INFORMATION TECHNOLOGY
CHALLENGES FOR LONG-TERM
PRESERVATION OF ELECTRONIC
INFORMATION

VIVECA ASPROTH
Dept. of Information Technology and Media
Mid Sweden University
SE-831 25 Östersund, Sweden
viveca.asproth@miun.se

Abstract
Information technology in combination with changes in organizational structures and methods of working has led to an increasing amount of information and also to totally new forms of information as well as in the creation of records. Furthermore the e-government services, where the whole idea is to exchange information electronically, will speed up this development. A 10-year period is a long time when dealing with the preservation of electronic records. This is of course due to the rapid technological development, for example the problems associated with obsolete software and file formats. However, since archives are records of permanent and enduring value, it is essential that methods must be developed to ensure that digital records can be preserved for a much longer time. Archival practice has for a long time been intertwined with information technology, dealing with the production, organization, storage, dissemination, and retrieval of records, using the available technology. This implies that the responsibility for research and development should lie upon researchers from both these domains in collaboration. In this paper some problem domains of long-term preservation and how ongoing research matches these domains is presented. It appears clear that although much of the challenges associated with digital preservation is strategic, organizational, and structural and not only technical, the research mainly concentrates on solving the technical issues.

Keywords: digital records, digital archives, electronic information, preservation

1. Introduction
Information technology in combination with changes in organizational structures and methods of working has led to an increasing amount of information and also to totally new forms of information and in the creation of records. Much of the information previously produced on paper is now being produced in electronic form, for example as e-mail and in databases. The Dutch archive theoretician Eric Ketelaar [1999] has invented the word “archivalisation” to describe the phenomenon where more and more material is deemed to be worthy of documenting and archiving. This means that a considerably larger amount of information requires handling. Electronically generated and stored information is also exposed to change and manipulation in quite a different way to that involved with paper-bound information. The e-government
services, where the whole idea is to exchange information electronically, will radically redirect paperbound information and documents towards electronically generated and managed information and documents. A 10-year period is a long time when dealing with the preservation of electronic records. This is of course due to the rapid technological development, and includes, for example, the problems associated with obsolete software and file formats. Goldstein [2004] stresses the need for a quick solution to the problem of archiving electronic information for the future. Furthermore, he states that the major threat to archived material today is not the fire hazard but the rapid development of different file formats for documents, sound and images.

Vital functions in society such as social insurance and national registration have to be accessible for long periods of time. Without authentic, accessible, searchable, and reliable records these functions in society are in fact impossible to maintain. The implication of technical developments and the transition to electronic media is that traditional methods have become obsolete or at least insufficient.

Access to the correct information is essential in any democracy. The rapid growth of public electronic services means that there will be an increasing amount of documentation and information, coming in to or being generated by the authorities, which will be in electronic form. This information must be handled and stored by the authorities and at a later stage be sorted out or transferred to long-term archives. Viewed from a longer perspective, it is also important that these electronic records are preserved as a cultural heritage and collective memory of human activity.

Companies are not ruled by the same regulations as authorities in spite of their requirements to store electronic documents such as contracts, patents, or blueprints for 30 to 40 years or even longer. However, the same requirements regarding accessible, authentic, and reliable information are valid for the companies. Both companies and authorities are faced with the same general problem concerning preservation and accessibility of electronic information.

Unlike paper-bound information, the means of preserving electronic information is determined at its creation, which is why actions must be taken at an early stage. Beyond the problem associated with the physical storage, the information must be structured in such a way so as to make it accessible during its entire preservation time. At present it is unusual for special consideration to be given to these needs in information systems.

Archival practice has for a long time been intertwined with information technology, dealing with the production, organization, storage, dissemination, and retrieval of records, using the available technology. The implications of this are that the responsibility for research and development should be a collaboration between researchers from both these domains. In this paper the problem domain of long-term preservation and how ongoing research matches this domain is presented.

