Neural Diseases" may better match "Introduction to Neural Diseases" than "A New Approach to Therapy of the Creutzfeld-Jakob Syndrome", making things even more complex.

As above, free texts (scope notes), images etc. support further the identification of a concept by a user. Hence the translation of concepts consists of an identification and a correlation problem of all concepts of all groups, which is obviously an open ended task, as continuously new concepts appear. Authorities, in particular thesauri, can be regarded and dealt with as knowledge bases, which comprise domain knowledge in form of terminological logic, combined with a linguistic layer for concept identification.

### *Current Situation*

From the point of view of implementation and system integration, the current situation can be described as follows:

- Either a separate, not integrated thesaurus tool is used, or there is an idiosyncratic implementation of a thesaurus management within the local collection management system or within a mediator component.
- Some libraries agree to use a foreign (typically English) thesaurus, as e.g. LCSH, ACM subject headings etc., thus giving poor support for the local language and any further specialization to local needs.
- Few systems support automatic query term expansion, in the same or to other languages.

- Evolution of the thesaurus on an external tool and consistent migration of new or changed terms **into** local collection management systems and into other external thesaurus tools is typically not foreseen.

Hence valuable information remains inaccessible, and retrieved information is incomplete and inconsistent with the request, at least by far more than necessary, even though thesaurus formats are standardized since a long time, and thesaurus merging and thesaurus federations are investigated by several groups.

### *The Architecture*

This article describes a proposal for an architecture, which should be able to render integrated terminology services on large federations of digital collections in a scalable and manageable way with similar quality as currently on some local systems. It builds on the experiences and system developments from several co-operations of the author [Doer96], among which the AQUARELLE project (see e.g. http://aqua.inria.fr, [Doer97]) is the furthest-going in this direction, and conforms with more general integrated intelligent access systems ([Wied94] and many others). The equally important question of knowledge acquisition and the effective creation of thesaurus contents is deliberately not addressed in the following (see e.g. [Doer97b]).

We regard the whole as a heterogeneous database problem with vertical distribution and partial data replication. Collection management systems and thesaurus management systems overlap on the terms as shared identifiers. As thesauri undergo slow changes, and adaptations to changes may include manual actions by autonomous groups, partial data replication of thesaurus contents is efficient in a large network. For optimal results, the terms used for asset classification, in a search aid thesaurus and in the experts' terminology should be consistent. This led us to a three level architecture of loosely coupled components cooperating within an information access network: (1) vocabularies in local databases, (2) local thesaurus management systems of wider use and (3) central term servers for retrieval support. (See figure).

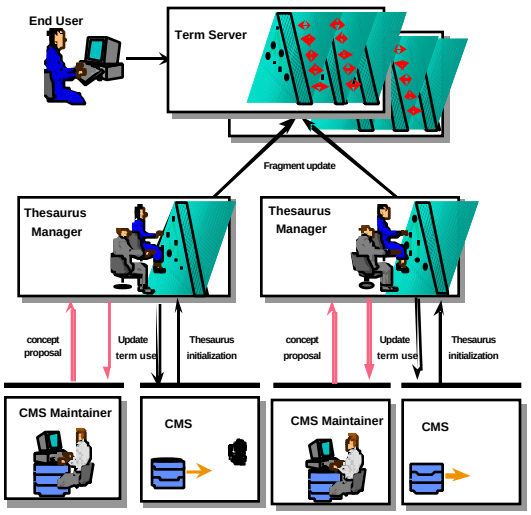



Figure: Terminology service at three levels

We foresee a separate thesaurus manager to which the vocabularies of several local databases can be loaded, and in the sequence organized as thesauri ("authorities") by some experts, following variations and extensions of the ISO2788 semantic structure. In addition, standard external vocabularies can be loaded to it. These authorities may be specific to one database, a user organization, or a whole language group. This thesaurus manager is locally consulted to enforce vocabulary control for classification and query formulation. This implies a knowledge (metadata) about which fields comply with which part of the authority. There are some advantages. A modern organization shares terminology within a group of heterogeneous applications – accounting, warehouse management, decision support, product development etc. Thesaurus software needs quite different structures and is in general maintained by other people than the individual application.

The local vocabularies and terms, that are already used for classification and are under the control of the thesaurus manager may need updating with the changes done at the thesaurus manager. This must be a semiautomatic process, which will be supported by a tool that compares the changes in the thesaurus manager and the use of terms in the local database, and makes proposals for the least changes to be made in the database. If e.g. a term is simply renamed, the change can be made automatically. If a concept is expanded into two, e.g. "tomography" into "CT" and "NMR tomography", a user may need to decide which one applies in each case.

Therefore the thesaurus manager must maintain a history of semantic changes from release to release. The same data can be used to translate or transform terms in a query formulated according to the new thesaurus release against a database consistent with an older release. The retrieval agent will know e.g., that the database understands only "tomography" instead of "CT". The returned answer is larger than expected (reduced precision), but not more imprecise than the database could answer anyhow. This is an essential element of this proposal. Without additional overhead for the thesaurus maintainers, we optimize the update of classification and allow for the simultaneous use of different thesaurus releases without loose of recall and precision.

Typically collection maintainers want to use more terms, than the managed terminology ever will contain (so-called local terms). If the local terms are related in the collection system with the next broader term of the authority, the collection system will give correct answers on the local and the broader term, and this relation can be updated as above. Even more, local terms could be automatically submitted to the thesaurus manager as proposals.

Term servers are used as search aids and need a limited management. Term servers are loaded with multiple thesauri from the local thesaurus management systems. Equivalence expressions will be introduced between the terms in the different thesauri, which on one side help users to select correct terms for databases using authorities he/she is not familiar with. On the other side retrieval agents should be able to make such "translation" automatically, in case many different databases are addressed simultaneously. Therefore term servers may give access to the necessary metadata about which data source uses at which field which part of a thesaurus. Term servers must be updated with new releases of local thesauri maintaining referential

integrity of the equivalence expressions. Again, the history of changes in the thesaurus managers is the key to that.

Equivalence expressions are not easily found, and their number increases with the possible combinations of thesauri. Scalability can be maintained however, if term servers are cascaded to support multiple translation steps (see e.g. [Dao96]). Of course, the precision will decrease over multiple steps. A suitable definition of the equivalence expression can allow to maintain the recall. If wanted, the precision may be maintained, but eventually empty answers are created. [Doer98]. Equivalence expressions may also be created by statistical and language engineering methods. The use of "interlingua concepts", as e.g. the European Education Thesaurus, reduce complexity as well. One may even think of cost-models for the cheapest translation.

In comparison to the size and complexity of multilingual thesaurus structures and management, the information needed to be transferred to classify a data item or to translate a query is small and has relatively simple structure. It is therefore quite efficient to access terminology resources through WANs. To standardize the interfaces between term servers and retrieval agents (a), between thesaurus management systems and local databases (b) and between term servers (c) may be more successful than the current effort to standardize the thesaurus structure itself in order to achieve interoperability. If such standards exist, thesaurus structure is only a matter of agreement between term server and thesaurus management system. In the sequence, more creativity of implementers can be tolerated and rapidly more intelligent services can be provided.

Within AQUARELLE we have enhanced our thesaurus management software SIS-TMS [Doer98b] to support cooperative development of multilingual thesauri. The system features release procedures with history of changes as described above and can also be configured as search aid thesaurus. By graphical visualization it allows for excellent understanding and control of complexly interlinked terminology structures. The solution consist of independent components with open interfaces. The system has found very good user response so far. Other groups have alternative and complementary systems necessary and useful in such an environment.

### *What to do now*

We advocate for international cooperation to implement and experiment with a full architecture as described above. It means

1) To provide solutions for term translation within complex query expressions, e.g. in Z39.50 protocol requests.
2) To develop the methods to manage the consistent operation of such federations and to investigate questions of quality of service.
3) To develop the appropriate network managers, retrieval agents, and data exchange utilities.
4) To define the three basic open interfaces ((a),(b),(c) above). These interfaces become really valuable, when an open communication protocol can be established, which allows to combine freely thesaurus management systems and their term servers with retrieval agents and collection classification systems at an international level.

These activities must be harmonized with improved methods for knowledge acquisition and term correlation (as e.g. aimed at by the TELEMATICS project "Term-IT), and with developments on schema integration and mediation. Under these conditions we believe, that the separation of the terminology service from the retrieval agents and collection management systems into an overall federated architecture as proposed here, has the potential to make effective retrieval from a large number of multilingual data servers a reality. As well, the provision of a correlated terminology rather than reclassification of all data can make even highly specialized data widely accessible.

### *References*


[Dao96] San Dao, B. Perry, "Information Mediation in Cyberspace: Scalable Methods for Declarative Information Networks", in: Journal of Intelligent Information Systems, 6, pp131-150, 1996.

[Doer96] M. Doerr, "Authority Services in Global Information Spaces." *Technical Report, ICS-FORTH/TR-163*, Institute of Computer Science-FORTH, 1996.

[Doer97] M.Doerr, I. Fundulaki, V. Christofidis, "*The specialist seeks expert views: managing digital folders in the AQUARELLE project*", in: Museums and the Web 97: Selected Papers", Archives & Museum Informatics, Pittsburg, 1997. ISBN 1-885626-13-4.

[Doer97b] M. Doerr, "Reference Information Acquisition and Coordination", in: "ASIS'97 - Digital Collections: Implications for Users, Funders, Developers and Maintainers", *Proceedings of the 60th Annual Meeting of the American Society for Information Sciences*, " November 1-6 '97, Washington, Vol.34. Information Today Inc.: Medford, New Jersey, 1997. ISBN 1-57387-048-X.

[Doer98] M. Doerr and I. Fundulaki. "A proposal on extended interthesaurus links semantics." *Technical Report* ICS-FORT/TR-215, March 1998.

[Doer98b] M. Doerr, I. Fundulaki "The Aquarelle Terminology Service", ERCIM News Number 33, April1998, p14-15

[Fosk97] D. J. Foskett. Thesaurus. In *Readings in Information Retrieval,* eds. K. Sparck Jones and P. Willet, publisher Morgan Kaufmann, 1997.

[Kram97] R. Kramer, R. Nikolai, C. Habeck. Thesaurus federations: loosely integrated thesauri for document retrieval in networks based on Internet technologies. In *International Journal on Digital Libraries* (1), pp. 122-131, 1997.

[Soer96] D. Soergel. "SemWeb: Proposal for an open, multifunctional, multilingual system for integrated access to knowledge about concepts and terminology." *Advances in Knowledge Organization,* 5, pp.165-173, 1996.

[Wied94] G. Wiederhold. Interoperation, mediation and ontologies. In *Proceedings of the International Symposium on Fifth Generation Computer Systems (FGCS94), Workshop on Heterogeneous Cooperative, Knowledge-Bases* (ICOT), Japan, December 1994, W3, pp.33-48.

# Beyond HTML: Web-Based Information Systems

**K. V. Chandrinos, P.E. Trahanias**

**Institute of Computer Science**

**Foundation For Research and Technology - Hellas (FORTH)**

{kostel|trahania}@ics.forth.gr

In this paper we briefly review current status of Web-Based Information Systems (WBIS) and present a number of different WBIS technologies and how they are being integrated during the ARHON project (Archiving, Annotation and Retrieval of Historical Documents). The project amounts to transforming a vast collection of OCR-resisting historical manuscripts belonging to the Vikelaia Municipal Library at Heraklion, Crete, to a fully functional Digital Library supporting the needs of casual users as well as scholar researchers [**Erreur! Signet non défini.**]. The solution we opted for, was the creation of an image database and the adoption of standard Web browsers enhanced with Java applets for browsing, querying and updating the database. This way we managed to provide an Intranet based querying and retrieval mechanism through a Web server. The access mechanism was designed to work with standard browsers so that at a later stage Internet access to the digital library should be trivial to implement. Initial decisions and results were presented in [**Erreur! Signet non défini.**].

**Introduction**

The widespread use of the Web for information management, as opposed to mere presentation, has revealed the weak points of HTML as a general purpose scripting language for information retrieval and query result presentation over the HTTP protocol. Early implementations of Web front-ends to libraries, that proposed an HTML form page in place of typing into a telnet connection, have rather disappointed end-users who had to face extended network latency for no apparent profit. We share the view that digital libraries should offer the end-user enhanced functionality compared to the physical library [12, 14, 19]. Despite recent attempts to extend the language specification (**Erreur! Signet non défini.**) and reverse its static nature utilizing server-side processing (e.g. **Erreur! Signet non défini.** [8], **Erreur! Signet non défini.** etc.) it is still hard to achieve the dynamic interactive environments we would like to have at our disposal when accessing large online repositories. Based on the experience gained during our project design, we intend to present the main problems of **Erreur! Signet non défini.** for Web-Based Information Systems (WBIS) and possible leeways for enhancement offered by recent developments. After reviewing current capabilities, we explain and justify our rationale for choosing a combination of standard HTML with **Erreur! Signet non défini.** applets and servlets to implement a 3-tier architecture for the manipulation of the digital library.

**Web-Based Information Systems**

Although most HTML programmers feel they could express their data of interest if they had just one more tag, one that would suit their particular project, a number of writers have identified the key problems of HTML that prohibit an effective transition of Information Systems to the Web [5, 7]. The chief shortcomings of HTML, which cannot be addressed by proprietary tags introduced from rival browser companies, boil down to lack of:

- *Extensibility*, since HTML does not allow users to define their own tags.
- *Structure*, since HTML does not allow deep structures needed to represent object-oriented hierarchies or database schemas.
- *Validation*, since HTML does not allow browsers to check data for internal validity on downloading.

These problems, combined with the nature of HTTP as a transfer protocol, have led to a number of difficulties affecting attempts to bring content to the Web. For example, they specify that only discrete transactions between client and server are allowed. Even worse, these discrete transactions can only be anonymous and stateless. As a result, current standard Web technologies can support distribution of only static multimedia documents. The introduction of plug-ins to remedy this, apart from being awkward, violates the openness and vendor-independent nature of the Web. Also, a discrete, stateless transaction model does not support either familiar interface elements (e.g. context sensitive help) or event-based interaction (e.g. error-checking on input, direct manipulation).

To avoid the shortcomings of HTML and HTTP most of the users first turned to server-side processing utilizing the Common Gateway Interface (CGI) scripting facility. Form processing answered, although clumsily, the problem of the client initiating a server response. Of course, the form validity could only be checked when it reached the server, a thing that added further delays to the already slow model of process-spawn-per-request, which CGI scripting imposed. An answer to data checking on input came from embedded JavaScript code or Java applets. As opposed to

with particular vendors resisting full compliance to the Java specification issued by Sun, but momentum seems to gather in that direction. Java endows the Web programmer with capabilities to express the most complex interface elements in the form of applets. However, the true power of network programming and network-oriented languages like Java, stems from server-side processing. As one should expect, not all the material that deserves exposition to the Web is already formatted in an appropriate way. Server-side programming allows content publishers to afford a considerable choice of middleware between their legacy-formatted data and the Web. The most promising version of middleware seems to be the Common Object Request Broker Architecture (**Erreur! Signet non défini.**) which stems from a large industry powered consortium, the Object Management Group (**Erreur! Signet non défini.**). This architecture allows existing or future components to be treated as objects, where the interfacing between these components is provided by static or dynamic methods. This way, communication with the components becomes independent of implementation (it could equally be C++ or COBOL code), provided interfacing to each component follows certain rules. Of particular interest are the CORBA services known as Internet Inter-ORB Protocol (IIOP) that allow components to communicate with each other over the Internet [17]. Using these services, one can hide implementation details or even execution platform and location of the components, making multi-tier architectures possible. CORBA is still in its infancy concerning deployment but despite the initial market reluctance to adopt it, spurred by the decision of Microsoft Corporation to pursue their own Distributed Common Object Model (DCOM), standardization moves at a fast pace. A lightweight version of middleware that can provide vendor independence can be implemented using Java servlets. Servlets are like applets, save for the graphical user interface. They only run in servers and, being genuinely multi-threading, avoid the caveats and latency of CGI scripting. Utilizing the Java Native Methods Invocation scheme, which allows direct calls of C-code from within Java programs, we have managed to implement a fast and reliable 3-tier architecture as we explain in a later paragraph.

Despite their shortcomings the combination of HTML & HTTP have managed to provide a more or less uniform way to transfer content over the Web. With the advent of **Erreur! Signet non défini.** they can also provide a minimum set of uniform presentation facilities. However, what they cannot cater for, is semantics. Descriptive markup provided by HTML tags such as <H1 is tightly bound to "structure". On the other hand, introducing markup with tags such as <AUTHOR falls into the category of "semantics". To achieve full semantics representation one could easily suggest the adoption of the Standard Generalized Markup Language (**Erreur! Signet non défini.**). Realistically speaking though, this is out of the question. Even if one managed to convince the browser producing companies to comply with the 500-page specification of SGML, spending time and money on a fully SGML compliant browser that they would probably have to give away for free, it would be outrageous to expect that the users will have to climb a slow learning curve to see their content semantically marked up on the Web. The World Wide Web Consortium (W3C) has acknowledged this deadlock and has initiated a Working Group that has come up with a rigorous subset of SGML which they call eXtensible Markup Language (**Erreur! Signet non défini.**) in their draft specification. XML manages to maintain an important number of SGML features in its 26-page draft specification and provides remedy for a lot of the HTML handicaps, e.g. XML advocates custom tagsets and n-ary links. All this without loss of the strictness attributed to SGML. XML intends to provide semantic markup facilities, thus allowing communities to formulate their common level of understanding. Such a development could lead to each community having its own ontology codified in an XML Document Type Definition (DTD), allowing future knowledge to be exchanged fast and reliably [5, 7, 11].

**The ARHON design: a multi-tier architecture**

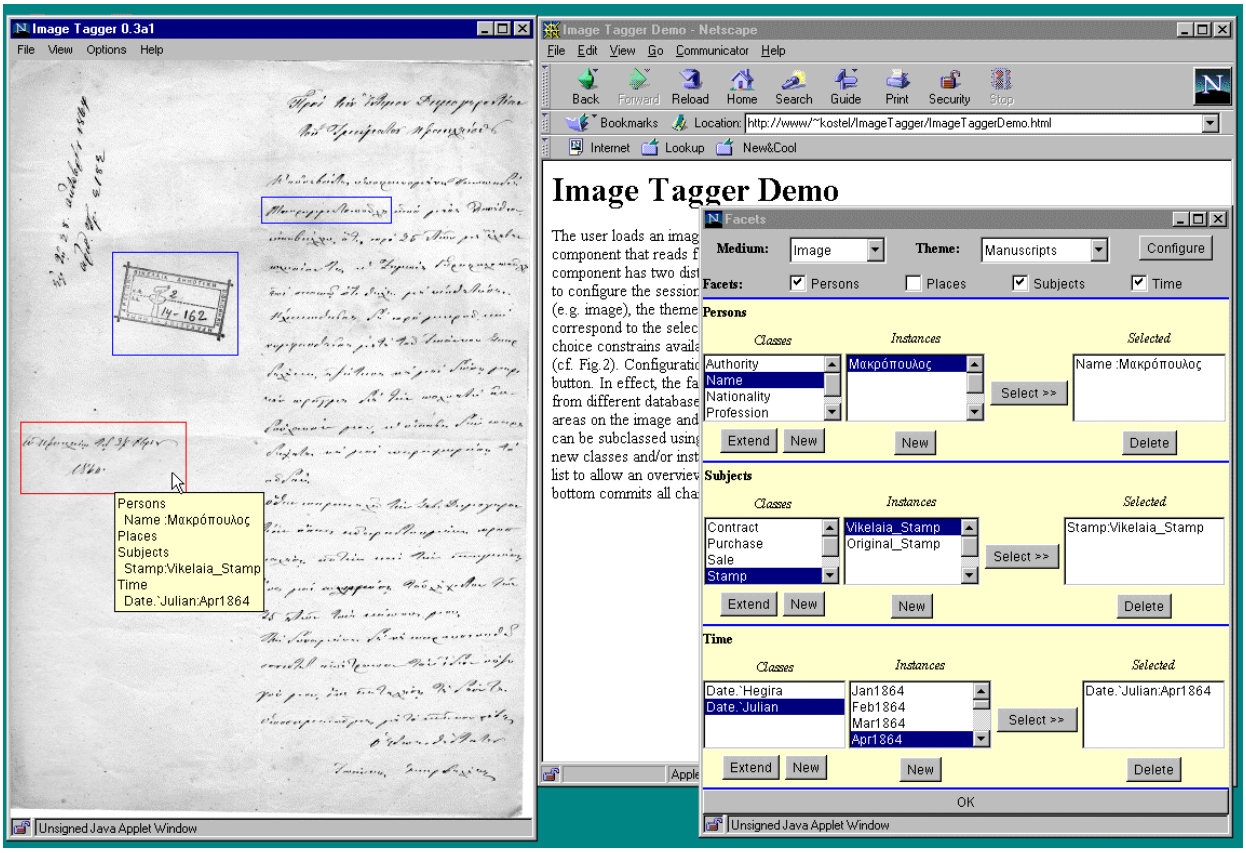



Figure 1*: ImageTagger* is a tool for rapid registering of data in image databases

Most client/server models (2-tier architecture) find hard to cater for the complexity of needs arising everyday. They give way to new models that include many servers, running on different platforms and physical locations distributing data and sometimes management of data. In our application, the first tier (client) is any number of Java-enabled browsers. These clients connect to a Web-server to download applets that handle complex interface tasks, consider error checking on input and present appropriately formulated query results. The second tier consists of servlets, Java programs running on a Web server, which encapsulate the rules, state and logic of our application, that is the management of a digital library. They also control access to the data repository (third tier) using a proprietary legacy API.