2. Archival information

The time horizon for the storage of electronic information stretches from a second to eternity. The distinction between business data and long-time preserved information becomes more and more diffuse. Archivists use the word record to separate archival information from information in general. Thomassen [2001] defines a record as “information that can be retrieved in the form of a document, i.e. the smallest unit of data that can independently function as information. Records are process-bound information, that is to say, information that is generated by and linked to work processes”. The opinion that records are made or received in some activity is also
expressed by Duranti [2000]. The International Council of Archives [2000] adds the
pursuance of legal obligations. Hoffman’s [1998] definition of a record includes
content, context (the creation process) and structure (form or manifestation). In
Livelton’s [1996] definition all these aspects are included. “All books, papers, maps,
photographs, or other documentary materials regardless of physical form or
characteristics, made or received by any public or private institution in pursuance of
its legal obligations or in connection with the transaction of its proper business and
preserved or appropriate for preservation by that institution or its legitimate successor
as evidence of its function, policies, decisions, procedures, operations, or other
activities or because of the informational value of the data contained therein”
[Livelton, 1996].

However, there appears to have been a paradigm shift in the view of electronic
are changes caused by technological development. These changes affect archival
methods and principles. In paperbound records the content and the context exist
together, while in electronic documents they are in fact separated and stored
independently of one other. Therefore documents do not exist as such, except in the
moment of capturing. Information is stored in bits and pieces, and can only, with help
of software, be connected and transformed to a document as required. An example is a
digital map, which consists of numerous objects organized in different layers. The
map, when it appears, is merely a view constructed at that moment for some specific
purpose.

The theoretical basis for archives is the principle of provenance, which means
that you must be able to trace the information back to its origin. The original function,
source and purpose must be evident. In spite of the paradigm shift the principle of
provenance is absolutely essential with reference to electronic records. When
reviewing recent literature in archival theory and practice some qualities or
characteristics of electronic information seem to be essential:

- **Evidence**
  Cox [2001] has noticed that information technology has resulted in a deviation
from traditional archival principle for records (the principle of provenance). He
 criticizes this change and is of the opinion that the original requirements with
regards to records must remain in spite of the technological development. One
important requirement is evidence. To fulfill these requirements, you must, for
example, be able to store and connect digital signatures and multiple signatures
to the record as long as the record is stored.

- **Reliability**
  Reliability refers to the ability of a record to be valid for the stated facts, i.e. the
trustworthiness of the record as to content. Duranti [2000]

- **Authenticity**
  The problem of authenticity is strongly connected to electronic information.
Duranti [2001a] considers information that is produced electronically very
difficult to validate.

- **Accountability**
  “Archives and record management share a simple goal: providing for
the concepts of evidence and accountability as being symbiotic. The possibility to trace and identify who is responsible for the information is essential for a democratic process. [Barata and Cain, 2001]

- **Readability and availability**
  Thomassen [2001] describes readability and availability as one of the objectives for archival theory. It must be possible to read and understand the content after a long period of time. The quality of process-bound information depends on the quality and the stability of the bond that links content data and generating processes. Only when content data can be retrieved and analyzed in an adequate form, in an adequate structure and within the context of its provenance (i.e. only when content data are linked to the relevant metadata) can the information provided have its intended quality. [Thomassen, 2001]

These characteristics are all concerned with availability and trustworthiness.

### 3. Problem domains

Within the area of long-time preservation of electronic information there are many different problems remaining to be solved. The problems concern technical issues but also organizational and legal issues. Karjalainen et al [2000] has in their study found that Enterprise Document Management systems can be seen as a comprehensive framework for managing information resources, including both human and technological systems in a holistic way. This view covers organizational and legal issues as well as the technological challenges.

#### 3.1. Technical problems

Issues at stake in this domain are of a variety of types, covering hardware, software and formats. The large amount of electronic records, which require to be stored for active use, in turn demands new mass storage media. Another problem is the fragility of the media which leads to low durability. There is often also a lack of knowledge concerning the durability of different media for storage. In addition, poor security exists for many electronic files [Wilson, 2000].

The software problem concerns the rapid development of new versions of the software product and occasionally changeovers to new software products. In addition, the software is not always appropriate for the purpose. Guercio [2000] writes that the technology industry ignores archive requirements. She claims that for years vendors have sold public administrations products that are not really useful.