**The problem**

The quality and load of historical manuscripts we had to face in designing this particular digital library has deferred us both from OCR attempts and typed-in transcriptions [1, 13]. On the contrary, we decided to scan the documents and apply image-enhancing techniques to improve their quality, sometimes further than the originals. Since the semantic information appeared in several parts of each document it had to be explicitly identified by scholar researchers at different levels. To facilitate their work we developed *ImageTagger*, a tool for interactive mark-up and annotation of image documents [**Erreur! Signet non défini.**]. Among the merits of this tool, as opposed to standard SGML mark-up editors, is that it tags and annotates the image of the document visually. There are a number of technically sound reasons why embedded mark up should be avoided when possible [6]. In our case, practical reasons have led us to the design of ImageTagger as a non-embedding markup alternative: it doesn't presuppose that you have a document in ASCII format and it doesn't affect the original, allowing for different treatment of the same data in the future (see figure 1). Also, since it is implemented as a Java applet it allows for remote tagging and annotation and does not depend on a database coming from a particular vendor. In our case, we used an object-oriented semantic net that supports semantic indexing (Semantic Indexing System [**Erreur! Signet non défini.**]) developed at **Erreur! Signet non défini.**. However, the overall design does not restrict us from using any other DBMS, be it relational or object-oriented. On the contrary, it allows concurrent use of more than one DBMS.

**The architecture**

Accessing the data for querying or retrieval is achieved by a Java/HTML user-centered interface. A different incarnation of an AccessApplet is served to each client according to his/her status (administrator / scholar / casual user) decided at login time. The dynamic nature of the interface allows the client to change status or even language during a session if appropriate (figure 3 depicts the working version of our interface).

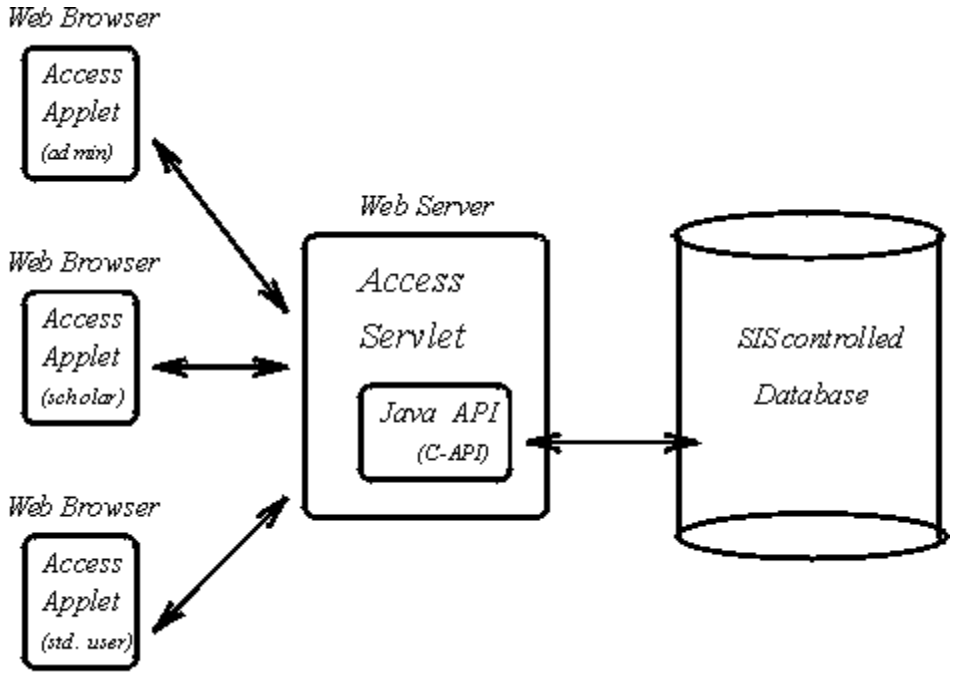



Figure 2: The ARHON architecture

To tackle the lack of state inherent to the HTTP protocol, we propose a server-side solution that is an improvement on "**Erreur! Signet non défini.**", the current standard solution for state preservation. We rely on a multithreaded AccessServlet running on the Web-server all the time, delegating after approval all client connections to the database. This servlet is also responsible for monitoring the consistency of transactions and integrity of the database, as it can roll back all unfinished transactions and track simultaneous attempts by users to update the database. Since servlets are programmed in Java we would have either to design and implement a Java API for the semantically indexed database, or convert an existing and proven C-API of the client-server model. This C-API has worked for a number of multi-linked thesauri installed in the past few years with the support and maintenance of the **Erreur! Signet non défini.** group at FORTH. Our choice was to utilize the existing C-API , by creating a Java API as a wrapper, in the Java Native methods Invocation scheme (JNI) which proved a fast and efficient way to utilize legacy code. Initial experiments show that the API interface latency is negligible.

The above sketched architecture (figure 2) has allowed us to surpass a number of problems presented by static HTML. At this point of the project we are turning our attention to digitizing a large amount of data and inserting it in the database so that we can perform efficiency analysis and improvements. Once a critical mass of digitized documents aggregates we intend to research on techniques concerning user profiles and query refinement.

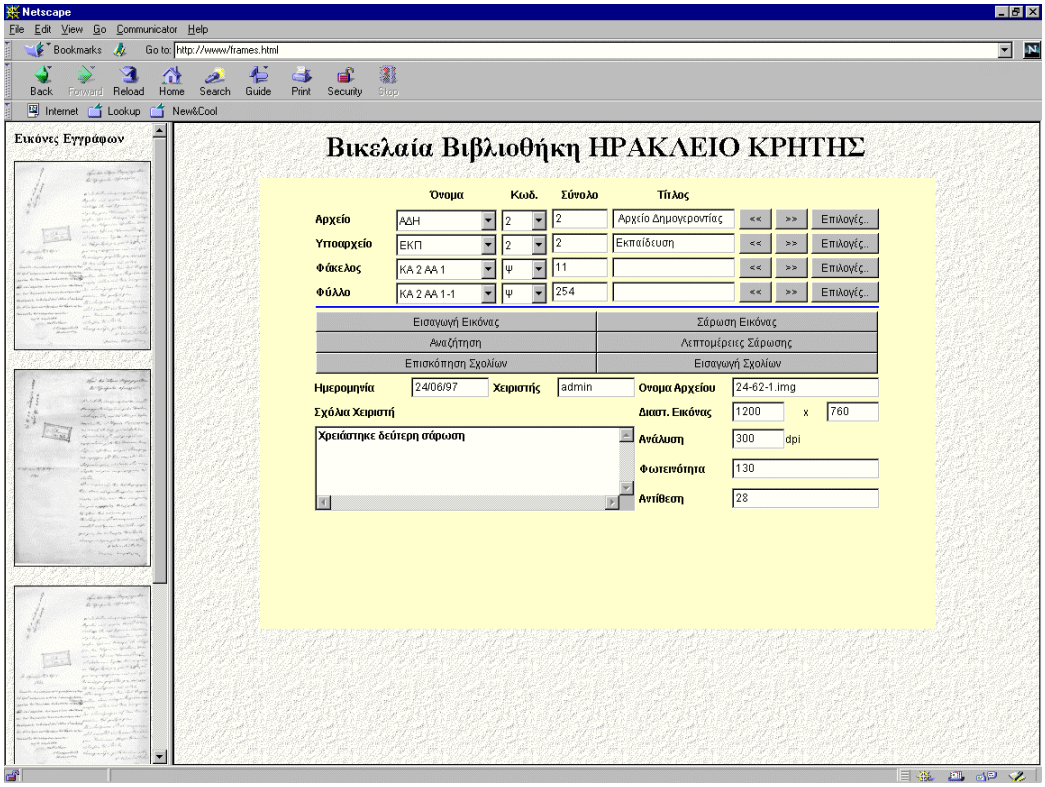



Figure 3: The AccessApplet acting as an interface to our image database

### A Note on Security

Although at this stage of the ARHON project Internet access is not be provided at least for an initial period, we have strived to provide the infrastructure that will ease the transition from an Intranet to an Internet use of the digital library. One crucial point to notice, particularly when working with multimedia databases such as ours, is the notion of security. Granting Internet access to a digital library is compared to opening a window at your house to allow other people to look in. Whether they will jump in to steal your belongings is a very complex social discussion. It is obvious that the level of security should reflect the value of the assets and the expenses the owner (or publisher, in our case) is ready to put themselves to. It is also obvious that with respect to image databases that can be accessed over the Internet there is no perfect security scheme, yet. In this light, what can be done is making the life and works of the potential misuser a bit more difficult. For our particular application, we have devised and implemented a digital watermarking technique that is meant to dissuade people from falsely claiming ownership of the images in our database.

Figure 4. The digital watermark superimposed on a document image

Watermarking techniques are used for the identification of the originator (owner) or the recipient of material. They are usually either visible or invisible [15, 16]. In our case, we devised a method that combines an invisible 128-bit digital signature, hidden in the image of the actual document with a visible watermark that informs potential misusers that this image is under copyright. We have made the visible watermark very robust by interweaving it with our image, making sure it resists statistical or other attacks, which will remove part of the text appearing on the document image along with the watermark (see figure 3). The only way to remove it is by using the actual digital signature hidden in it as the key to an algorithm that we trust with the archivists of the library. This way, either them or the persons they pass their trust to, are the only people who can remove the watermark in order to examine the original document. Of course, this security scheme, like almost every other, will collapse if someone is determined to sacrifice an enormous amount of human and technological resources. However, we feel that it suffices for our cause, which is to make our digital library less vulnerable to malicious attacks.

**Conclusions**

A Web-based architecture should serve as the connecting glue between subsystems developed or even executing on different platforms. It presents excellent modularity and openness per component. It also allows for extremely fast prototyping. Since nowadays the time factor seems to have shifted from the speed of software execution to the time-to-completion of a reliably working application (or project) we consider this latter aspect of Web-based architecture very critical. The same architecture imposes uniform interfacing to the data over different platforms, except for minor differences in the layout of each browser. Additionally, with the expected gradual adoption of the Unicode encoding scheme we anticipate enhanced automatic support for multilinguality. Given that the Web is a widely distributed environment the extension of a digital library with new material or even the concatenation of different digital libraries can be easily and transparently achieved [4, 9, 10]. Also, we should point out the great advantage of open standards governing the Web: one is not trapped through this solution to a proprietary format or even a particular vendor and his/her support habits. The number of people currently working on applications of Web technologies is rising, and this, apart from investment protection for the future, guarantees low cost and reliable support during development. Most important, this support can be provided remotely. Until all content is prepared with the Web in mind, there is a clear need to utilize existing technologies and augment them so as to make current content accessible and presentable. In this paper we have provided an example of a multi-tier architecture implemented with state-of-the-art technologies. We feel that this example can provide inspiration for further refinements that will set up the foundations of WBIS relying on existing content.

The authors would like to thank Dr. John Immerkaer for constructive reviewing of the image database architecture and technical discussions. The prototype implementation of the digital watermark techniques mentioned in this document were carried out by Ghislain Bidaut.

**Production Note:**

This document was originally written in HTML. It can be found at **Erreur! Signet non défini.** with its links and references active, where appropriate. It will be maintained and updated by **Erreur! Signet non défini.**

References:

[1] **Erreur! Signet non défini.**.
K.V. Chandrinos, J. Immerkaer, P.E. Trahanias, EUSIPCO 98, Special Session on Multimedia Signal Processing

[2] **Erreur! Signet non défini.**.
K.V. Chandrinos, J. Immerkaer, Martin Doerr, P.E. Trahanias, ERCIM 5th DELOS Workshop on Filtering and Collaborative Filtering, Budapest, Oct. 97

[3] **Erreur! Signet non défini.**

[4] The German National Bibliography 1601 - 1700: Digital Images in a Cooperative Cataloguing Project.
M Doerr, H. Haddouti and S. Wiesener, in Proccedings ACM Digital Libraries '97

[5] XML, Java and the future of the Web.
Jon Bosak in D. Connolly eds., XML: Principles, Tools and Techniques, World Wide Web Journal, Vol 2 (4) 19'97

[6] Embedded Markup Considered Harmful,
Theodor Holm Nelson in D. Connolly eds., XML: Principles, Tools and Techniques, World Wide Web Journal, Vol 2 (4) 1997

[7] XML: A door to Automated Web Applications
Rohit Khare, Adam Rifkin, IEEE Internet Computing, July-August 1997

[8] **Erreur! Signet non défini.**
Brian J. Fox (http://www.metahtml.com)

[9] Building a Digital Library: The Perseus Project as a Case study in the Humanities
Grecogory Crane, in Proccedings ACM Digital Libraries '97

[10] **Erreur! Signet non défini.**
Sandy Ressler and Bill Trefzger, IEEE Internet Computing, September-October 1997

[11] **Erreur! Signet non défini.**
Maintained by the W3C XML Special Interest Group

[12] **Erreur! Signet non défini.**
E. Doerry, S. Douglas, T. Kirkpatrick and Monte Westerfield, Tech. Rep. CIS-TR-97, Computer & Information Science Dept. U. of Oregon

[13] **Erreur! Signet non défini.**
C. Palowitch, Darin Stewart, in Proceedings of DL 95

[14] Browsing is a Collaborative Process
M. Twidale, D. Nichols, C. Pace, Information Processing and Management, Vol. 33 (6) pp. 761-783, 1997

[15] **Erreur! Signet non défini.**
F. Mintzer, J. Lotspiech, N. Morimoto, D-Lib Magazine (http://www.dlib.org), December 1997

[16] **Erreur! Signet non défini.**.
Martin Kutter, Frederic Jordan, Frank Bossen in Proceedings SPIE-EI97, 1997

[17] Client/Server Programming with Java and CORBA,
R. Orfali and D. Harkey, 2nd ed., Wiley, 1998

[18] Protecting Ownership Rights through Digital Watermarking,
Berghel, H. and L. O'Gorman, IEEE Computer, 29:7, pp. 101-103 (1996)

[19] Database Backed Websites, P. Greenspun, ZD-Press, 1997

# Retrospective Conversion of Old Bibliographic Catalogues*

A. Belaïd

LORIA UMR, Campus Scientifique, B.P. 239

F-54506 Vandœuvre-lès-Nancy Cedex, France

e-mail : abelaid@loria.fr

**Abstract**

This paper describes a framework for retrospective document conversion in the library domain. Drawing on the experience and insight gained from the MORE project launched over the present decade by the European Commission, it outlines the requirements for solving the problem of retroconversion of old catalogues in UNIMARC format. Based on OCR technique and automatic structure recognition, the system proposes a direct schema for the conversion of references in machine readable records. Furthermore, as the system is meant for a real production chain, the paper describes the industrial constraints and gives a complete benchmark realised on this chain for 11 volumes and 4568 references. Without any manual intervention, the recognition rate of the system is greater than 75%.

**Keywords** : Retrospective Conversion, Library Catalogue, Reference Recognition, Structure Analysis, OCR, UNIMARC

*This work was funded by the EEC libraries programme LIB-MORE

# 1 Introduction

The success of library automation, resulting in user-friendly on-line catalogues[1] integrated with the WEB and other circulation-systems facilities, has created an urgent need for retroconversion of the older parts of catalogues [1, 6, ?, 11, ?]. As users get used to the new catalogue medium, the documents not registered in machine-readable form become "invisible" and unreadable. This has meant for many libraries the relegation of an important part of their rich stock of documents to a state of inaccessibility.

Such obvious waste of library collections in addition to the cost difference between manual handling and an equivalent set of automatic routines has made a strong case for the need to convert a library's entire collection of works to machine-readable records, in the interest of ensuring an efficient use of the investment in the new technology.

This has led to the search for cost-effective tools for the conversion of old catalogues into machine-readable forms. This search has not been limited to the sole problem of conversion but has been extended to embracing other objectives such as ensuring very high rates of distribution and sharing of documents between several libraries.

Drawing heavily on the experience and insight gained from MORE[2] [2, 3, 10] this paper outlines the main phases of retroconversion for a real production chain and states the relevant requirements of the retroconversion operation in such a chain.

# 2 Automatic Retroconversion

The use of generic tools to manipulate bibliographical information almost invariably poses the same problems. These are related to the following facts :

- *Heterogeneous Content.* The reference catalogue have usually been produced over a long period of time during which the cataloguing rules have changed. It contains references produced by different cataloguing agencies each applying

[1] A catalogue is a list of bibliographic descriptions of works.

[2] Marc Optical REcognition

their own rules. Many catalogues to be converted contain many different types of references: main entry references with headings representing authors or titles. Added entries by secondary authors, title, subjects, etc. Entries covering more than one reference. The system will have to be able to differentiate between these types and handle the information according to the type.

- *Typographic imperfections*: Bibliographical information is made up of text containing a large number of abbreviated words, not only in the document language but in the cataloguing language as well. It also contains numerical information, sometimes in Roman numerals, and an important quantity of names. To these must be added the multiplicity of languages used and the use of a wide range of stressed characters not in keeping with Latin writing styles. There is higher frequency of punctuation marks than in ordinary text. In addition to their natural role, punctuation marks are used as separations to delimit logical elements of information. The presence of several similar character sets such as hyphens and long dashes, parentheses and square brackets, further increase their frequency. Printed catalogues make use of typography to differentiate between sets of elements belonging to the same logical category. Unlike card catalogues, the layout is more elaborate, including systematic justification of text, variable spacing, and at times word cutting at the end of line. Some of the word cuts belong to the very publication language covered by the catalogue to be converted.

- *Linguistic variabilities*: The recognition of some fields depends on the recognition of some key words in specific lexicons. In these lexicons we can find all the cataloguing vocabulary and all the words that exist in bibliographical work titles and insertions concerning the "authorship responsibility". Punctuation is currently less reliable than that of ISBD [7]. Some words are related to the publication language (title fields, edition, address, collection) and others are related to the cataloguing language (collation and notes). Finally, all the words have to be taken into account in a complete form and also in an abbreviated form, knowing that they were not normalized at the time of the tests.

- *Higher Density of Structure*: The main problem posed by the bibliographical references resides in the density of their logical structure and the multiplicity of choice of information sequences. In fact, several cataloguing entities are optional and repetitive. These information elements are required only for the cataloguer, if the information exists in the catalogued document. Furthermore, these elements can depend on the kind of the document and of course on the kind of references, such as "monograph" or periodical publications, or as in certain catalogues, on "principal" or "secondary" reference. Finally, a practice inherited from printed catalogues is at the root of the current use of punctuation marks as a means of condensed representation of information. The ISBD normalization on the international level further reinforces this.

# 3 The Belgian Catalogue

## 3.1 Structural Aspects

The Belgian Bibliography is presented as a series of monthly catalogs on paper. Each catalog is divided into two parts. The first part contains the bibliography body while the second is filled with indexes leading to authors, subjects treated (titles, collections, rubrics in French and in Dutch, etc.).

### 3.1.1 Layout Reference Structure

The layout structure is very poor; it is partitioned into five areas (cf. figure 1). The first area, composed of the first line, contains on the right hand side, the "CDU" code (Classification Décimale Universelle) which gives some information about the library classification of the reference. The second area contains the reference body. It is composed of a series of fields describing the work referred in the reference such as : "*heading*" (author name or beginning of title), "*title*", "*address*", "*collation*" (material description of the work : location, editor, year, format, etc.). The body is often typed

in many lines. The third area contains the "*Collection*" field (description of the series, volume, etc.). The fourth area contains the "*Note*" field which gives information about, for example, the title (abbreviated, complete, original, etc.). These last two areas are optional and so are not always present in some references. The last area, located on the last line of the reference, contains the "*reference*", on the left hand, and the "*order number*", on the right hand.

| | | |
|---|---|---|
| | 159.962 | UDC |
| **Liger-Belair (Gérard).** Je suis fakir. ([Par] Gérard Liger-Belair). (*Verviers, Editions Gérard & C°*, 1973), 32° *carré, couv., ill.*, 158 *p.* (30 *fr.*). | | Body |
| (Marabout-flash, 352). | | Collection |
| [Titre introductif : Souvenirs, révélations, conseils]. | | Note |
| B.D. 14.814 352 | *73-2108* | Ref |

Figure 1: Example of a library reference.

### 3.1.2 Logical Reference Structure

The logical structure is, on the other hand, more dense. A "heading area", representing the first author or the beginning of a title is always located at the begining of the "body". As for the rest, there is an enormous number different possibilities. We can find, for example, depending on the references, "principal authors" or "secondary" (introduced by some characteristic expressions) which can be physical persons or legal entities, some "main titles", "parallel" (printed in different languages), or "partially", "sub-titles", "publishers" with their "addresses" and the "date" of publication, an area "collation" describing the characteristics of the work (number of pages, format, supporting documents, etc.).

# 4 Automatic Recognition System

Figure 2 shows the main phases of the recognition process of references. In the following, we briefly describe these different components for the Belgian Library.

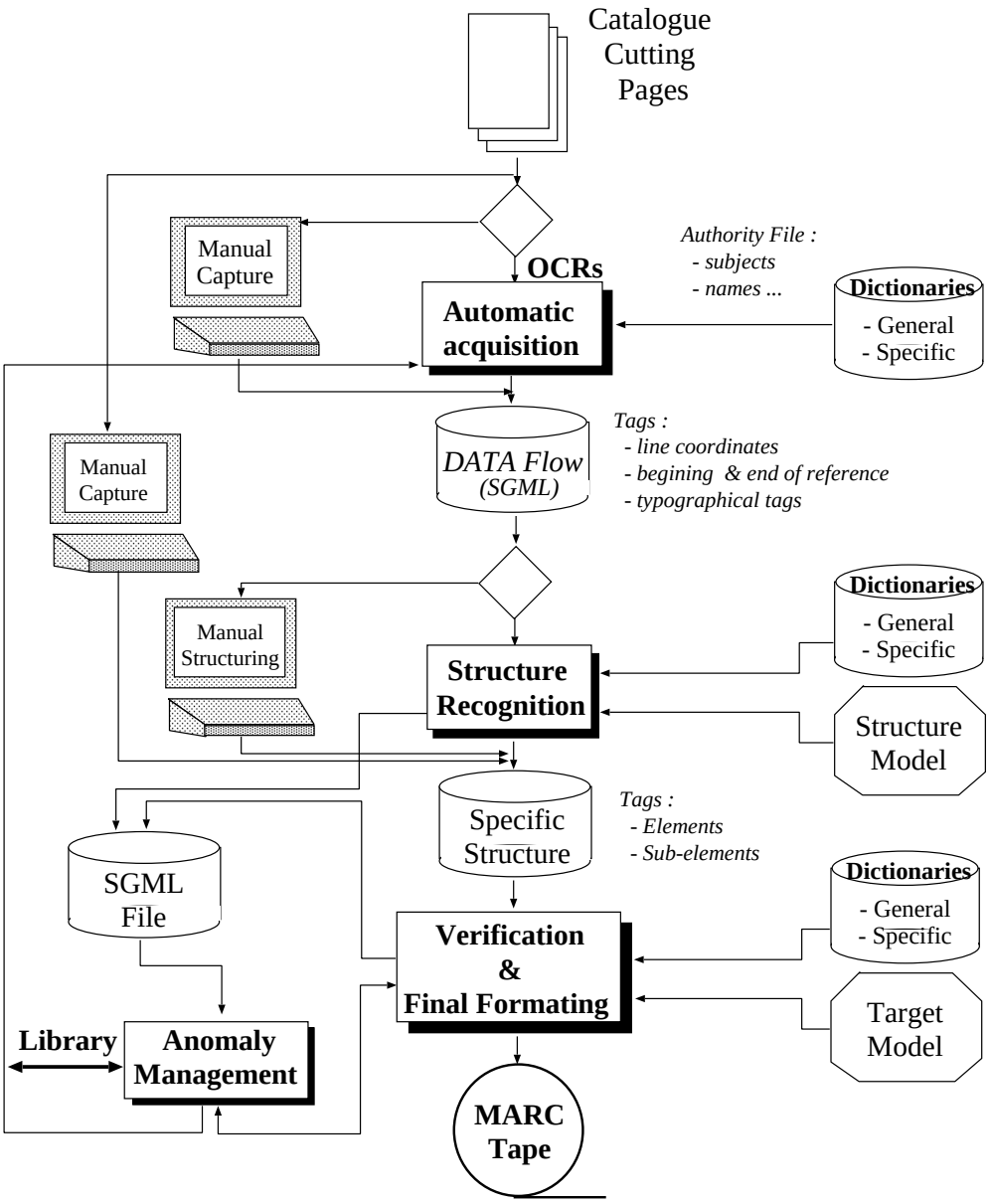



Figure 2: System Overview.

## 4.1 Data Acquisition

The main problems with handling catalogues are related to the automatic feeding of the pages or cards, the existence of cards printed on both sides, and the variable quality of well used machinewritten cards.

For catalogues in volume, as used in MORE project, pages are separated and feeded separately. The project has identified scanners able to handle a great quantity of pages at acceptable speed. In fact, the speed of the scanning process does not depend entirely on the scanner itself, but also on the controller page as well as the speed of the controlling system. For the resolution and because of the variations in printing quality, many tests were operated in order to determine the average resolution (here equal to 400 dpi) which can be used for all the catalogue pages without changing during the feeding.

Data acquisition also includes data formatting. Being individually pasted into pages, the reference images are altered (skew angle, font changing, cut or connected characters, etc.). Specific algorithms had to be developed in order to take into account these particularities [2]. At the end of this process, each reference is extracted from the page image and given to the recognition system as a list of successive lines.

## 4.2 Text Conversion

Each reference is passed through a series of commercial OCRs. The results of these OCRs are combined to obtain the best response. The reason for this is that references contain a lot of different symbols (such as punctuations, indices, exponents, and multilingual words typed in different sizes and styles) which are very difficult to recognize using only one OCR. We thought that combining the results from different specialized OCRs will give a maximum of information on the text, its style, its language, and on its separators.

The result of these tasks is a data flow containing the reference text coded in SGML [8]. The tags separate the lines and different information such as style or lexical

class corresponding to each word (token). Figure 3 shows the flow corresponding to the reference of figure 1.

The reference is located in this flow between two successive tags "<NOT" and "</NOT>". Useful tags for the document analysis are "LEX" which gives the lexicon affiliations of words, "I" for italic style, "B" for bold style, and "S" for the number of spaces. The defaults style is standard and as such not tagged. It is possible to have some errors during this first conversion (especially in recognition of style and punctuation). For example, the exponent "o" in $c^o$ is replaced by the character "o". Anther initial recognition error concerns the style of the end of the secondary title which is identified as "standard" instead of "italic".

## 4.3 Structure Modelling

Knowing that the problem is to find the sub-fields within reference areas, the model specification concentrated on the description of sub-field properties, by the distinction of their typographic styles, the existence of particular words or group of words and their appearance in certain lexicons, and essentially their limits (type of initials and finals such as capital letters, particular words or type of punctuation separating the sub-fields).

The model is given by a context-free grammar written in the EBNF formalism. The format of a production rule is as follows :

| | | |
|---|---|---|
| Term | ::= | Constructor subordinate_Objects[Qualifier] \| |
| | | Constant \| Terminal |
| Constructor | ::= | seq_td \| seq_lr \| seq \| aggr \| cho \| import |
| Separator | | Name subordinate_Objects |
| Attributes | | [Name Weight]$^+$ |

```
<DOC   % image source        1st reference last reference  Directory
TY=N PROV=ENRLEX EG=OK NPN=2085 NDN=2114 IMA=users/brb/juin73/images>

     <PAG % number              bounding box
    NP=1 NOM=0008.ima> <COL XHG=63 YHG=1900 XBD=1027 YBD=2912>

        <NOT % number coordinates
            NON=2108 EN=OK>
         <LIG XHG=870 YHG=2215 XBD=1000YBD=2266 YBSL=2256 ST=t>
         <REDF=85.69>159.962</LIG>
         <LIG XHG=149 YHG=2260XBD=1001 YBD=2313 YBSL=2298 ST=p><B>Liger-Belair</B>
           <I>(Gérard).</I><LEX L=GFR,GNL><REDF=50.00>Je <LEX L=GFR>suis <LEX
           L=GGB,GFR>fakir.<LEX L=GGB,GFR,GNL><RED F=99.99>([Par] <REDF=99.97>Gé-</LIG>
         <LIG XHG=148YHG=2304 XBD=1002 YBD=2356 YBSL=2342 ST=p>rard Liger-Belair).<RED
           F=89.99> <I>(Verviers, <LEX L=GGB> <RED F=100.00>Editions <LEX L=GNL>
           <REDF=99.99>Gérard</I> <I>\&amp;amp; </I></LIG>
         <LIG XHG=151 YHG=2350 XBD=1000 YBD=2403 YBSL=2388 ST=p>C0,<RED F=83.33>1973),
           320<LEX L=GFR,GNL><I>carré, <RED F=66.66>couv.,ill.,</I><RED F=99.97>158
           <RED F=25.00>p.30 <I>fr.</I>).</LIG>
         <LIG XHG=149 YHG=2406 XBD=549 YBD=2457 YBSL=2443 ST=p><LEX L=GGB,GFR,GNL>
           Marabout-flash,352).</LIG> <LIG XHG=148 YHG=2447 XBD=1001 YBD=2499 YBSL=2485
           ST=p> <LEXL= GFR><RED F=99.99>[Titre <LEXL=GFR>introductif:<LEX L=GGB,GFR>
           <RED GFR>F=89.99>Souvenirs,<LEX L=GFR>rivilations, <LEX GFR> L=GGB,GFR,GNL>
           con-</LIG> <LIG XHG=149 YHG=2494 XBD=245 YBD= 2545
           YBSL=2530 ST=p>seils].</LIG>
         <LIG XHG=148 YHG=2546 XBD=1002 YBD=2599 YBSL=2584 ST=t> <RED F=43.75>B.D.
         14.814 <REDF=99.97>352<SN=15><I>73-2108</I> </LIG>
       </NOT>
     </PAG>
</DOC>
```

Figure 3: Flow of the reference given in figure 1.

### 4.3.1 Constructors and qualifiers

A term, the left hand of a rule, can be either simple (constant or terminal) or composed of subordinate objects. In the last case, a constructor describes the relationship between objects. The constructor precises the order of the appearance of subordinate objects such as SEQUENCE : top-down (*seq_ td*), left-right (*seq_ lr*) or logical (*seq*), AGGREGATE (*aggr*) or CHOICE (*cho*). A special constructor "*import*" is used to inherit for the term some or the total description of another existent and similar term. Furthermore, to express the object occurrence in the term, each object may be accompanied by a qualifier such as OPTIONAL (*opt*), REPETITIVE (*rep*) and OPTIONAL-CONDITIONAL (*optc*) precising the condition under which an object may appear.

### 4.3.2 Separators

As the structure is not enough sufficient to characterize the fields and to separate them, the limits between consecutive field are introduced to reinforce the field description. Separators can be *specific punctuation marks* as point, comma, bracket, parenthesis, etc. *Mode changing* (Capital letter in the beginning of the field), numeric area, *font style changing*, etc.

### 4.3.3 Attributes

Because of the weakness of the physical structure and the multitude of choices represented in the model, we add to the previous description some attributes given by the library specification to better precise the description of the reference components.

Several kinds of attributes have been defined, among them, *Type* (string, line, word, char, etc.), *Mode* (capital, numeric, alphabetic, punctuation, etc.), *Style* (bold, italic, standard, etc.), *Position* (beginning of line, inside, end), *Lexicon* affiliation (author index, countries, towns, abbreviations, articles, etc.), *Weight* which specifies the degree of importance of subordinate objects, etc.

## 4.4 Structure Analysis

The structure analysis is based on the model and on the entry data flow. For the model, the grammar rules are converted by a compilation procedure into a working structure. The input data flow is also reorganized into a working table by a filtering task. This table contains useful tokens extracted from the flow such as style, token, size, etc. and a pointer to a buffer containing the corresponding content. Figure 4 summarizes the principal functioning mode of the structural analysis.

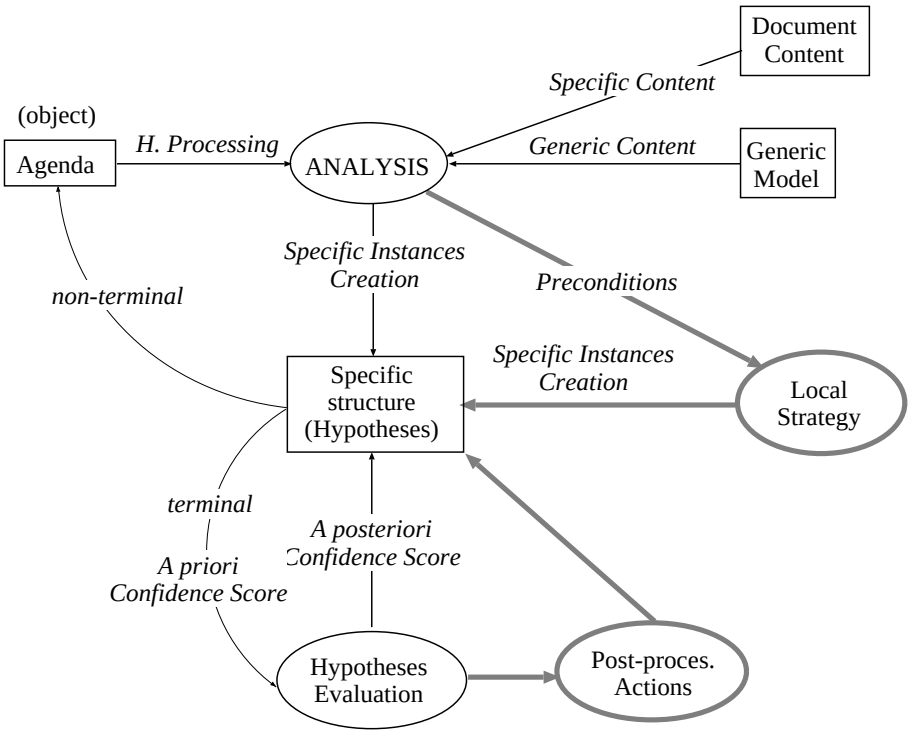



Figure 4: Functioning Scheme of the Structural Analysis.

### 4.4.1 Model Compilation

This step allows to adapt the analysis process to the application model. It generates working files containing the specific terms, actions and attributes for the application. References as well as indexes (containing authors and subjects) are modelized as three different applications. During the analysis, these files are converted into dynamic tables of terms where the entries correspond to term codes. Each term is given by a

list of characteristics gathered in a characteristic table. This allows the system to read rapidly the characteristics of each analyzed term.

### 4.4.2 Hypotheses Management

At each step of the analysis, the system proposes for the current object different choices for its decomposition (analysis). These choices which are not already verified are called *hypotheses*. We use a structural tree to store these hypotheses. A confidence score (*a priori* score) is computed for each generated hypothesis. This score allows to choose, in an *agenda*, among all the current hypotheses which one to process first. The score computing is initialized by the weights given in the model for the current object (for its attributes and subordinate objects). This score is successively updated as the hypotheses are verified and becomes a recognition score. At the end of the analysis, each tree path corresponds to a possible structure (for the input reference) weighted by a recognition score. This qualitative reasoning allows to reduce errors and to isolate possible doubtful areas.

The hypotheses are chosen from the agenda according to the importance of their *a priori* scores (apr). Thus, the analyzer is said to function in an opportunistic mode. Terminal terms (tree leaves) are directly verified. On failure or success, the *a priori* score is up-dated and becomes an *a posteriori* score (aps) which is propagated from bottom to top in the corresponding path (see figure 5).

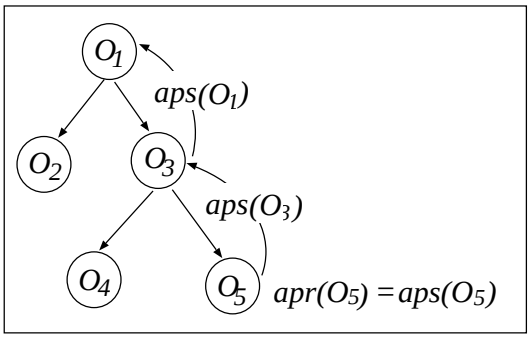



Figure 5: Score Propagation.

The *a priori* score of a current object $o$ depends on the result of the observation

of its attributes ($a_t$) for each token $t_k$ of $o$ ($C(a_t, t_k)$). It is also function of the tokens length ($L$) and of the weight $W$ of each attribute.

$$apr(o) = \frac{\sum_{a_t} \sum_{t_k} C(a_t, t_k).W(a_t).L(t_k)}{\sum_{a_t} W(a_t).L(o)}$$

The *a posteriori* score of $o$ is updated from the *a posteriori* scores of its subordinate objects ($o_i$) by taking into account their corresponding weights ($p$).

$$aps(o) = \frac{\sum_{i} p(o_i).aps(o_i).L(o_i)}{\sum_{i} p(o_i).L(o_i)}$$

With this method, the different objects and attributes influence the final score according to their importance given in the model (weight) and in the input data string (length).

### 4.4.3 Local Strategies

We show here some examples of actions executed before the general analysis. Depending on the status returned by these actions, they can play the role of pre-conditions, in such a case, the analysis continues normally, or of local strategies, stopping general strategy. When an action plays the role of a local strategy, it has the control of new hypotheses (possible decomposition of the current object) to submit.

**Author searching** . In library references studied, it was fitting to identify the secondary authors of the publication. Contrary to principal authors, secondary authors are introduced by a particular expression ("par", "introduit par", "illustration de", etc.). It suffice to recognize this expression and to verify that what follows corresponds to an author. The problem here comes from the fact that authors are not necessarily presented in the same format in indexes and within the references. Furthermore, the list of expressions is not exhaustive. It is fitting to apply a fine syntactical analysis to recognize these secondary authors, as shown by this example:

| | | |
|---|---|---|
| ZATZME | ::= | Seq ZAT ZME? |
| Sep | | Ponct1 |
| Action | | +InitAuteurs(Expressions,IndexAuteurs,...) |

Parameters `Expressions`, `IndexAuteurs`, etc. correspond to a list of lexicons used by the local strategy `InitAuteurs`

**Style Searching** . In order to minimize the hypotheses number submitted during the analysis, we have developed some heuristics allowing to cut an area by searching typographic characteristics. The following example shows an action which cuts the current object at the first punctuation preceeding the beginning of the italic area. This pre-cutting allows, in fact, the analysis part by part and makes the economy of several hypotheses which, in all cases, will failed.

| | | |
|---|---|---|
| ZATX | ::= | Seq ZATZME ZIC |
| Sep | | Ponct |
| Action | | +SplitField(italic,Ponct) |

**Suppression of Useless Hypotheses** . Some objects to recognize are easily identifiables (for example, a town found in town dictionary). In this case, it is interesting to delete all hypotheses in the queue which contain, in an another context, the same search area. The action `KillAmbiguities` in the example below is activated if the object `MotEd1` is perfectly recognized. It goes through the specific structure tree and suppresses all waiting hypotheses that contain the same content as `MotEd1` and that do not belong to other instances of `MotEd1`. This action may be used carefully because every new hypothesis on this area, which is not an instance of `MotEd1`, will be forbidden.

| | | |
|---|---|---|
| MotEd1 | ::= | Terminal |
| Alex | | Edition //opl. tir. uitg. éd, etc. |
| Nature | | mot |
| Action | | KillAmbiguities() RestituteField() |

### 4.4.4 Output Flow Restitution

When the analysis is finished, it is necessary to go through the structure tree to produce a structured flow corresponding to the result. This running is realised depth first. The structure is represented by a mark up format like SGML. Each tagged field is given by a confidence score.

Figure 6 gives the analysis result of the reference given in figure 1. All the sub-fields were correctly localized. They are coded and tagged in UNIMARC. "QSTR" indicates the evaluation score (maximum 10 000).

```
<675 I=bb QSTR=10000> <$a QSTR=10000>159.962</$a> </675>
<200 I=0b QSTR=9834>  <$f QSTR=9487>Gérard Liger-Belair</$f>
       <$a QSTR=10000>Je suis fakir</$a> </200>
<700 I=b0 QSTR=10000> <$a QSTR=10000>Liger-Belair</$a>
       <$b QSTR=10000>Gérard</$b> </700>
<210 I=bb QSTR=9705>  <$a QSTR=10000>[Verviers]</$a>
       <$c QSTR=9519>[Editions Gérard & Cî]</$c>
       <$d QSTR=10000>[1973]</$d> </210>
<215 I=bb QSTR=9750>  <$d QSTR=7353>32f carré</$d>
       <$c QSTR=8601>couv., ill.</$c>
       <$a QSTR=10000>158 p.</$a> </215>
<010 I=bb QSTR=10000> <$d QSTR=10000>30 BEF</$d> </010>
<225 I=2b QSTR=10000> <$a QSTR=10000>Harabout-flash</$a>
       <$v QSTR=10000>352</$v> </225>
<517 I=0i1 QSTR=10000><$a QSTR=10000>Souvenirs, révélations,
       conseils</$a> </517>
<900 I=bb QSTR=9772>  <$a QSTR=10000>B.D. 14.814 352</$a>
       <$b QSTR=9285>73-2108</$b> </900>
```

Figure 6: Structural Analysis Result of the Given Reference.

In the event of errors, the system generates a fictive UNIMARC code 903 which it uses to demarcate the zone it should have recognized for a field but which does not quite fit the characteristics as specified by the user. This helps in modifying the model to take care of exceptional cases or to really determine that the reference was badly

formed as a result of OCR errors, the printers devil or outright bad transcription of the reference.

When the system finds more than one solution for a given zone, it equally generates a fictive UNIMARC code 902 that it puts around each of the possible solutions which are then presented to an operator who has to make a choice.

## 4.5 Results and Discussion

The global evaluation of the prototype is made up after the treatment of all the 11 catalogue volumes of the Belgian Library, e.g. 4548 references. The volume of june is discarded because it was used in the first phase for the control quality evaluation. Performances will be discussed in the following points:

- *OCR/ICR.* 6.69 doubts per reference for only the body of the bibliography and 9.87 doubts per reference if include the rest: main and secondary entries.

- *Structure Recognition.* 67% of references have been recognized automatically by the system.

- *Attribution of language and country codes.* 77.7% of references have their codes created automatically by the system.

  However, considering all the operations of correction provided for the automatic structure and code generation, as well as the corrections effected on references with "risk", only 47.5% of references have been entirely recognized automatically without any manual intervention.

- *Speed up.* The speed up of the prototype is about 1'30 per notice. This depends on the complexity of the structure and the correction procedures launched by the system.

  The table 1 give the time spent by the system for the different modules for all the 4548 references.

| Automatic Module | Time in hours | % total time |
|---|---|---|
| OCR/ICR | 16.5 | 14% |
| Structure Recognition | 99 | 83.5% |
| Others | 3 | 2.5% |

Table 1: Time spent by the Automatic Processing.

**Manual Intervention**

The table 2 gives statistics on manual interventions either for OCR correction or for re-treatment of the structure or the codes generation.

| Module | Defect Cases | manual interventions |
|---|---|---|
| OCR/ICR | 44920 doubts examined | 9.87 doubts per reference |
| Structure | 1494 references unstructured totally or partially | 33% of references |
| Codes Country Language | 1014 references with in less one non-generated code | 22.3% of references |
| Structure + Country Codes + language | 2083 references corrected in less one time | 52.5% of references |
| Anomaly after Quality Control | 246 references returned to the Library | 5.4% of references |

Table 2: Statistics on Manual Interventions.

### 4.5.1 Problems encountred

The main problems encountred in the technical realisation of this project concern the treatment by OCR of the bibliographic information, the structure modelling of the

Library catalogues and the moving to an industrial production.

**OCR and Bibliographic Information** . The variability of the typography seriously handicaped the straightforward conversion of the bibliography by OCR techniques. Many reasons have been signaled in section **??**. The main deficiencies encountred in the Belgian catalogues are:

- *Typographic aspects*: connected characters for bold data and use of standard numeric characters within textual areas in italic;
- diacritics added by hand,
- use of long dash line for all the parallel areas and for someones of the collection sub-areas;
- intensive use of square brackets.

3.6% of references were returned to the Library because of the presence of non latin charaters to transliterate by the Library.

**Bibliographic Catalogue Modelling** . This problem is already encountred in a traditional retrospective conversion process in which the Library writes specifications for the conversion of its catalogue. These specifications must be validated on several references and modified in a continuous manner in order to adjust the model in order to take into account the exceptions and new encountred problems. In the MORE project, these specifications were very detailed but with a point of view oriented more for the cataloguing than for the automatic conversion by computer. This needed a more adaptation of the two populations (from Library and Laboratory) to better harmonize their dialogue.

In the other hand, the structure of the bibliographic information is very difficult because of this three main characteristics:

- Catalogues are written before the apparition of the standard ISBD, leading to different structures with particular rules for layout and punctuation;
- The correspondance between the pré-ISBD cataloguing rules is sometimes difficult to establish with the UNIMARC format for the transcription of the titles areas and responsability mentions. This difficulty is lower in USMARC where the main cataloguing elements are grouped into three sub-areas non-repetitive in only one possible sequence. In UNIMARC, the same information can be shared in six sub-areas all of them repetitive and with a high number of possible sequential combinations. The same difficulty is encountred in the modelling of the edition and collection areas.
- Some catalogues has a cataloguing with hierachical levels in the case of the monography in severals volumes, with significant titles for each volume. The model has to take into account some specific considerations for the treatment of these volumes. In the Belgian Library, the cataloguing of volumes belonguing to a monography in many volumes, as well as the treatment of collectif titles was very difficult to model and had created some anomalies returned to the client (the Belgian Library). The presence of many official languages in this bibliography (French ans Dutsh) have also led to a great number of parallel mentions in titles and notes, with a very complex structure of titles and responsability mentions. 78.4% of the 246 references have been returned for manual control because of bad structure, 46.53% for the title structure and 12.25% for the validation of collectivies authors.

# 5 Conclusion

The aim of this paper was to enhance understanding of the issues involved in the retroconversion process and to show the advances in the field of character recognition and structure interpretation and their usefulness in the development of solutions to the retroconversion problems.

The system presented here gave good results on tested library references. The errors encountered were due to incomplete specification (reference not falling into any of the categories we were provided information on) or OCR errors. The ambiguities encountred were partly due to a combination of incoherence in the specification (which allows different legal segmentations) as well as OCR substitution errors. The evaluation allows the observation of the quality of each reference and each field in the reference and allows the user to intervene or not for manual correction.

# References


[1] Beaumont J., Cox J. P.: Retrospective Conversion. A practical Guide for Libraries. Meckler, Westport/London. 1989. 198 p.

[2] Belaïd A., Chenevoy Y., Anigbogu J. C.: Qualitative Analysis of Low-Level Logical Structures. In *Electronic Publishing EP'94*, volume 6, pages 435–446, Darmstadt, Germany, April 1994.

[3] Belaïd A., Chenevoy Y.: Document Analysis for Retrospective Conversion of Library Reference Catalogues, ICDAR'97, ULM, Germany, August 1997.

[4] CEC, DG XIII B: Libraries Programme, Telematics Systems in areas interest 1990-1994: Libraries, Synopses of Projects. http://www2.echo.lu/libraries/en/libraries.html

[5] Council of Europe: Guidelines for Retroconversion Projects prepared by the LIBER Library Automation Group, Council of Europe, Council for Cultural Co-operation, Working Party on Retrospective Cataloguing, 1989.

[6] Crawford R. G., Lee S.: A prototype for fully Automated Entry of Structured Documents. In The Canadian Journal of Information Science, (15)4, pp. 39–50, 1990.

[7] ISBD (G): General International Standard Bibliographic Description: Annotated Text. Prepared by the Working Group on the General International Standard Bibliographic Description set up by the ILFA Committee on Cataloguing. London, 1977. 24 p.

[8] International Standard Organization: Information processing, text and office systems, standard generalized markup language (sgml). Draft International Standard ISO/DIS 8879, International Standard Organization, 1986.

[9] ISO 8859-1 to 7: Information Processing - 8-bit single-byte Coded Graphic Character Sets - Part 1-7: Latin Alphabet No. 1 to 7. International Standards Organisation. 1987.

[10] Lib More: Marc Optical Recognition (MORE), Proposal No. 1047, Directorate General XIII, Action Line IV: Simulation of a European Market in Telematic Products and Services Specific for Libraries, 1992.

[11] Schottlaender B.: Retrospective Conversion: History, Approaches, Considerations. Haworth Press, NY. (1992).

[12] Süle G.: Bibliographic Standards for Retrospective Conversion. In IFLA Journal (16)1, pp. 58–63, 1990.

[13] Valitutto V. and Wille N. E.: A Framework for the Analysis of Catalogue Cards. FACIT Technical Report no 2). Statens Bibliotekstjeneste, Copenhagen. October 1996.

# TopicMark

A Topic-focused Bookmark Service for Professional Groups

Hui Guo and Hans-Ludwig Hausen
GMD - German National Research Center for Information Technology
FIT - Institute for Applied Information Technology
Schloss Birlinghoven
D-53754 Sankt Augustin, Germany
{hui.guo, hausen}@gmd.de

## Abstract

Information overload emerges as a critical problem for information gathering as the resources on the Web grow rapidly in both scale and complexity. Networked information discovery aims to assist people's search tasks with centralized information services. User interests and user relevance judgements of Web pages with regard to these interests can be used to improve the resource discovery process. Based on this observation, an interest-topic concept model is presented. It illustrates how group processes can be enhanced with support for cooperation between users with various professional affiliation but similar interests. The design of a domain-independent bookmark service– TopicMark– is then presented that specifies the specific topic generation process, and the autonomous aggregation of information resources.

**Keywords:** resource discovery, topic, user interest, group process, bookmark collection, social agent.

## 1. Introduction & Motivation

As the Internet and the World Wide Web grow at an explosive speed, people are bothered by information overload and lacking structure when searching for interesting information sources. Information resource discovery emerges as a topic of active research with the goal of finding solutions that facilitate people's information gathering process and that meet their information needs. On the Web, various specialized collections, bookmark list pages, searchable catalogues such as Yahoo, information discovery systems such as Harvest [3] and recommender systems [13] were created to collect Web resources in one place to ease access of information. Search engines appeared as a trial to index the resources on the Web and provide retrieval services for the general public.

However, easy-to-use tools that support the users' information gathering tasks with minimal overhead are still lacking. The systems mentioned above are usually ineffective to provide relevant results with high precision for users' queries. Factors such as massive storage requirement, Web page spamming, up-to-date validation of index and ranking criteria make it difficult for search engines to keep pace with the growing Web and to provide high-quality results with regard to users' requests.

More importantly, in a professional environment, e.g. a research institute for computer science, people are accommodated to use a particular terminology (or language) when conducting information gathering tasks. This is especially true for professional groups as these groups always develop their own terminology to ease communication; the situation is similar in leisure time groups, where gangs and cliques always have their own languages. An effective personalized tool for information gathering, therefore, should allow users to use such private terminology or language. For the professional environment standardized term lists or thesauri can be used; for most hobby groups or private interest groups, their own bylaws or terms of reference can be adopted. On the other hand one might want to construct a term list or thesaurus for a particular domain or search request. Therefore termlist creation or thesaurus construction should be supported in a useful information discovery tool.

General resource discovery tools such as search engines, however, are using languages different from that of the users (or the group the user belongs to). Users don’t know how to formulate queries effectively to get better results. These tools focus on pro-

viding results with high recall but low precision, and focus on the interests of the search service provider (e.g. collect and sell information on the search engine user) without providing means to control operation or results with respect to user needs or interests. In addition, the contextual information of user search requests are not considered in the search process, thus many irrelevant results are presented to user making these services hard to use for a serious search task.

Another aspect of the information gathering process is, that people not only search information individually, but also as members of particular groups in loose or tight cooperation. To support the special information needs in a collaborative environment, an information gathering service that allows the information discovery process to be controlled by users, needs to be developed so as to support seamless integration of information systems with the resources dispersed on the Web. Wrapping search engines or digital libraries are considered ineffective to meet such needs.

This paper presents an approach to information discovery which utilizes the contextual knowledge of users interests, activities and collaboration with others. Based on this approach, web resources are harvested regarding people's shared information needs. Document indexes of the collected resources are constructed and associated with the shared user interests which we call topics. Vector space model-based retrieval can be performed within the topic-related document collections regarding the context of users' queries. The similarity among users' interests, documents' content and groups' topics are explored to support advanced information discovery processes. Such processes can be navigation in document collections by browsing topics, or recommending persons with matching interests.

There are some other ongoing research efforts adopting related technology. Subject-based search engines such as Yahoo deploy subject-focused organization of documents, but are mainly based on predefined categories or those generated from document classification. Knowledge management and search system such as Verity[9] provide means to organize intranet documents regarding user-defined categories, but lack extensibility to tackle Internet resources. Various bookmark organization tools[14] allow users to attach personal bookmarks to defined categories with little support of searching relevant documents on the Web. In the area of information retrieval, user profiles are studied and mainly used for query extension or relevance feedback. Text mining tools such as IBM[4] apply feature extraction and hierarchical clustering to sample documents so as to provide an integral view of document content without taking into account users' personal interests. Various agent-based applications [6][11] are created to support personal information discovery of Web resources by exploring user profiles.

In order to provide relevant results with both high precision and recall value, we explore the usefulness of implicitly/explicitly capturing shared interests of users in different contexts, and the construction of topic-structured collections. An Interest-Topic model is thus presented as a foundation for a "meta-search" approach to access information on the web. It is integrated in a system for "collaborative sharing and discovery " of bookmarks-TopicMark. We also examine how it can be used to augment collaboration.

Our project at the CSCW group in GMD was set up to create an open distributed platform[17] supporting various Web-situated applications using software agent technology. Under the scenario of the Social Web Programme[1], the TopicMark system built upon this platform is designed to fulfill people's various information needs during a collaboration process. It is designed to be used with low overhead, to be domain independent and self-organized, and to be an add-on service for various collaboration applications that access resources on the Web.

## 2. Topic Model & Group Process

This section describes the Interest-Topic model and investigates the usefulness of the model to support group processes.

### 2.1Users' information needs in a collaborative environment

When people work in a collaborative environment, they usually have different information needs than when working individually. In the latter case, it might be unclear what kind of information to look for and what are the proper questions to ask at the starting point. People may at first simply browse through Web pages, following interesting links and gradually come up with some specific questions or queries.

As figure 1 shows, in a collaborative context, people have very specific information needs. These needs arise from the interaction occurring during their cooperation work and closely relate to the topics they

1. URL:http://orgwis.gmd.de/projects/socialweb/

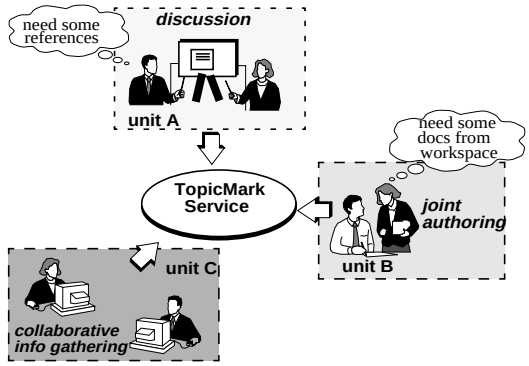



Figure 1. Information needs in collaborative environments

are focusing on. The critical requirement for such a search request is that highly relevant results are delivered in real-time so that cooperation can smoothly proceed rather than be interrupted or delayed. An information service within or across the collaborative environment should contain compact information resources relevant to people's current tasks and interests so as to meet their specific requests. In addition, people sometimes distribute the search task among each other and concentrate on particular aspects. Presenting information relevant to the overall search context will facilitate the collaborative gathering process. To this end, users' shared interests and the context can be captured to support their information gathering tasks.

## 2.2 Interest-Topic model

To realize the information service mentioned above, one critical problem has to be solved: organize the documents in a way that highly relevant documents are selected for a user's particular search interests. Traditional statistical classification of documents is not feasible here since it is mainly dependent on the statistical characteristics of the document collection that may be different in different users' view of the contents. Thus, search results based on this method are very likely to be irrelevant with respect to the users' real interest.

The proposed Interest-Topic model is shown in figure 2. The major concepts are:

- capturing users' personal interests in a collaborative environment.
- extraction of shared user interests as topics.
- categorization of documents to topics based on content relevance.

Within collaborative environments, users' personal interests (long-term/short-term) and special needs are captured. Internally they are formalized as text vectors. Based on user-selected terminology or community language such as ACM Classification System for expressing interests, these interests are aggregated and clustered. Shared interests are extracted as topics and normalized by a standard thesaurus.

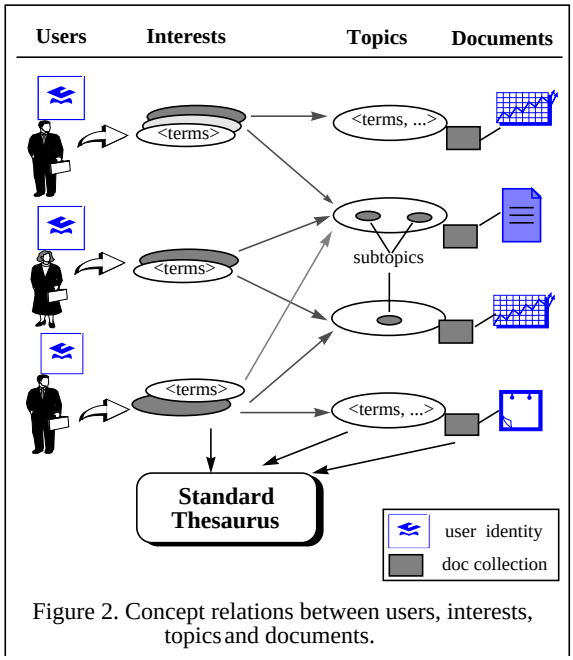



Figure 2. Concept relations between users, interests, topics and documents.

Topics are also formalized as text vectors. Regarding these generated topics, relevant web pages on the Web are harvested, collected, indexed and ranked as sample documents relevant to the topics. By analyzing the content of these web pages, important terms are extracted based on term occurrence frequency and added to topics. Incoming documents are categorized to existing topics and topic-controlled harvesting is continued. According to this approach, gradually users, user interests, topics and documents are associated to each other as illustrated in figure 2.

Documents are categorized according to the users' shared interests. When a user issues a search request, the contextual information of this request is matched to registered topics; the query itself is matched with the documents associated with those topics. Results generated in this way are expected to be relevant since they match the query based on content, and the categories that they are associated with match the context of the request. All these matching processes adopt a correlation coefficient function, e.g. cosine function or distance measure[16].

### 2.3 Specification of user interests

By user interest, we mean the information needs aspects of a user's work. These needs constitute the context of users' search requests. One example is the need to search some information relevant to the topics users are concerned about. To capture such needs, the three-dimension user model [15] is adopted as. Based on that, users' short-term/long-term interests both as an individual and as members of a group can be captured implicitly.

User interests can be indicated by the user's membership within various physical organizations. Firstly users' professional affiliation in the real world represents the scenario of their interests as a whole, e.g. a computer scientist's major interests fall into the computer science area. Users' institutional affiliation can serve as a symbol of interests, e.g. a researcher in GMD-FIT mainly works in the HCI or CSCW area. In addition, when people use particular collaborative applications to assist their work, the user identity in a group also reflects interests in particular contexts, e.g. a person invited to a "Social Web Program" workspace in the BSCW shared workspace system [2] should have interests in that research direction. Such characteristics of a user are the basic part of their profiles when they join a collaborative environment, and thus can be directly captured.

On the other hand, users' search interests are usually created as queries which specify the keywords or phrases to appear in the document. These queries represent a snap shot of the user interest at a given time. In the case where users are offered some means to store or organize the queries together along with the collected documents, e.g. a Netscape bookmark folder, the context of each query can be represented by the container name, depending on the way that queries and documents are grouped and the normalization to name the containers. Usually users attach several highly-rated documents as very relevant documents of particular interests. The content of these documents also behave as a description of the interests. Using query-routing technology as introduced in [10], important terms can be extracted and treated as descriptors of the user interest. Therefore a user interest consist of three element-queries, tasks, and relevant documents.

A user interest is represented as an object with some attributes: UserID, semantic context, denotation, attached documents, and timestamp. One user may have multiple interests. Each aspects of user interests are captured in the user interface and are transformed into this standard representation without losing semantic consistency. In this definition, UserID refers to the user owning this interest. Semantic context means the context in which a user delivers the search request, e.g. the task he/she is doing, the content of the document he/she is authoring. Denotation is defined as list of descriptors or terms corresponding to the user's queries. The timestamp or other attributes are used to indicate changes in the interest and thus trigger the topic-formulation process. An example in the experimental TopicMark system is:

```
Angi's interest ::=
   { ANGI, {"GMD-FIT",  "social web"},
   {{"virtual places", "avatar"},
    {"social construction of knowledge",
      "conceptual index"},
    {"collaborative filtering",
      "recommender system"}},
   list of high-rated documents, Mar.98};
```

### 2.4 Aggregation of interests into topics

A topic is defined as the shared aspects of a certain number of users' interests. Thus it mainly captures the collaborative aspects of the users' work. It can be identified by the frequently co-occuring terms that appear in the users' interests. Since a few community languages or public dictionaries such as FOLDOC [7] are available, only those user interests expressed in a same language are clustered to generate topics and topics are thus grouped regarding the language. Once generated, depending on whether it exists in standard thesaurus or appears in user interests as an agreed phrase, a topic is registered in existing topics group.

A topic is described in a similar way as interests:

```
Topic :: = {Users, aggregated context,
denotation, listof subtopics, attached
documents, timestamp};
```

The subtopics of a topic can be derived by looking for partially shared interests. Subtopics are also made uniform using a standard thesaurus or a user-controlled vocabulary, e.g. words appearing in users' interests as a phrase statistically. Each topic generated is associated with a collection of documents according to the relevance calculated as a correlation between the topic's description and the contents of the documents. Once a new topic is generated, the system can either attach existing documents to the topic, or initiate a documents collecting task and start content-based attaching process.

A topic can be reformulated by extracting high-frequency terms in the attached documents as additional terms in the denotation. It could also be revised once users change their interests or define new interests. A fine-grained topic structure can be defined by aggregating structured personal interests. More specifically, topics can be classified as topics for a group, or topics for several groups sharing interests.

### 2.5 Augmentation of collaboration

The TopicMark solution for information gathering introduces several important ways to support collaboration. Basically, since information resources on the Web are collected and organized regarding users' interests, TopicMark offers an easy-to-use central service matching users' requests accurately. Secondly, users' shared interests are captured and it is possible to classify users into groups and to bring people together or to recommend person with similar interests, e.g. recommend a technical expert. The internally identified groups can also help users cooperate across physical group boundaries. Furthermore, by exposing the generated topics to the users and by allowing them to manipulate or navigate through the topic space, users are offered a good chance of mutual learning, information discovery and constructing knowledge of their interest.

The topics generated in the TopicMark system are adapted dynamically according to the changes in users' interests so as to capture users' behavior. It is possible to build personalized agents for individual users, and socialized agents for the groups. Mediated by TopicMark system, the cooperation between various kinds of information systems on the Web can be facilitated.

To explore the usefulness of the model, we conducted some experiments using the raw data in a recommender system-LiveMarks[8]. The results shows that user interests show a lot of common aspects, thus making it feasible to construct topics. However it also exhibits the necessity to specify structure of topics. We intend to make use of all the web pages in GMD intranet indexed by TopicMark robot to achieve more understanding.

## 3. An information gathering service for groups – TopicMark

This section gives a detailed description of the TopicMark service based on the topic model presented before.

### 3.1 System functionality

TopicMark is based on vector space model[16] for information storage and retrieval. It provides:

1. Users or group-specific indexing of web resource;
2. Interest and topic-focused search and retrieval;
3. Enrichment of web resource description and automatic updation of resource’s validity.

It supports above features by registration of users, groups and web resource with interests or group topics, and doing indexing of these artifacts via user-specific term lists or thesauri.

More specifically, in TopicMark user requests and web items are registered with both interests and topics. For the indexing (i.e. the mapping of the raw web item or the request onto the set of descriptors) user-selected term lists or mature thesauri are used. In an initialization phase one can use a standard term list or thesaurus (e.g. ACM computing reviews thesaurus, ISO multilingual term list) or a dedicated subset thereof. It is assumed that indexing with user/group-approved thesauri improves user satisfaction as well as recall and precision. A request is processed by matching it against the user’s interests or topics or both or (if that match was not successful) approaching the web via public search engines and feeding the results into the interests and topics. TopicMark updates its data base on interests, topics and associated web items on defined time intervals and concurrently tests the data on these web items. This ensures availability and validity of the marked web items. Depending on user defined preferences users are notified on updates and changes on web items as well as on other users or on groups. As a consequence users participating in TopicMark will be aware of users, groups, interests, topics as well as on available and valid web items.

TopicMark can be extended by introducing other functionality. For example, the contextual information of users’ interests as well as particular web resources can be explored so as to match documents with users’ request more comprehensively and present results in a more understandable way; a group matcher can be constructed by specifying shared user interests to support group-specific information fathering. A full integration paradigm can be applied to integrate conceptual index [17], TopicMark and recommender LiveMarks.

### 3.2 User interface

The screendump of the TopicMark user interface is shown in figure 3. TopicMark is currently integrated with a recommender system as an add-on information service. TopicMark provides several places for users to conduct collaborative information gathering.

- Tasks page: a place to join predefined groups and express search interests in task level;
- Results page: a place to input queries, browse URLs and share annotations;
- Topics page: a place to access the captured topics;
- Search interface: a place to conduct context-related search using ACM Topic Tree .

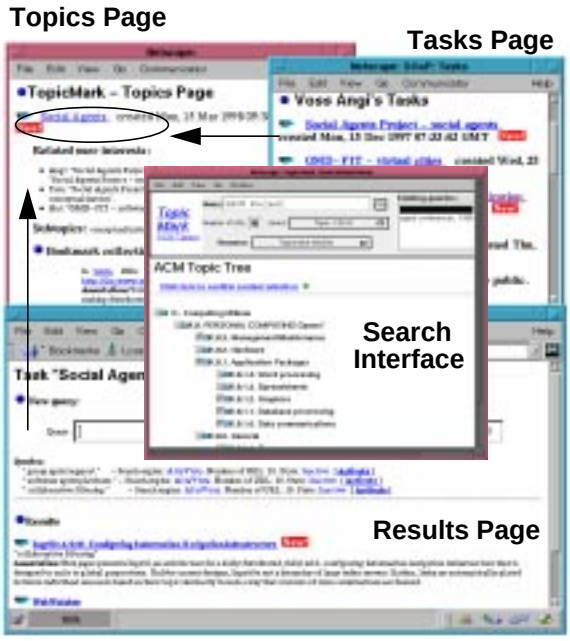



Figure 3. TopicMark User Interface

Under user-specified privacy constraints, the tasks and queries in the first two places are captured by TopicMark and are used to maintain the place for topics. As figure 4. show, the topic *Social Agents* are generated out of user Tom's personal interests and user Voss Angi's interests represented by the *"Social Agents Project - social agents"* task, queries such as *"group agent support"* and *"software agent platform"*, and the highly-rated bookmarks within results page. The topics page provides a central view for users to be aware of the activities occurring within/outside the groups by navigating through the topics and the associated information such as user interests, subtopics, and bookmark collections.

TopicMark also automatically matches users' queries against the topics and bookmark collections and puts relevant bookmarks into users' particular task page as retrieval results. In addition, in TopicMark each artifact as a part of user interests has an event indicator, e.g. when a new topic is generated regarding registered queries, the indication icon *NEW* will be highlighted in the topics place to inform the users of the event. Such events can also be "pushed" to users in order to support real-time service.

### 3.3 Topic, shared interest and user interest

We intend to develop TopicMark as an information gathering system for specialized groups. The system constructs bookmark collections of reasonable size for a group's common needs which are represented as topics. The key points are identification of the topics out of individual users' interests, and during the construction procedure taking users' personal view of the relations between bookmarks into consideration. This is dependent on the specific user model and group model of applications. For instance, in LiveMarks, users express their interests in the form of "groups-tasks-queries-bookmarks". In the collaborative applications like BSCW[2], users are able to formulate queries, review results and organize the information into folder-based shared workspace. Thus, the representation of users' interests varies with different user interface. From the retrieval point of view, however, they can be generalized as presented in the model. The denotation can be lists of descriptors. In such lists each descriptor carries a weight indicating its relevance with the particular interest.

A topic is then defined as the shared interest within a group or across groups. It can be identified by clustering user interests or text mining approach. Topics are dynamically created and updated according to the changes in users' interests which are also captured dynamically, e.g. based on a periodical schedule or on users' demand. Both user interests and topics are made uniform by structured descriptor lists or thesauri, or using an ontology. For each topic, a user group is defined associating users to topics so as to form a basis for retrieval. It could also be used to recommend users with similar interest by calculating similarity of user interests, or to support collection-mediated collaboration.

### 3.4 Bookmark collection and repository

For each topic, relevant web pages are collected to form a topic-specific collection. An agent-based crawler generates a set of queries out of the topic's description and initiates gathering process. It wraps meta search like search engines to acquire preliminary materials. From a recommender service, it explores socially constructed knowledge to obtain highly-rated URLs. Then it performs link following to fetch the original web pages. These Web pages are filtered by a wide range of criteria, e.g. spamming, semantic redundancy, hostname alias, syntax error, , robot exclusion protocol, etc. Additional information concerning a Web page such as hyperlinks, meta content and lastModifiedDate, etc., are extracted as its properties to enrich its description based on the statistics introduced in [1].

Each fetched Web page is parsed into a uniformed text vector of <descriptor, weight> and internally represented as a bookmark with properties like title, excerpt, lastModified, etc. The weight of a descriptor is derived based on the relative term occurrence frequency. A bookmark is then created for each Web page. The definition of bookmark is intended to be compatible with standards of web content such as MCF and RDF [5]. It is classified regarding existing topics or can be used to generate new topics. In this way each topic gradually accumulates its own collection of bookmarks with varying relevance in content. The frequently appeared descriptors within the bookmarks are extracted and put into the topic's descriptor lists to extend the topic description.

All the collected bookmarks are stored within a single repository. The repository deploys an efficient storage scheme and develops maintenance services to guarantee the validity of the bookmarks. A bookmark storage and access architecture is designed to support database-transparent access of the bookmarks collections in an object-oriented way, and to enable interoperation between repositories, e.g. exchange of bookmarks.

### 3.5 Indexing, classification and retrieval

To normalize the vocabulary used to express users' interest, topics, and bookmarks, published glossaries, e.g. ACM Computing Reviews Dictionary, are taken as standard termlist. A larger scope of vocabulary for queries and the indexed bookmarks is supported by extracting the words from the collection itself to construct topic-specific thesaurus. Collected bookmarks are stemmed and indexed by the same set of descriptors. A varied inverted file organization is constructed for the descriptors, i.e. each descriptor in the dictionary is related to a group of relevant topics each of which is associated to a group of bookmarks. In this way those bookmarks that don't contain the descriptor still can be recommended as relevant documents for a particular user's queries, whereby user s are able to get good results even for poorly-formulated queries. Thus a better recall/precision value is achieved. Similarly hyperlink indexes are also generated to support full-fledged retrieval.

To provide better performance, a topic can be splited and the corresponding bookmark collection can be reclustered. The classification scheme is based on statistical term association [16]. All the descriptors are compared and high frequency keywords are extracted either to extend the topic or to identify a new topic. Each bookmark is then analyzed to determine which topic it belongs to. The classification can be overlapped, i.e. one bookmark may belong to multiple collections. On the other hand, bookmarks can be classified based on the form rather than the content, e.g. entry page, bookmark list page in order to provide user with an option of intended Web page types.

When a user issues a new query, it is parsed into vector model-based representation and matched with collected bookmarks according to a two-step retrieval scheme. It firstly matches with relevant topics then the bookmarks within the collection designated by the topic. The contextual information of a query is also delivered to match more topics for which the user hasn't register yet. The matching algorithms use the correlation coefficient, e.g. cosine function or distance measure [16].

### 3.6 Coordinated agents for TopicMark

Since social agent technology[17] is very fit for the autonomous work needed by the TopicMark system, we decide to select it as the base technology for implementation.

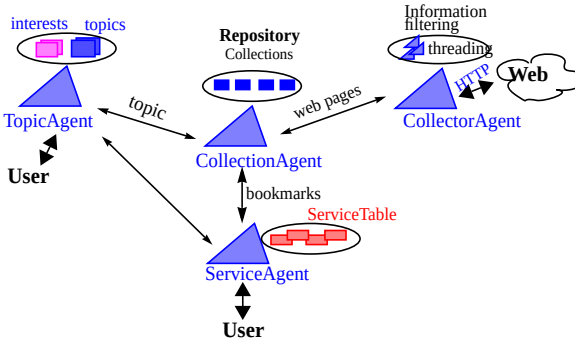



Figure 4. Agents' interaction in TopicMark.

As figure 4. shows, the agents in TopicMark are specialized with regard to behavior and function. Each type of agent uses the knowledge of its environment and has specialized task to accomplish. They plays different roles in the overall system. For instance, TopicAgent is designed to formulate and manage topics and it cooperates with CollectionAgent in order to build collections for particular topics. They interact with each other by exchanging messages of certain types ("performative"). The interaction- synchronous or asynchronous--can be conceived of as a conversation. Such conversation patterns may be formalized as finite-state-machines or in a distributed environment as high-level Petri nets [19]. The FSM specification of each conversation can be described by conversation tables which specify the state transitions as well as the message transferred in particular states. They may be used to formally verify that the conversation is free of deadlock even in the presence of message delay and mixed initiative of the conversation partner (not a simple turn-taking protocol). In addition their activities are monitored by a coordinator agent (not shown in the figure). It synchronizes the state of TopicMark agents so as to ensures the reliable and predictable behavior of the whole system.

## 4. TopicMark implementation

As a service allowing groups to accumulate topics and bookmark collections, TopicMark is implemented on top of an agent open infrastructure [18] as an application service. The infrastructure is a distributed platform for interacting social agents providing a runtime environment with system functionality such as an agent naming/addressing schema, an agent messaging mechanism, a directory service, etc. This platform consists of the kernel modules–an agent engine and system communication services. Both TopicMark and the platform are implemented in Java.

TopicMark is implemented as a domain-independent service. Various collaboration system can integrate it by following the protocols specifying the user interest and retrieval interface. It has basic architectural support such as request scheduling, transaction process and persistence support. By distributing functionality over network, the scalability of the service is expected to be guaranteed. In addition the information exchange protocol between individual TopicMark system and other information service such as a recommender service like LiveMarks, is specified to support the system's interoperability.

The first prototype is developed and experimented as a service in LiveMarks within our research group at GMD. It includes agents which manage topic formulation, build bookmark collections, and mediate service requests. This prototype is intended for demonstration, exploratory use, and evaluation in cooperation with an industrial partner from the oil business. Prospective users are members of project teams operating in oil field development. Team members with different background respectively belong to special professional groups. They are usually spread around the world, and may belong to several teams at the same time. Information retrieval and exchange is central to their work. Construction of indexes of the Web resources relevant to their interest and preference is essential to meet their needs and facilitate the information gathering process.

## 5. Conclusion

This paper presents a user-centric approach for collaborative information gathering. An interest-topic model is described briefly. It focuses on exploiting the shared aspects of users' work. Base on the model, a TopicMark system is developed to support collaboration within/across groups through information sharing. The presented approach represents a novel way for resource discovery on the Web in order to meet people's information needs within various collaborative environment.

Although TopicMark represents a possibly promising solution for collaborative information gathering, there are some important issues to investigate in order to build a useful information service. First of all, the personal aspects of users' interest might be lost in the topic-construction process and make it difficult to support users' information gathering as individuals. Building personalized topic structure with regard to each user's preference is a possible solution for this problem. Secondly, since users may use a different vocabulary to express their interests, it is critical to integrate these vocabularies in order to ensure topics to have an unambiguous meaning.

The distributed bookmark services presented in this paper form a network of topic-focused bookmark collections and emerge as a valuable information resource for users to access through popular Web browsers. In the future, we propose to integrate TopicMark service with BSCW shared workspace

system [2] to achieve more understanding within extensive collaborative contexts. Some important issues to be investigated include user feedback, topic-reformulation, media migration and so on.

## 6. References


[1]Allison Woordruf, et al. "An Investigation of Documents from the World Wide Web," in Proc. 5th Int. WWW Conf. May 6-10, 1996, Paris, France.

[2] Bentley, R., Appelt. W., Busbach, U., Hinrichs, E., Kerr, D., Sikkel, K., Trevor, J., Woetzel, G. "Basic support for cooperative work on the World Wide Web," Int. J. Human Computer Studies 46 (1997), 827-846.

[3]C. Mic Bowman, Peter B. Danzig, Darren R. Hardy, Udi Manber and Michael F. Schwartz "The Harvest Information Discovery and Access System," Computer Networks and ISDN Systems 28 (1995), 119-125.

[4]Daniel Tkach, "Text Mining - Turning Information Into Knowledge, A White Paper from IBM". http:// www.ibm.com. Sept, 1998.

[5]David Singler, et.al. "Resource Description Framework (RDF), "WD-RDF-Schema-1998011, W3C 1998.

[6]Eui-Hong(Sam) Han, Daniel Boey, Maria Gini, Robert Gross, Kyle Hastings, George Karypis, Vipin Kumar, Bamshad Mobasher, and Jerome Moore, "WebACE: A Web Agent for Document Categorization and Exploration," Proceeding of Agents'98. 1998.

[7]FOLDOC-Free On-Line Dictionary of Computing, http:/ /www.easynet.de/resources/foldoc/index.html

[8]Guo, H., Kreifelts, Th., Voss. A. "SOAP: Social Filtering through Social Agents," ERCIM Workshop report No.98/ W001. 5th Int. DELOS Workshop: on Filtering and Collaborative Filtering Budapest, Hungary, 10-12 November, 1997.

[9]J.O. Pedersen, C. Silverstein, C.C. Vogt (Verity, Inc.) "Verity at TREC-6: Out-of-the-Box and Beyond,", Overview of the Sixth Text REtrieval Conference (TREC-6). 1997.

[10]Julian A. Yochum, "Research in Automatic Profile Generation and Passage-Level Routing with LMDS", In D. Harman, eidtor, *Overview of the Third Text REtrieval Conference (TREC-3)*, pp-289-297.1995

[11]Liren Chen, Katia Sycara, "WebMate: A Peronal Agent for Browsing and Searching,", Proceeding of Agents'98. 1998.

[12]Peter Pirolli, Patricia Schank, Marti Hearst, Christine Diehl, "Scatter/Gather Browsing Communicates the Topic Structure of a Very Large Text Collection," Proceeding of CHI'96. 1996.

[13]Resnick, P., Varian, H.R. "Recommender Systems, " Comm. ACM 40, 3(1997), 56-58.

[14]Richard M. Keller, et.al, "A Bookmarking Service for Organizing and Sharing URLs,", Proceeding of Sixth International World Wide Web Conference. 1997.

[15]Robert B. Allen, "User models: theory, method, and practice", Int. J. Man-Machine Studies 32 (1990), 511-543.

[16]Salton, G. "Automatic Text Processing", Addison-Wesley, Reading, MA, 1989.

[17]Voss, A., Kreifelts, Th. "SOAP: Social Agents Providing People with Useful Information," in Proc.Int. ACM SIGGROUP Conf. on supporting group work, ACM, New York NY, 1997, pp-291-298.

[18]Voss, A, Guo, H, Hausen, H.-L., Juhnke, M, Nakata. Kreifelts, Th., and Paulsen, V.(1998) "Agents for Collaborative InformationExploration", in Proceeding of the Third International Workshop on CSCW in Design (CSCWID'98), Tokyo (to appear).

[19]Woetzel, G., Kreifelts, TH. "Deadlock freeness and consistency in a conversation system," in B.Pernici, A. A. Verrijn-Stuart (eds) Office Information Information Systems: The Design Process, Proc. IFIP WG 8.4 Work. on Office Information Systems: The Design Process, North-Holland, Amsterdam, 1989, pp.239-253.

# Information Preservation in ARIADNE

**6th DELOS Workshop – Preservation of Digital Information**

Nuno Maria, Pedro Gaspar, António Ferreira, Mário J. Silva
{nmsm | pmag}@ui.icat.fc.ul.pt {asfe | mjs}@di.fc.ul.pt

ICAT/FCUL

Instituto de Ciência Aplicada e Tecnologia
Faculdade de Ciências da Universidade de Lisboa
Lisboa – Portugal

## ABSTRACT

Preserving digital information is a necessary commitment to the future. In ARIADNE, a project of the Digital Publishing Group at ICAT, we are developing an integrated information system for news processing and publishing. In this paper, we present our perspectives on various information preservation issues addressed by this research. These include the semantic preservation of information classification schemes, preservation of the layout of dynamically generated documents, preservation of the linkage to external collections, and the economic sustainability of the news archive.

## INTRODUCTION

Preserving or archiving digital information is, today, an important and necessary commitment to the future. How will we make available today's news to future readers? Giving the importance of this topic, the Digital Publishing Group of ICAT is studying new methods for preserving and processing heterogeneous information in organizations, combining multiple databases under a common framework. In project ARIADNE, jointly developed with *Público*, a national daily newspaper, we are building a new digital publishing structure, where all the information used and produced by journalists is organized in a common database (containing both data and metadata for collections maintained outside the organization). From the information in this digital library, we generate publications in digital format.

*Público* already maintains an archive of all the editions of its paper publications. This is a profitable unit within the company. The archive is used by the newspaper journalists and provides services to external entities. However, as we move into on-line publishing (we are about to release several new publications which will be available exclusively on-line) there is a need to define the processes for archiving and retrieving previously published on-line information. These new on-line publications differ significantly from the previous generation, where we were doing little more than creating on-line replicas of the paper

editions. Our new publications are beginning to behave more and more as interactive user interfaces to databases of multimedia presentations.

In the next section we present the global architecture of our system, and then proceed with a more detailed discussion of the preservation issues in our digital library.

## ARCHITECTURE

The architecture of ARIADNE is based on large multimedia data repository, which holds various collections of documents, newspaper articles, databases of readers and authors, places and events. For some collections, namely external publications, we only keep metadata and links for the articles.

The global architecture and the main information flow are shown in Figure 1. Several sources, including news agencies feeds, articles created for the paper edition of *Público* and external publications provide news items to the ARIADNE repository. Each article received is submitted to a preprocessing stage, where its metadata is extracted. The articles and news feeds are then converted into a common format based on XMLXML Specification Version 1.0, December 1997, , and archived in the collection repository with the software module *Loader*. Editions of electronic publications are built on a second stage, using another module, *Generator*, which selects a group of articles archived in the collections repository and packs them into presentations (or editions). This process finishes by converting the XML sources to HTML, making articles viewable from the current generation of web browsers. With this strategic approach we intend to overcome possible changes in data format standards and sustain our archival missionIssues and Innovations in Preserving Digital Information, in Transforming Libraries, Issue 5, ARL March 1998, <http://www.arl.org/transform/pdi>;.

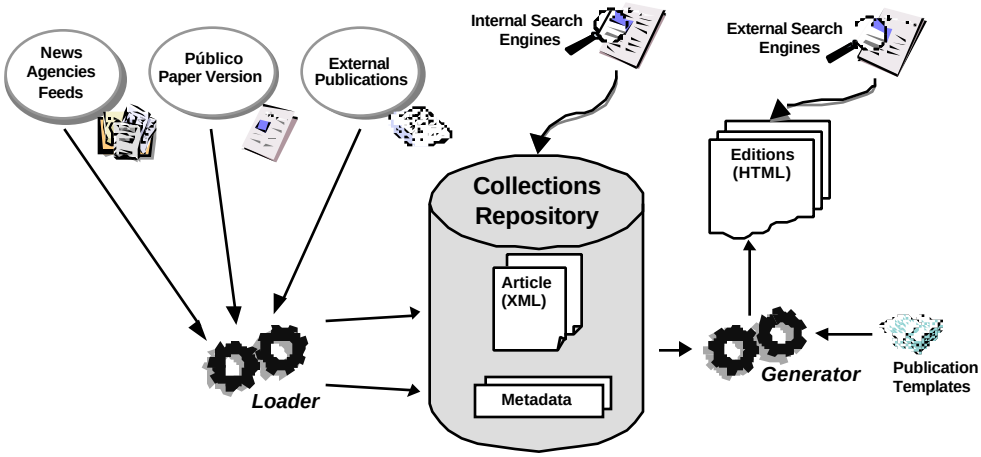



**Figure 1. ARIADNE global architecture, and its main information flow. Each new article is classified, converted to XML and archived. Electronic publications are generated by picking articles from the various collections maintained in the repository.**

Figure 2 shows the UMLUML – Unified Modeling Language, <http://www.omg.org/library/schedule/Technology_Adoptions.htm#tbl_UML_Specification> class diagram of ARIADNE's collections repository, with its main entities. The *article* represents the major information unit in this model. All other main classes are directly associated to it.

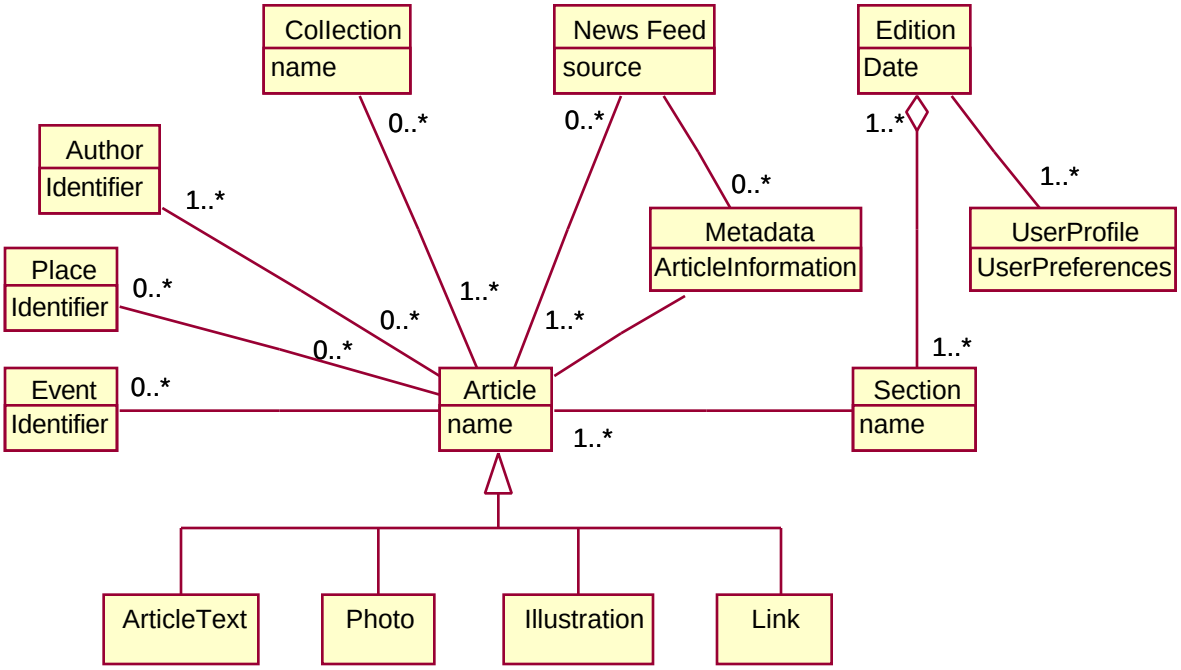



**Figure 2. UML model for the ARIADNE data repository architecture. The article is the major information object in this model. All other entities are directly related with this object.**

ARIADNE has also an internal search facility, combining information retrieval techniques with relational queries and data mining agents. Journalists and registered users will use this facility to retrieve information from the collections' repository. As our publications are available on the World Wide Web, the main Internet search engines also index them.

Personalized publication is another major feature of ARIADNE. We use information-filtering techniques and maintain dynamic user profiles. These profiles are updated as a result of the data mining of access logs and user specified preferences. With this scheme, we intend to track users' preferences as they change over time. However, this raises the problem for preserving personalized editions.

## PRESERVATION ASPECTS

We face several problems associated with the archiving of the publications maintained in ARIADNE. First, we need to provide access to past editions of publications, with the additional complexity of preserving each of the personalized editions. As the personalized editions are dynamically generated from

queries to the repository, we also need to preserve old information classification semantics when retrieving past editions.

Secondly, we need to preserve the layout of these personal editions. The layout format is one of the most characteristic aspects of a publication. The location and format of the articles within a page provide many visual clues, which are later used by readers to recall the same items.

Finally, as with all other digital libraries, we also face the problem of maintaining articles from external publications, handling their copyright restrictions, and recovering the cost of archiving the information.

The remaining of this section discusses our views on these preservation-related topics.

**Preservation of Access Paths**

Information classification is a complex, but necessary task for an easy discovery of resources. In ARIADNE, we use part of the work developed in the Dublin Core Workshop Series, and apply the 15 elements of the Dublin Core (DC) metadata element setDublin Core Metadata, <http://purl.oclc.org/metadata/dublin_core>; to each article in the data repository. In addition, we created complementary schemes for the specific domains of some of our collections. For instance, in our recipe collection, we added five additional descriptors: *recipe type, region of origin, preparation cost, difficulty and time.* This approach makes it easier and much more efficient to search and retrieve information in ARIADNE's data repository, but introduces the problem of preserving access paths to information items as the classification schemes and the retrieval interfaces change.

As concepts evolve over time, it is unrealistic to expect that the key characteristics of each information domain will remain static. In the previous example of our recipe collection, we started with a classification containing just four descriptors. Later on, *Público* decided to print several regional culinary books and we added the new *region of origin* attribute to the recipes loaded from these books. As the recipes already in the collection, were not classified with this descriptor, when a reader uses the new interface, with the new classification topic to select some regional recipes, these unclassified recipes will not be selected.

Another problem with this classification scheme raised when it was necessary to separate soup recipes from "*Açorda*" recipes (*Açorda* is a traditional Portuguese recipe made of soup and bread). Initially, both soup and *Açorda* recipes were classified only with the "soup" attribute, but the new retrieval interface distinguishes between these two kinds of recipes. This re-classification introduces a new problem: we now need to preserve the semantics of search results.

Information semantics based on information classification is a very important aspect in our architecture. We are studying in ARIADNE two different approaches to this problem. First option is to complement the retrieve mechanism, plugging two or more different components in the engine, which then will be responsible for the information retrieval in each single semantic classification domain. In this approach, the archive remains unchanged and there is no need for digital records reclassification. Another option

would maintain a single and simple retrieve mechanism, but reclassify each digital record in the main repository according to the new semantic classification scheme.

Each solution has advantages and drawbacks. With the first, we have a scalability problem, as the introduction of new semantics implies adding new retrieve mechanism plugins, which delay retrieval operations. However, this solution has high maintainability because its centralized retrieval mechanism can be easily upgraded. With the second approach, we have a fast retrieval mechanism, but maintaining the collections will be hard, as we predict a huge, distributed data repository and the application of a conversion function for reclassification of a large number of digital records may be impracticable.

In the recipe collection, we fixed the "*Açorda*-soup" problem using the first approach, by inserting a plugin in the search engine that controls queries in this collection. The search engine understands each query and collects all related recipes, presenting them in the order implied by users' queries. Nevertheless a similar approach will not be efficient with the regional recipes' problem, because the criteria for identifying its region are not clear. In this case, a reclassification of the recipe collection would be a better choice.

**Layout Preservation**

How many times did we search for that particular second item on a report list, generated by a web search engine, even though we did not remember what was really written there? This shows the importance of preserving the order of presentation of information items within a digital document.

In a traditional repository of digital publications we might archive each edition with its user interface and it would remain forever that way. However, in ARIADNE, the existence of personalized editions raises the problem of also preserving the profiles that generate these editions. We use a method to archive each edition in a way that allows storage space consumption to increase only incrementally as we add new readers. In our architecture, a user profile maps a reader into one of a set of presentation styles (or templates). Publication editors define these styles to match typical profiles that they have identified among their readership. This approach still requires the preservation of the presentation styles used by each reader over time, so we can re-create the user interface and the layout of pages dynamically generated when a reader visits the same edition of a publication.

**Technology Preservation**

Another key topic in information preservation is the technology. We strive to make ARIADNE resist to continuous changes in representation formats, by adopting the latest standard and converting information into this format. When necessary, because some browsers may not yet support the new standard, we generate the information in the older formats. It may sound incoherent but, with this design, we can let readers access past editions, even if meanwhile they have migrated into new browsers, and we find it easier to convert data back into the original format when required. In ARIADNE we are already storing articles in XML, which is richer then HTML and will be directly viewable with the next generation of

browsers. However, as almost our readers cannot read this format, we are currently generating editions in HTML.

We will keep on generating HTML editions until only a small fraction of our readers remains without the capability to process XML documents. It is a growing common ground the increasing conservatism of web users. Jakob Nielsen showed recently the increasing conservatism of web users, which frequently don't have the latest client software available Nielsen, Jakob, The Increasing Conservatism of Web Users, March 1998, <http://www.useit.com/alertbox/980322.html>;. So it is important to keep ARIADNE's publications easily accessible not only by the fifth generation Internet browsers but also by the first generation browsers. There is no point in converting information to the latest format if most of the readers cannot access it.

**Linkage Preservation**

It is impossible to us to track all of the journalistic information available on the web. We can not rely on available Internet search engines, as they do not index other newspapers of interest to us with the required frequency. As a result, we need to create our own index, and establish our own approach for preserving this linkage. For several of the external sites, we keep linkage and metadata information. This strategy overcomes several legal and scale problems, but it also raises other problems related to the preservation of the referenced contents. What will happen if the publisher changes information location or even ceases business Issues and Innovations in Preserving Digital Information, in Transforming Libraries, Issue 5, ARL March 1998, <http://www.arl.org/transform/pdi>;? How about the intellectual preservation of remote resources, there is no way to make sure the article maintains the integrity and authenticity of the information as originally recorded .

Facing these disadvantages we decided in some cases to collect entire external publications. As there is widespread uncertainty about legal requirements for managing intellectual property in digital environmentIssues and Innovations in Preserving Digital Information, in Transforming Libraries, Issue 5, ARL March 1998, <http://www.arl.org/transform/pdi>;, we restrict availability of this information to the journalists of Público. With this approach, we intend to increase our digital library functionality, avoiding corrupted information contents and outdated linkage.

**Preservation Costs**

The electronic archive being developed brings many costs and must sustain itself economically. Gillian Laughton presents an interesting comparative study of the archiving cost of an ASCII text based electronic journal with that of a new generation journal in HTML Laughton, Gillian, Archiving of Electronic Journals, <http://solaris.cis.csiro.au/im/ejournal/archive.htm>;. In an example of his study, once indexing and the increased staffing required to maintain the more complex formats are included, the overall costs jump from $435 to $1,000 per title per year. In our publishing system, access to recent news is free, and advertising is the only source of income. However, the use of all other services will be charged. These services include searches on our indexes and collections, and notifications sent by user

subscription basis. We are starting the development of a micro-payments system, so that we can charge individual news articles. However, we do not have an idea of how much revenue these services will produce. This may be only be obtained through observation of the readers reaction to our pricing policies.

## CONCLUSIONS

In ARIADNE we are concerned with some specific aspects of information preservation.

We attempted to describe in this paper our current concerns and our approach to building an archive of digital publications. This is becoming an increasingly harder problem to solve, as we move into personalized editions, which invoke dynamic queries and incorporate richer information types. However, in our view, the exact information available and its presentation organization at the time of publishing are essential aspects to preserve in the archive of a reference daily newspaper, such as *Público*.

A common concern of librarians is that computer science success is “in part because of its luxurious ignorance of the past”, and this generation choose to delay the problem of archiving, “mainly because it is simply not as interesting as developing the wonders of future” Issues and Innovations in Preserving Digital Information, in Transforming Libraries, Issue 5, ARL March 1998, <http://www.arl.org/transform/pdi>;. In ARIADNE we are feeling the same pressure from our partner journalists.

## REFERENCES


[1] Graham, Peter S., Preserving the Digital Library, Long Term Preservation of Electronic Materials Workshop, November 1995, <**Erreur! Signet non défini.**>

[2] Hall, Barbara, Archiving Electronic Journals, presented in the American Library Association Annual Conference, June 1997, <**Erreur! Signet non défini.**>;

[3] Issues and Innovations in Preserving Digital Information, in Transforming Libraries, Issue 5, ARL March 1998, <**Erreur! Signet non défini.**>;

[4] Laughton, Gillian, Archiving of Electronic Journals, <**Erreur! Signet non défini.**>;

[5] Nielsen, Jakob, The Increasing Conservatism of Web Users, March 1998, <**Erreur! Signet non défini.**>;

[6] Dublin Core Metadata, <**Erreur! Signet non défini.**>;

[7] HTML - HyperText Markup Language, <**Erreur! Signet non défini.**>;

[8] UML – Unified Modeling Language, <**Erreur! Signet non défini.**>

[9] XML Specification Version 1.0, December 1997, <**Erreur! Signet non défini.**>;

# Preserving the U.S. Government's White House Electronic Mail: Archival Challenges and Policy Implications

David A. Wallace (daval@umich.edu)
University of Michigan, School of Information

Sixth DELOS Workshop: Preserving Digital Information
Lisbon, Portugal – June 19, 1998

## Introduction

This paper examines the archival and policy implications resulting from a decade of litigation over the creation, use, management, and preservation of electronic mail technology in the Executive Office of the President of the United States government from the mid-1980s onwards. While the context under examination is particularistic in many respects -- such as the applicable recordkeeping statutes and its political and organizational contexts -- is does speak substantively and powerfully to broader issues that any records management and/or archives program will need to confront as it struggles with digital information resources. Among others, this case surfaces concerns over the relationship between computing and continuing governmental accountability via recordkeeping, distinguishing official from unofficial records, distinguishing between different types of official records, evaluating the distinctive qualities between electronic records and their printed counterparts, the need to develop and implement electronic recordkeeping systems, assigning appropriate disposition schedules that ensure that records of continuing value are preserved while providing for the appropriate destruction of temporary records, and the detrimental impact to archival programs that attempt or are required to perform «salvage archiving» of computer generated data.

After a brief discussion of the introduction and use of electronic mail technology in the U.S. National Security Council (NSC), discussion turns to a description and analysis of the court case arising from a dispute over the propriety of the policies overseeing that use and of the archival preservation challenges it presented. This paper then closes with a series policy and technology lessons applicable to other contexts.

## Electronic Mail Use and Management in the U.S. National Security Council: 1985-1989

In 1982, the U.S. National Security Council (NSC) installed an electronic mail (email) system on a pilot basis. In April 1985, email was made more widely available throughout the NSC using IBM's proprietary "Professional Office System" (PROFS). Later, other email systems would be introduced, including the VAX-based All-in-One package. The PROFS system allowed users to exchange email, transfer text documents, and share calendar information. PROFS email functionalities provided users with the ability to log on to the

system and compose, transfer, display, receive, store, file, forward, print, and delete electronic messages. Backup tapes of all messages stored on the system were performed on a rotating nightly and weekly basis.[1]

The PROFS email produced by the White House's NSC gained wide public notoriety in late 1986 and throughout 1987 with the exposure and eventual investigation into the «Iran-Contra Affair,» an illegal initiative that sold arms to Iran to obtain the release of U.S. hostages and then used the profits from these sales to fund the U.S.-created Contra army in Nicaragua in its efforts to overthrow the Sandinista government. The PROFS system provided the primary communications conduit between the two key participants in this diversion scheme – NSC staffer Oliver L. North and his boss, National Security Advisor John M. Poindexter. In April 1985, Poindexter made it possible for North to send him email messages directly, bypassing the normal flow and filtering of email through the NSC's Executive Secretariat. It was through this unique email communications channel that North and Poindexter were able to secretly conduct their work related to Iran and Nicaragua.[2]

When the diversion of funds from the Iranian arms sales to the Contras became public in November 1986, both North and Poindexter began destroying documents, including email messages, associated with their role in the Affair. Right before they were to become the subjects of intense investigatory scrutiny, North deleted 736 email messages from his user storage area and Poindexter deleted an astounding 5,012 messages. Such deletions are all the more remarkable in light of the fact that each message had to be individually deleted. While the messages may have been deleted from the live system, they still existed on backup tapes that had been pulled aside by the White House Communications Agency (WHCA) which oversaw management of the PROFS system. These backup tapes and the existing live system provided three chronologically separate snapshots of the PROFS system immediately preceding and following the exposure of the scandal. By comparing the user storage areas for North and Poindexter across these three snapshots investigators were able to identify and examine those messages that North and Poindexter had deleted once the scandal became public and investigation was imminent.[3] These «recovered» PROFS messages became crucial evidence in the subsequent Congressional and other investigations into the scandal as well as the criminal trials of both North and Poindexter.

In the wake of initial investigation into the Iran-Contra Affair, the NSC adopted a formal policy for its email. It directed staff to store as little information as possible on the email system and to retain only those messages that would be needed for future reference. In the event that a staffer was «tasked for action» via an email message, they were directed to print the message out onto paper and incorporate it into the package they forwarded to their principals. The attitude towards email at this point was that it was merely designed to serve as a surrogate/substitute for «information that would be otherwise handled by phone.»[4] NSC staff were reminded that email usage was not intended to create official government records, nor was the system itself to be thought of as a formal recordkeeping system. In the odd event that an official record was created via email – if it had «enduring value, or if it documented agency functions and transactions» -- it was to be printed out onto paper and filed or its content was to be «memorialized» in a written memorandum or letter.[5] Staff had to be later admonished to keep the length of their email messages to a minimum and to create a typed

formal memorandum instead of composing long and complex email messages (unless «time [was] truly of the essence.»)[6] During the preparation for the transition between the Reagan and Bush administrations in January 1989, White House employees were instructed to «take care» and review their computer data, including their email user storage areas to «ensure» that they had made «hard copy of all 'record' material….»[7]

After the initial Iran-Contra investigations in 1986 and 1987, the PROFS system receded back into obscurity. Given what it felt was a clear and sound policy for managing email messages, the government had expected to erase all Reagan-era electronic versions of email messages stored on the PROFS system to free up disk space for the incoming Bush administration. The accidental discovery of this proposed erasure threw open the NSC's management of its email to public scrutiny and led to a decade long series of lawsuits that continue up to the present.

## A Decade in Court: The Impact of Technology on Recordkeeping Law and Practices, 1989-1998

In the waning days of the Reagan administration, the National Security Archive (NSA), a nonprofit research library of declassified U.S. government records, discovered informally by chance from an employee of the U.S. National Archives and Records Administration (NARA) that all non Iran-Contra-related email backup tapes would be erased and recycled and that the live email system would be purged to make room for the work of the incoming Bush administration. All of the Iran-Contra backup tapes uncovered during the initial investigation into the scandal were slated to be saved as evidence for other ongoing investigations. Upon receiving official confirmation that this in fact was the government's plan of action the NSA sought legal relief. NARA's position at the time was that anything of record significance would have been printed out and filed into a formal recordkeeping system, hence anything that remained electronically would have been either a redundant «convenience copy» or non-record material. In addition, since it was standard NARA practice to not accession any electronic records that had not been converted from a proprietary format into a hardware and software independent format and that all of the electronic versions of the email messages it had approved for erasure existed in proprietary email software packages, NARA was of the opinion that the erasure was clearly in line with both policy and law.[8]

In their initial legal action in January 1989, the NSA sought a Temporary Restraining Order (TRO) to prevent the government from moving forward with the proposed erasure that was to occur in less than 48 hours. In the NSA's opinion the government's argument that all official record material had been printed out and filed was not congruent with the recovery of North and Poindexter's electronic email messages from the backup tapes set aside by the White House Communications Agency (WHCA). There was no indication that any of the messages deleted by North and Poindexter had been printed out and filed into a recordkeeping system. Naming President Ronald Reagan, President-elect George Bush, the National Security Council, and the Archivist of the United States as co-defendants in their lawsuit, the NSA contended to a U.S. District Court judge on the eve of the presidential transition that the erasure would violate the Presidential Records Act (PRA),[9] the Disposal of Records Act (a

component of the broader Federal Records Act (FRA)),[10] and the Administrative Procedures Act (APA).[11]

The PRA stipulates, in part, that the President «shall take such steps as may be necessary to ensure that the activities, deliberations, decisions, and policies that reflect the performance of his constitutional, statutory, or other official or ceremonial duties are adequately documented and maintained.» The PRA also directs that at the end of a President's term in office the Archivist of the United States «shall assume responsibility for the custody, control, and preservation of, and access to the presidential records of that President.» In order to dispose of any presidential records during his term in office, the PRA requires that the President obtain the views and approval of the Archivist of the United States. And upon the expiration of a presidential term of office when the Archivist has taken custody of an administration's records, the Archivist can only appraise records for disposal once he/she has publicly announced the proposed destruction sixty days before it is to take place.[12] The FRA defines recordkeeping responsibilities for both federal agencies and NARA. The FRA requires the head of a federal agency «to make and preserve records containing adequate and proper documentation of the organization, functions, policies, decisions, procedures, and essential transactions of the agency and designed to furnish the information necessary to protect the legal and financial rights of the Government and of persons directly affected by the agencies activities.» In order to accomplish this the FRA directs federal agencies to «establish and maintain» a records management program which provides for «effective controls over the creation…and maintenance and use of records in the conduct of current business.»[13] The FRA directs the Archivist to provide «guidance and assistance to Federal agencies with respect to ensuring adequate and proper documentation of the policies and transactions of the Federal Government and ensuring proper records disposition.» It also requires the Archivist's approval for any agency records destruction.[14] The APA defines the scope of administrative rulemaking and decisionmaking in the executive branch of the government and also defines the parameters of judicial review of administrative decisions. It also makes it a violation of law for elected or appointed official to act in an arbitrary, negligent, or capricious manner.[15]

In their January 1989 complaint, the NSA requested that the court prevent the government from erasing the backup tapes and wiping existing messages from the White House's live email system.[16] The government countered that the NSA did not have a legal right to make the challenge they were proposing, that their action would «gravely impair» the Presidential transition, and that all that was occurring in this instance was the «removal of extraneous and unnecessary communications.»[17] After granting the NSA and the government an hour to make their best arguments, U.S. District Judge Barrington D. Parker granted the NSA's request for a TRO and assigned the case to U.S. District Judge Charles Richey.[18] Little did anyone involved at this point realize that this simple act would open up a decade long legal battle that has persisted to the present.

Remarkably, it took four full years of litigation before the court ruled on the adequacy of the defendants' recordkeeping guidelines and the Archivist's performance of his statutory obligations. In the interim the government argued that both their oral and written recordkeeping guidances amply demonstrated that their recordkeeping practices were in

accord with the FRA. They pointed out that since 1987 the NSC had provided oral guidance to employees on their recordkeeping responsibilities both when they started working for the NSC and again when they departed. Employees were explicitly instructed that when an email message was a «record» it was to be printed out and logged into the formal paper recordkeeping system. They also pointed out that since February 1990, departing NSC employees were required to read and sign a certification that they had «met their recordkeeping obligations and [had] handled their electronic mail in accordance with the prescribed requirements.» In May 1992, the NSC modified the PROFS software so that when a user wanted to send a message they first had to assign a record status – presidential record, federal record, or non-record – before the system would route it. If a message was assigned the status of either «presidential record» or «federal record» a copy of it was automatically transmitted to the NSC's records management office for printing out and filing in to a recordkeeping system.[19] In their response to defendants' claims of proper behavior, the plaintiffs argued that the FRA required that all records, regardless of medium, had to be preserved unless NARA had fist authorized their disposal. They claimed that in this instance the defendants had «arbitrarily» deemed email as non-records without first making any effort to evaluate their content in order to justify such a determination, and that employees were not provided adequate guidance on how to identify a federal record generated by email system and how to distinguish record from non-record material. The plaintiffs also rejected the government's claims that the electronic versions of email messages were merely extra copies and not official government records. By declaring them to be extra copies and not records under the FRA, the plaintiffs contended that the defendants had «erroneously instructed» staff of their legally binding recordkeeping responsibilities. The plaintiffs asserted that electronic records were not extra copies because their «form and content are unique» and printouts did not necessarily capture all of the information associated with a particular document. Items such as the identity of the sender and the recipient, acknowledgement receipts which provide the sender with a confirmation that their message was received, as well as the date and time of receipt and system usage statistics such a user logon/logoff and connect times were some of the types of electronically stored metadata that appeared nowhere on printouts. The plaintiffs further contended that the existence of a paper printout did not invalidate the record status of the electronic record version and that instead of being an extra copy the electronic version continued to be a record in its own right.[20] In a counter-reply, the defendants criticized the plaintiffs for asserting that the government was «somehow affirmatively obligated under existing law to do more than simply preserve 'records' contained on the PROFS system in hard copy paper format.» The government argued that the defendant agencies had consistently employed a «paper system as its primary means of maintaining agency files.» As such, the defendants had been totally within their legal discretion to not designate the PROFS system as a recordkeeping system for filing and managing records. They claimed that they had always treated PROFS as a communications system which sometimes was used to transmit records, but which for the most part communicated non-record material. Regarding the plaintiffs' contention that the electronic versions of PROFS materials contained information not available on the printouts, the defendants countered that when a PROFS note, calendar, or document is printed out the resulting paper copy contains, with the exception of function keys, all the information that had appeared on the user's computer screen. The defendants stated that they were «unaware of any authority…for the proposition that defendants [we]re obligated to do more….[T]here

is certainly no requirement that individuals spell out abbreviations in their paper letters and memoranda, or track down the times of receipt of the documents they create[d] or note when acknowledgements in the form of return notes were received, all prior to 'archiving' such letters or memoranda in traditional agency files.»[21]

Taking into account all of the above arguments, U.S. District Judge Richey issued his ruling on the matter in January 1993. In brief, Judge Richey determined that the defendants had violated the FRA and that their recordkeeping practices were «arbitrary and capricious» under the Administrative Procedures Act (APA) because they permitted the improper destruction of federal records. He also ruled that the Archivist had failed to fulfill his statutory duties as mandated by the FRA and directed the Archivist to take immediate action to preserve the «electronic federal records» that had been the subject of the case and develop new guidelines for managing email. Richey specifically faulted the Archivist for not preventing the destruction of federal records. On the issue of the record status of the electronic versions of the email messages, Richey ruled that despite the fact that not all information stored on the defendants email systems were records, he could not «read the FRA to exclude computer systems such as those at issue here.» To buttress this contention he noted that the FRA had been designed to include materials regardless of physical format. On this issue of the adequacy of paper printouts as surrogate records to the electronic versions, Richey determined that paper printouts did not reproduce information that existed in electronic versions. He specifically referred here to information about who received a message and when it was received as well as distribution lists, lists of individual senders and recipients, times of acknowledgement, and logon/logoff times. Richey rejected the defendants argument that such items do not rise to the level of a record by noting that «[d]efendants' argument misses the point because this information does not stand alone. This information must be saved because, in combination with the substantive information contained in the electronic material, it will convey information about who knew what information and when they knew it.» Since the electronic versions could be federal records in their own right, he ruled that they must be saved, regardless of whether or not a paper copy was made. This determination made obsolete the defendants' continuing contention that the electronic version was merely an extra copy of the paper printout. Richey also ruled here that the defendants recordkeeping procedures and recordkeeping guidelines violated the APA because they provided an inadequate records management program or supervision of staff decisions on the record and non-record status of their email messages and that they also allowed the improper destruction of federal records.[22]

Upon receiving the above decision, the defendants immediately appealed and sought relief at the next higher level of the judiciary. [The plaintiffs also appealed a portion of Richey's decision, however, their appeal dealt with issues which are not of direct concern to this paper.] In August 1993, the U.S. Appeals Court ruled. They affirmed Judge Richey's January 1993 decision that the defendants electronic records management guidelines were in violation of the FRA, that paper printouts of electronic versions of records are not acceptable substitutes for the electronic versions as the strip off relevant contextual information, and that the existence of a paper printout did not invalidate the record status of the electronic version. In specific reference to the defendants recordkeeping guidelines, the Appeals Court found that the instruction to print hard copy paper versions of electronic records was «flawed because

the hard copy printouts that the agencies preserve may omit fundamental pieces of information which are an integral part of the original electronic records, such as the identity of the sender and/or recipient and the time of receipt.» In exploring this issue in more detail, the Appeals Court reasoned that by 1993, nearly 1,000 federal employees had access to Executive Office of the President (EOP) and NSC email systems and apparently used them to «relay lengthy substantive – even classified 'notes' that, in content, are often indistinguishable from letters and memoranda.» The paper printouts made from an email message would not necessarily contain all of the information associated with the same document that resided on a computer system. «Directories [for deciphering oftentimes cryptic user ID's and nicknames], distribution lists [which provide simple aliases that might include many users], acknowledgement of receipts and similar materials do not appear on the computer screen – and thus are not reproduced when users print out the information that appears on the screen.» Hence, a subsequent reader of the hard copy version may have trouble distinguishing «basic facts» about the document such as its sender, recipient, and time of transmission. And if the electronic version was erased then such contextualizing information would be forever unavailable. In addition, the fact that the electronic version was reduced to a paper copy did «not affect the record status» of the electronic version and render it an extra non-record copy unless the printout «include[d] all significant material contained in the electronic records.» The record compiled as a result of the case demonstrated to the Appeals Court that, as currently constructed, a printout and electronic version of a message could not appropriately be called copies of one another and, consequently, the electronic version continued to retain its federal record status even after it had been printed out. As such, «all of the FRA obligations concerning the management and preservation of records» still applied to the electronic version. To the Appeals Court mind, since the defendants' agencies employees had never been instructed up to the time of the Judge Richey's January 1993 order to include «integral parts of the electronic record in any paper printout, there is no way [they] could conclude that the original records are mere 'extra copies' of the paper printouts.» The Appeals Court therefore found that the District Court's January 1993 ruling was «fully justified in concluding that [the government's] recordkeeping guidance was not in conformity with the [FRA].»[23] With this ruling the parties entered in settlement negotiations regarding the development of new recordkeeping guidelines. While at the time this may have appeared to have led all concerned to see the light at the end of the tunnel, new controversies would emerge that would lead to new litigation.

As a consequence of the above judicial determinations, the National Archives was required to develop a new government-wide policy for email management. In March 1994, they published a «notice of proposed rulemaking» and invited comments from any and all interested parties. The guidance was drafted by NARA with the goal of instructing federal agencies across the government on the «proper means of identifying, maintaining, and disposing of Federal records created or received on an email system.» When the final draft was to be issued it would be designed to provide agencies with the means to «develop specific recordkeeping policies, procedures, and requirements to fulfill their obligations» under the FRA and NARA regulations. The draft guideline pointed out that email messages were not to be considered non-records materials «merely because the information they contain may also be available elsewhere on paper or in electronic files.» It also stated that email messages could not be deleted without prior disposal authority from NARA. This applied equally to all

versions of an email message, including the original electronic version. The draft guideline encouraged agencies to consider maintaining their electronic mail generated federal records in electronic form. The advantages noted for electronic maintenance included ease of storage, searching and manipulability, and the simultaneous availability of the records to many different users. The guidelines, however, did not mandate the electronic maintenance of email generated federal records. They instead placed the focus on the need to maintain email records in a proper recordkeeping system. While the proposed guideline advocated electronic recordkeeping systems, they offered agencies the option to print out email records and file them in a paper recordkeeping system. If an agency email system was not designated as a formal recordkeeping system, the draft guideline instructed that the email in question «must be copied or moved to an appropriate recordkeeping system for maintenance and disposition.» The guideline approved of such action as long as the appropriate «transmission and receipt information,» such as sender, recipient(s), message date, and read receipt were attached to the printout. Only when the electronic mail record was stored in a proper recordkeeping system, whether electronically or in paper, would the original electronic version on the 'live' email system be «appropriate for deletion.»[24]

This draft guideline elicited an enormous response. Over ninety-two separate comments totaling over 1,500 pages of written comments were received by NARA from federal agencies and private organization and individuals. This dwarfed previous replies to other notices of proposed rulemaking. Over 80% of the responses were from federal agencies, the «vast majority» of which were critical of its stipulations. Comments on the draft regulations revolved around several themes, including: that they would be too expensive and burdensome on agencies (agencies were under the impression that the guidance required electronic preservation, which they argued was not possible given the poor records management functionalities of commercial off-the-shelf email packages – as a consequence the final rule was revised to «provide realistic requirements that agencies can meet immediately); that they rendered too many email messages as records; that it would have a «chilling effect» on agency email usage; that they inflated the significance of email; that NARA could not impose upon agencies the format in which they chose to preserve their records; and that the requirement to preserve transmission data was too complex, especially for distribution lists. The «final rule» was issued in August 1995. By that time the original lawsuit had effectively run its course and with the issuance of the final guideline Judge Richey dismissed the case from his court. A major change that resulted from the federal agency commentary was that references to electronic recordkeeping became muted in the final rule. NARA argued that discussions of electronic recordkeeping was something to strive for in the future and something which was better suited to a separate guidance. The final rule made agencies responsible for providing adequate training to staff, instructing them on distinguishing between records types and on how to transfer electronic mail messages into agency recordkeeping systems, be they electronic or paper. It also contained the following stipulations specific to the management of electronic mail. Transmission data (identification of sender, recipient(s), date sent) had to be preserved if the message's context was to be decipherable in the future. Agencies needed to determine what, if any, other transmission data should be linked to messages. Lists of nicknames in directories and/or distribution lists needed to be retained so that the identity of individuals on the system could be known. And for systems that provided them, read receipts needed to be preserved as well. Agencies were

specifically instructed to not store copies of federal record email on an email system unless the system: enabled grouping of related records into relevant categories; permitted «easy and timely retrieval» of both individual items and groupings; was accessible to individuals who required access to them; was maintained in a usable format as specified by a NARA-approved records retention schedule; preserved transmission and receipt data; and, provided for the transfer of permanent records to NARA. Agencies whose electronic management of their email did not meet these standards were required to transfer electronic federal record email messages to a proper recordkeeping system. Transfers to a paper-based recordkeeping system required that proper transmission data be attached to individual messages. The final rule also forbade the destruction of electronic versions of email messages, whether they were records or not, without «prior disposition authority from NARA.» Once an electronic mail message was transferred to a recordkeeping system, «identical versions» such as the remaining electronic copy could be disposed of under General Records Schedule 20 (GRS 20), which dealt with the disposition of electronic records.[25]

The lesson taken by the government from the January and August 1993 court rulings was not how to enable electronic recordkeeping. Rather, it was how to make better paper printouts of email messages that included suitable transmission data. The plaintiffs lead, Michael Tankersley, later complained that by allowing agencies to rely on GRS 20 to dispose of their electronic email records, NARA was enabling the wholesale destruction of these messages regardless of their content or their qualitative differences to paper printouts.[26] General Records Schedules provide for «disposal authorization for temporary records common to several or all agencies of the federal government. They include records relating to civilian personnel, fiscal accounting, procurement, communications, printing, and other common functions, and certain nontextual records.» Agencies are permitted to dispose of records covered under a General Records Schedule without additional approval by NARA and without public notice. Such records are believed to constitute one-third of all records created by federal agencies. The remaining two-thirds of federal agency records – substantive program records – need to be covered by General Records Schedules specifically created for such program records.[27] Tankersley's objection to the government's reliance on a General Records Schedule for disposal of electronic records was that it classified all electronic records as a uniform type of record based on their format, whereas General Records Schedules were supposed to deal with classes of information based on their function. This objection represented a fundamental chasm of opinion between the plaintiffs and the defendants – a chasm that was to lead a new lawsuit in December 1996.[28]

This most recent episode of PROFS-related litigation resulted from NARA's issuance of a new General Records Schedule 20 on the disposition of electronic records in conjunction with their final email regulations in August 1995. A draft of GRS 20 issued by NARA in October 1994 yielded 37 comments, 14 of which were submitted by federal agencies who were generally supportive of it. The 23 non-government submissions were largely critical, claiming that it would provide for the inappropriate deletion of electronic versions of records that had been converted to either paper of microform. The final August 1995 GRS 20 provided, in part, for the deletion of electronic versions of records created on word processing and electronic mail systems once a recordkeeping copy had been made and filed into either an electronic or paper-based recordkeeping system. This new GRS 20 was applied for the first

time to office automation systems. Previously, such records were covered by General Records Schedule 23 – Records Common to Most Offices within Agencies.[29]

On December 23, 1996, roughly the same group of plaintiffs that had entered into the initial PROFS litigation in 1989 sued the government once again, this time over the new GRS 20. In this complaint the plaintiffs alleged that the new GRS 20 purported to «authorize destruction of electronic mail and word processing files at all federal agencies if a hard copy of the record had been created on paper or microfilm.» The suit reported that on December 17, 1996 the Archivist of the United States had endorsed an Executive Office of the President decision to dispose of electronic records under this new GRS 20, including electronic records from the Office of the U.S. Trade Representative that were supposedly preserved pursuant to the prior PROFS litigation.[30]

On October 22, 1997, U.S. District Judge Paul Friedman ruled against the government and declared GRS 20 to be «null and void» and ordered the defendants to «not destroy electronic records created, received or stored on electronic mail or word processing systems pursuant to General Records Schedule 20.» Judge Friedman determined that in issuing GRS 20 the Archivist had exceeded his authority under the Records Disposal Act[31] section of the FRA in three ways. First, he inappropriately authorized the destruction of federal agency «program» records under GRS 20 while General Records Schedules were «unequivocally limited…to administrative records.» Second, he found that the Archivist «abdicated to the various departments and agencies of the federal government his statutory responsibility under the Records Disposal Act to insure that records with administrative, legal, research or other value are preserved by federal agencies.» And third, he determined that the Archivist was remiss in identifying, as required by the Act, a specified retention period for the electronic records scheduled under GRS20. Pointing out that some word processing systems provide for a document annotation summary that would provide information on the document's author, purpose, date drafted and revised, etc., Friedman, like the District and Appeals Courts before him in the initial PROFS case, underscored the unique value of electronic versions of documents that are not converted to paper printouts. To Friedman's mind, such electronic records «do not become valueless duplicates or lose their character as 'program records' once they have been printed on paper; rather, they retain features unique to their medium.» Friedman went on to call the Archivist's actions in this case:

> irrational…and one that is necessarily premised on the illogical notion that a paper copy adequately preserves the value of an electronic record. While, in some cases, paper copies may in fact adequately preserve the administrative, legal, research or historical value of an electronic record, there is no rational basis for the Archivist's conclusion that a paper copy invariably adequately preserves such value in all cases and that electronic records never retain any administrative, legal, research or other value once such records have been copied to paper….By categorically determining that electronic records possess no administrative, legal, research or historical value beyond paper print-outs of the same document or record, the Archivist has absolved both himself and the federal agencies he is supposed to oversee of their statutory duties to evaluate specific electronic records as to their value. The Archivist has also given agencies carte blanche to destroy electronic versions without the Archivist's

> approval when the agency believes they are no longer needed by the agency. Because GRS 20 leaves the destruction of electronic versions of records unchecked by the Archivist, it fails to meet the requirements of Section 3303a(d) [of the Disposal of Records Act].[32]

In December 1997, the government field an appeal challenging Friedman's order. As of this writing that court has yet to render an opinion. In April 1998, nearly six months after issuing his rather scathing opinion and order, Judge Friedman answered a motion for action by the plaintiffs and found that the defendants had «flagrantly violated» the above order. At issue were post-October 1997 published issuances by the Archivist in the *Federal Register* that agencies could continue to rely on GRS 20 to dispose of electronic records despite an order to the contrary that «could have not been more clear.» In an order striking the Archivist's issuances down, Judge Friedman ordered the Archivist to issue a new statement to federal agencies that GRS 20 has been rendered null and void.[33]

As a means to placate the Judge and develop a means to resolve the issues raised by the case, the Archivist of the United States convened an «Electronic Records Work Group» in November 1997 and tasked it with assisting the National Archives in developing a strategy to respond to the Judge Friedman's October 1997. It has been specifically charged to: review the current version of GRS 20; identify appropriate areas for revision; explore alternatives for authorizing disposition of electronic records; identify methods and techniques that are available with current technology to manage and provide access to electronic records; and, recommend practical solutions for the scheduling and disposition of electronic records. This working group is composed of NARA staff, federal agency records officers, and outside experts. The most recent options paper presented to the Archivist in May 1998 offered three proposals for bringing government practice in line with Judge Friedman's order: schedule all program records in all formats and eliminate electronic mail and word processing program records from GRS 20; revise the entire GRS to cover all formats of administrative records; and, revise GRS 20 to cover only those systems administration/management and operations records.[34] The timetable for completing the work of the working group calls for issuing the proposed options to federal agencies for comment in the first week of June 1998 and then publishing them in the *Federal Register* for public comment during the week of July 20, 1998. The final version will be publicly issued by the Archivist to meet the court's deadline of September 20, 1998. Given what it calls a «tight timeline,» the electronic records work group has decided to forego any exploration of electronic recordkeeping and has instead chosen to concentrate its efforts on developing a scheduling approach that is compliant with federal recordkeeping law.[35]

What started out as a seemingly simple challenge to the proposed erasure of the Reagan administration's electronic mail messages has developed over the following decade of litigation as a detailed and thorough examination of the records and archival management of computer generated information throughout the federal government via office automation systems across three presidential administrations and their relation to the very specific requirements of federal recordkeeping statutes. The archival preservation of the electronic records preserved as a result of this litigation has proven to be enormously complex and has provided detailed information on the actual physical and organizational challenges involved.

Preserving Electronic Mail Records: Policy and Technology Issues Raised by the PROFS Litigation

As noted above, one central aspect of Judge Richey's January 1993 ruling against the government was his instruction that NARA take immediate action to preserve the «electronic federal records» that had been the subject of the case. As part of this order NARA worked throughout the transition of power between the Bush and Clinton administrations transferring the Reagan and Bush materials from the Executive Office of the President to the National Archives. On January 28, 1993, the defendants entered a post hearing submission to Judge Richey in response to his question regarding what actions the Archivist had taken to comply with the ruling he issued earlier that month. This submission noted that the Archivist had taken physical custody of nearly 5,700 backup tapes (in cartridge, reel and helical scan formats) and over 150 personal computer hard drives. This massive volume resulted from the Temporary Restraining Orders (TRO) issued by the court which prevented the destruction of any messages from the defendants email system. Since the work of the government had to go forward upon the granting of the initial TRO in January 1989, the government had to save every backup it made in the interim until the case was resolved one way or the other. These contents of these items were to be eventually evaluated to distinguish the federal records from the presidential records from the non-records that resided on them.[36]

While the plaintiffs had learned on January 28, 1993 that the materials had been transferred to NARA, it would not be until the following month that they discovered the circumstances under which the transfer occurred. At that point in time they obtained access to a memorandum written by five NARA employees who participated in the transfer of the materials. Termed the «Armstrong Materials Task Force,» these employees reported on the difficulties they faced throughout the transfer. Given the time crunch they were required to operate under and the fact that they were not fully informed as to what materials were covered, their physical location, and their exact volume, the Task Force reported that «it would be impossible to establish intellectual control over the materials» and that they would instead have to rely on the inventories compiled by the White House and verify them against the labels on the materials themselves. Unfortunately, as they collected materials from the National Security Council, the Executive Office of the President's Office of Administration, and the White House Communications Agency, they noted that they «did not receive an adequate description of any system that would allow the Archives to operate the system or review the data contained in the system.»[37] Judge Richey was not patient with NARA's explanations of the problems and challenges posed by the backup tapes and hard drives. Despite the government's assertions that their actions were in accordance with the Court's January 193 ruling, on May 21, 1993 Judge Richey found the defendants in contempt, determining that the government's plans to preserve, copy, and repair the materials in need of immediate action was inadequate. To vacate this contempt ruling Richey ordered the defendants to take «all necessary steps…to preserve the tapes transferred to the Archivist. These steps shall include all necessary preservation copying and the repair and enhancement of any damaged tapes; [and, demonstrate] to the Court that the materials are being stored under conditions that will ensure their preservation and future access….»[38]

Kenneth Thibodeau, the head of NARA's Center for Electronic Records (CER), has provided details on the challenges presented in preserving the materials that had been transferred to NARA in January 1993. Initially, NARA's Office of Presidential Libraries was given custody of the materials. It soon became evident that they did not have the capability to handle the preservation demands and the Acting Archivist soon handed responsibility for these materials over to the CER. Unfortunately, the CER also had no systems capable of either reading or copying these materials. Fortunately, though, they had recently awarded a contract for an in-house preservation system and although the desired system had not yet been developed, the contractor was able to provide some technology to assist the CER in its efforts. The CER had previously, and still does, only accession electronic media that is handed over to them in a hardware and software independent environment. All of the materials transferred to their possession in this instance was hardware and software dependent. In order to develop a preservation plan of action the CER decided to examine each one of the nearly 6,000 tapes and 150 hard drives in their possession. Unfortunately, they had no staff with the appropriate security clearance to review the actual material on the tapes themselves and also had no secure facilities to store the tapes. Prior to this accession the CER had a total of perhaps five reels with classified information on them. According to Thibodeau, the CER eventually received approval to work with these materials «only if we configured the systems so that there was no way to output any of the data.» The CER was not allowed to bring the information on the tapes up on a screen and could not print any of it out. If they ran across an error on a tape, and there were many (see discussion below), there was no way to look at the actual contents of the tape to figure the error out. They just put the tape aside and moved onto the next. The CER did eventually received permission to look at the tapes to decipher the errors that they ran across, however, once they began analyzing the errors they discovered that some tapes did not conform to industry standards and that they often exhibited properties that the CER did not even know were possible for tapes.[39]

From the Archives perspective, they were tasked with a chore of monumental proportions that literally dwarfed all of their previous work. According to Thibodeau, «there is no comparison between everything we had ever done and what we did in that short time period with the PROFS [materials].» Copying the files off the NSC's personal computer hard drives alone was, by volume, larger than everything the CER had collected over the previous twenty years. Copying the information on the hard drives was complicated by the fact that the NSC's removable hard drives were «handcrafted hard drives» not compliant with industry standards. Fortunately for the CER, the NSC had saved one of each of the five different types of personal computers required to load and read these hard drives. In order to retrieve the data off of them the CER had to place the hard drives back into the appropriate personal computer and then output the contents onto industry standard removable hard drives. Backup tapes had their problems as well. Creased tapes had to be ironed, ripped tapes had to be spliced, and tapes with unwanted moisture on them had to be literally baked. There was only a 5 degree Fahrenheit window of opportunity for correctly baking a tape and exceeding that range would cause damage the tape. In order to properly calculate that 5 degree window, the CER had to know the specific chemical makeup of a specific tape based on the manufacturer's batch number because different batches of the same make of tape could have a different chemistry. In addition, in order to be able to copy a backup tape one needed to know the

system configuration at the time that the backup was made in order to properly reload the tape and read it. Sometimes even getting that far was not enough. At times the CER had to contact the tape manufacturer in order to understand how particular types of tape stored the date of the backup. All of these types of preservation issues had never been dealt with by the CER. Previously, if an agency sent a bad tape to the CER they returned it and required the agency to submit a new readable copy. Given that these tapes came to the CER as part of a Court order and that they documented various iterations of the defendants systems at different points in time over the previous decade – iterations that no longer existed – they had no such luxury. And all the while that the CER was performing these preservation tasks they did not receive any new appropriations to offset their costs. At one point, CER-head Thibodeau issued a stop work order for the CER's non-PROFS work because it was expending its budget too rapidly on PROFS related activities. Despite all of these challenges, the CER was actually quite successful in copying the materials. According to Thibodeau, the eventual success rate was over 99%. This success, however, did not come without a serious cost to the CER's other work. Thibodeau reports that the PROFS work essentially «brought the rest of the [CER's] program to a stop…I basically had to tell staff [to] stop accessioning stuff [from the rest of the government] because we [could] not do anything with it if we accessioned it. All of our capability was going to PROFS.»[40]

What is striking when viewing the above from an international perspective is that all other national archival electronic archiving programs in Europe have committed fewer human resources than was available to the CER throughout the above litigation.[41] This would seem to indicate that any national European program finding itself in a similar situation would likely see itself quickly overwhelmed to not only cope with the mass of data but also to keep its other program efforts moving forward.

## Conclusion

After nearly a decade of litigation the U.S. federal government is still attempting to manage its electronic records in a manner compatible with federal recordkeeping law. The issues raised by the various legal battles point out salient features of electronic records management and electronic archiving that are relevant to any institution seeking to effectively manage its computer generated information resources. The most salient to policy lessons from the PROFS-related litigation include the following salient points:

- Electronic mail software can produce official government records.
- Computer systems need to accommodate an electronic recordkeeping functionality at the front end during systems design if back end archival processing and digital preservation is to be accomplished in a timely and economical manner.
- This electronic recordkeeping functionality needs to be able to create and/or capture metadata that identify record status and provide for appropriate subject, function, and genre classification.
- Policies that rely on print to paper can strip out critical systems metadata and also can violate the law if the printout is used as a justification for deleting electronic versions.
- Backup tapes are not a suitable format for archival preservation.

- Archival management of electronic records needs to explore strategies and tactics that retain original systems functionalities as hardware and software independent environments may decontextualize records and harm their evidential value and authenticity
- Attempts at salvage archiving of computer-generated data are likely to require resources that are beyond what is available in most institutions and will likely be unsuccessful unless substantial additional resources can be concentrated on the salvage effort. Such efforts, though, are likely to significantly hamper other electronic archiving program elements, especially the critical need to address electronic archiving issues at the front end of the records life cycle.

---

ENDNOTES:

[1] Armstrong, et al. v. Executive Office of the President, et al., (Civil Action No. 89-1042), Declaration of George Van Eron, February 6, 1989, p. 2. Since 1979, Van Eron had been the Director of the NSC Secretariat, the «information management office of the NSC»; United States of America v. Poindexter (Criminal No. 88-0080-01), Volume IX, Transcript of Trial, Morning Session, March 15, 1990. Testimony of Kelly Williams, pp. 1722-1814. Williams served in the White House Communications Agency (WHCA) – a joint U.S. military organization which provides communications support to the President and other entities within the Executive Office of the President.

[2] See: U.S. Congress. House. Select Committee to Investigate Covert Arms Transactions with Iran, and U.S. Congress. Senate. Select Committee on Secret Military Assistance to Iran and the Nicaraguan Opposition, *Report of the Congressional Committees Investigating the Iran-Contra Affair, with Supplemental, Minority, and Additional Views* (Washington, D.C.: Government Printing Office, 1987), p. 138.

[3] United States of America v. Poindexter (Criminal No. 88-0080-01), Volume IX, Transcript of Trial, Morning Session, March 15, 1990. Testimony of Kelly Williams, pp. 1755-1765, 1800-1801.

[4] U.S. National Security Council, Memorandum for the NSC Staff from Grant S. Green, Jr., «PROFS and A1,» March 5, 1987. Undated fact sheet entitled «PROFS/VAX» attached. The telephone analogy was to become a primary argument by the government during the PROFS litigation

[5] Armstrong, et al. v. Executive Office of the President, et al. (Civil Action No. 89-0142), Declaration of Gordon Riggle, February 6, 1989. Riggle served as the Director of the Office of Administration within the Executive Office of the President from September 1987 through the close of the Reagan administration (January 20, 1989).

[6] U.S. National Security Council, PROFS note from Paul Schott Stevens to All PROFS Users, «PROFS Notes,» January 21, 1988 at 15:34:13.

[7] U.S. Executive Office of the President. Office of the White House. Office of Counsel to the President, Memorandum for White House Staff from Arthur A. Culvahouse, Jr., Counsel to

the President, «Presidential Records Act Obligations of Departing White House Staff,» December 1, 1988.

[8] Interviews with Eddie Becker, Scott Armstrong, and John Fawcett, July 25, 1995, May 27, 1997, and May 27, 1997 respectively. At that time Becker was a researcher and consultant with the NSA, Armstrong was the NSA's Executive Director, and Fawcett was NARA's Director of Presidential Libraries. As a general rule, NARA only considers 1-2% of all the records that are ever created by the federal government to be worthy of archival preservation.

[9] 44 United States Code §§ 2201-2207. The PRA was enacted in 1978 and established for the first time that the records created by a President and his staff were the property of the United States, not the President's to do with as he pleased. A «presidential record» is defined in the PRA as «documentary materials, or any reasonably segregable portion thereof, created or received by the President, his immediate staff, or a unit or individual of the Executive Office of the President whose function is to advise and assist the President, in the course of conducting activities which relate to or have an effect upon the carrying out of the constitutional, statutory, or other official duties of the President.»

[10] 44 United States Code §§ 3301 et. seq. This act is part of what is commonly considered the broader Federal Records Act (44 United States Code §§ 2101-2118, 2901-2909, 3101-3107, and 3301-3324). Since the passage of the Federal Records Act in 1950, the FRA has come to embody a series of interlocking statutes that cover the entire lifecycle of a record – creation, use, management, and disposition – and that stipulate recordkeeping responsibilities for both NARA and federal agencies. The FRA defines a federal «record» as including «all books, papers, maps, photographs, machine readable materials, or other documentary materials, regardless of physical form or characteristics, made or received by an agency of the United States Government under Federal law or in connection with the transaction of public business and preserved or appropriate for preservation by that agency or its legitimate successor as evidence of the organization, functions, policies, decisions, procedures, operations, or other activities of the Government or because of the informational value of data in them. Library and museum material made or acquired and preserved solely for reference or exhibition purposes, extra copies of documents preserved only for convenience of reference , and stocks of publications and of processed documents are not included.»

[11] 5 United Sates Code § 551 et. seq. The APA was enacted in 1946 to govern «practice and proceedings before federal administrative agencies.» See: *Black's Law Dictionary, Abridged Sixth Edition* (St. Paul, Minnesota: West Publishing Co., 1991).

[12] 44 United States Code § 2203.

[13] 44 United States Code §§ 3101 and 3102.

[14] 44 United States Code §§ 2904 and 3303.

[15] 5 United States Code § 551.

[16] Armstrong, et al. v. Executive Office of the President, et al. (Civil Action No. 89-0142), Complaint for Declaratory and Injunctive Relief, January 19, 1989.

[17] Armstrong, et al. v. Executive Office of the President, et al. (Civil Action 89-1042), Transcript of Temporary Restraining Order Hearing Before the Honorable Barrington D. Parker, United States District Judge, January 19, 1989.

[18] Armstrong, et al. v. Executive Office of the President, et al. (Civil Action 89-1042), Temporary Restraining Order, January 19, 1989.

[19] Armstrong, et al. v. Executive Office of the President, et al. (Civil Action No. 89-0142), Supplemental Brief in Support of Defendants' Motion for Summary Judgment, June 12, 1992.

[20] Armstrong, et al. v. Executive Office of the President, et al. (Civil Action No. 89-0142), Plaintiffs Opposition to Defendants' Motion for Summary Judgment and Memorandum in Support of Cross-Motion for Summary Judgment on the Adequacy of Defendants' Recordkeeping Guidelines and the Archivist's Failure to Perform His Statutory Duties, July 6, 1992.

[21] Armstrong, et al. v. Executive Office of the President, et al. (Civil Action No. 89-0142), Reply Memorandum in Support of Defendants' Motion for Summary Judgment and in Opposition to Plaintiffs' Cross-Motion for Summary Judgment on Counts II and IV (Recordkeeping and Archivist's Duties), August 8, 1992.

[22] Armstrong, et al. v. Executive Office of the President, et al. (Civil Action No. 89-0142), Opinion, January 6, 1993. (810 F.Supp. 335).

[23] Armstrong et al., v. Executive Office of the President, Office of Administration et al., (Civil Action Nos. 93-5002, 93-5048, 93-5156, 93-5177), Opinion, August 13, 1993. (1 F.3d 1274).

[24] U.S. National Archives and Records Administration, «Electronic Mail Systems, Notice of Proposed Rulemaking,» *Federal Register*, March 24, 1994, pp. 13906-13910.

[25] U.S. National Archives and Records Administration, «Electronic Mail Systems, Final Rule,» *Federal Register,* August 25, 1995, pp. 44633-44642.

[26] Interview with Michael Tankersley, May 28, 1997.

[27] U.S. National Archives and Records Administration, «Introduction to General Records Schedules,» Transmittal No. 7, August 1995.

[28] Public Citizen, Inc., et al. v. Carlin et al. (Civil Action No. 96-2840).

[29] United States. National Archives and Records Administration, «General Records Schedule 20 -- Disposition of Electronic Records,» August 14, 1995. Available June 1, 1998 at: gopher://gopher.nara.gov:70/00/managers/federal/grsfr.txt.

NARA rationale for their position is captured in the following passage: «For records to be useful they must be accessible to all authorized staff, and must be maintained in recordkeeping systems that have the capability to group similar records and provide the

necessary context to connect the record with the relevant agency function or transaction. Storage of electronic mail or word processing records on electronic information systems that do not have these attributes will not satisfy the needs of the agency or the needs of future researchers….Search capability and context would be severely limited if records are stored in disparate electronic files maintained by individuals rather than in agency-controlled recordkeeping systems. Furthermore, if electronic records are stored in electronic information systems without records management functionality, permanent records may not be readily accessible for research. Unless the records are adequately indexed, searches, even full-text searches, may fail to find all documents relevant to the subject of the query. In addition, numerous irrelevant temporary records, that would be segregable in systems with records management functionality, may be found. Agency records can be managed only if they are in agency recordkeeping systems….The respondents who [criticized GRS 20] mistakenly concluded that the proposed GRS 20 authorized the deletion of valuable records. On the contrary, GRS 20 requires the preservation of valuable records by instructing agencies to transfer them to an appropriate recordkeeping system. Only after the records have been properly preserved in a recordkeeping system will agencies be authorized by GRS 20 to delete the versions on the electronic mail and word processing systems. As indicated, most agencies have no viable alternative at the present time but to use their current paper files as their recordkeeping system. As the technology progresses, however, agencies will be able to consider converting to electronic recordkeeping systems for their records….The critical point is that the revised GRS does not authorize the destruction of the recordkeeping copy of the electronic mail and word processing records. The unique program records that are produced with office automation will be maintained in organized, managed office recordkeeping systems. Federal agencies must have the authority to delete the original version from the "live" electronic information system to avoid system overload and to ensure effective records management.»

[30] Public Citizen, Inc., et al. v. Carlin et al. (Civil Action No. 96-2840), Complaint for Declaratory and Injunctive Relief, December 23, 1996. Other plaintiffs in the case include the: National Security Archive, American Historical Association, American Library Association, Center for National Security Studies, Organization of American Historians, Scott Armstrong, and Eddie Becker. In addition to Archivist of the United States John Carlin, other defendants include the: Executive Office of the President (EOP), the EOP's Office of Administration, and the Office of the U.S. Trade Representative.

[31] 44 United States Code §§ 3301-3324. Friedman specifically pointed to section 3303(a)d which authorizes the Archivist to «promulgate schedules authorizing the disposal, after the lapse of specified periods of time, of records of a specified form or character common to several or all agencies if such records will not, at the end of the periods specified, have sufficient administrative, legal, research, or other value to warrant their further preservation….»

[32] Public Citizen, Inc., et al. v. Carlin et al. (Civil Action No. 96-2840), Opinion and Order, October 22, 1997.

[33] Public Citizen, Inc., et al. v. Carlin et al. (Civil Action No. 96-2840), Memorandum Opinion and Order, April 9, 1998.

[34] U.S. National Archives and Records Administration. Electronic Records Work Group, «Summary of Draft Proposals for Action on the Management of Electronic Records for May 18, 1998, Public Meeting,» May 18, 1998. Available June 1, 1998 at: **Erreur! Signet non défini.**.

[35] U.S. National Archives and Records Administration. Electronic Records Work Group, «Public Meeting of May 18, 1998. Available June 1, 1998 at: http://www.nara.gov/records/grs20/minutes.html.

[36] Armstrong, et al. v. Executive Office of the President, et al. (Civil Action No. 89-0142), Archivist's Post-Hearing Submission, January 28, 1993. Another part of this submission reported on an agreement between the Archivist and soon to be ex-President Bush wherein the Archivist provided Bush with «exclusive legal control of all Presidential information, and all derivative information in whatever form, contained on the materials.» This aspect of the agreement led to another and separate lawsuit which resulted in the judicial determination that the Archivist had violated both the Presidential Records Act, portions of Article II of the U.S. Constitution (under the Constitution the Archivist was obligated only to the sitting President, not a former President who was now a private citizen), and that the «Archivist's decision to enter into the agreement…was arbitrary, capricious, an abuse of discretion, and contrary to law.» American Historical Association et al. v. Trudy Peterson, in her official capacity as Acting Archivist of the United States, and George Bush (Civil Action No. 94-2671), Memorandum Opinion and Order, as amended, February 27, 1995 (876 F.Supp 1300).

[37] U.S. National Archives and Records Administration, «Armstrong Materials Task Force – Summary Report of Actions,» February 16, 1993.

[38] Armstrong, et al. v. Executive Office of the President, et al. (Civil Action No. 89-0142), Order, May 21, 1993.

[39] Interview with Kenneth Thibodeau, May 30, 1997. Thibodeau has been the Director of the Center for Electronic Records since December 1988.

[40] Ibid.

[41] See: International Council on Archives. Committee on Electronic Records, Electronic Records Programs: Report on the 1994/1995 Survey (Paris, France: International Council on Archives, February 1997).