Many software systems do not handle different versions of data or do not encourage storage of different versions. Dollar [1992] states that a major archival problem concerning most geographic information systems applications is the general practice of updating maps electronically without saving previous versions.

Connected to the problem of rapidly changing software, is the additional problem of changing formats, the consequence of which in the long run can be that archival information become unreadable [Gibbs and Heazlewood, 1999].

The archivists have used different strategies for the technical refreshment of digital information. The strategies cover a range from rather simple copying to migration and emulation. Marcum [1996] identifies the following strategies:
• Medium refreshing – copying data from a physical carrier to another physical carrier of the same type
• Medium conversation – transferring data from one medium to another medium
• Format conversation – converting the data format
• Migration of technical environment (hardware and software)
• Emulation of technical environment – systems which run in a new operating environment but emulate a previous obsolete environment

The available preservation strategies have their strengths and weaknesses. We must be aware of the fact that we have to “choose what to lose” [Rothenberg, 2000]. Until reliable solutions exist, people responsible for archiving will oppose a change from paperbound to electronic [Warner, 2002]. The drawback concerning both migration and emulation is that information can be lost. Repeated migration or emulation increases the risk of loss of information. [Mannerheim, 2000]

3.2. Legal issues
The rapidity and ease with which changes can be made means that there are challenges associated with ensuring the continued integrity, authenticity, and provenance of digital materials [Beagrie, 2002]. Several authors emphasize the problem of how to ensure authenticity and how to preserve the evidence [Bearman, 1994; Duranti, 2001b; Rothenberg, 2000; Wilson, 2000].

3.3. Organizational problems
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. [Marcum, 1997]

There are many different stakeholders over the lifecycle of digital resources and uncertainties over who should be responsible for preservation activities [Beagrie, 2002]. At the Erpanet training seminar in Marburg it was stated that: “senior management are not aware of the consequences if their organizations do nothing” [Erpanet, 2003]. Meijer [2001] agrees with this, and states that the consequences of poor electronic management are not well understood. At the InterPARES symposium Barbedo [2000] summarized the results from a Portuguese study which found that agencies have no strategies at all for maintaining continuous access to electronic records.

Barbedo [2000] argues that records management is not fully recognized as an organizational function. He also claims that there exist organizational and administrative problems that hinder the implementation of correct record keeping. Bearman [1994] and Meijer [2001] write about problems of information sharing that make it unclear as to which information belongs to which organization.

Many authors claim that there is a lack of collaboration between archivists and IT specialist [Cox, 2000; Guercio, 2000; Margaret Hedstrom, 1996; MacFarlane, 2000]. “Archives, government organizations and IT solutions providers need to work together to implement sound systems and procedures to maintain authentic records “ [MacFarlane, 2000]. Archivists also need to broaden their competence. They have insufficient knowledge concerning information technology [Cox, 2000]. Gilliland-Swatland [2000] thinks that there is a need for a new kind of professional, namely one with information technology competence as well as the archivist competence. Wettengel [1996], on the other hand, says that “the archivist should not work with
creation, they should remain the keepers of the records” [Wettengel, 1996]. Though
the latter view is changing, it is still embraced by several archivists.

3.4. Context and metadata
How is it going to be possible for someone in the future to understand the context of the
information and the processes that created it? This is a question of metadata, a
question now considered to be of much greater importance in the digital age. How
can we determine what amount of evidential historicity is required and how to capture
and retain it [Bearman, 1994]? “Where should the necessary metadata reside (within
the digital information system, in paper form or both)” [Gilliland-Swetland, 2000]. A
document consists of both medium and content. For paperbound documents the paper
is the medium. With virtual documents, it is now the metadata which form the real
medium [Delmas, 2001].

4. What is done about the problems?
There is a great deal of ongoing research which aims to solve the technical problems
associated with long-term preservation of digital information. However the projects
are very seldom directed towards this particular area. An example of research on new
mass storage media is found in Kalbitzer [2004]. The project CAMILEON (Creative
Archiving at Michigan & Leeds: Emulating the Old on the New) deals with the
development and evaluation of different technological strategies for long-term
preservation [Holdsworth and Wheatley, 2001]. Quite a number of papers concern
research about digital signatures and other security issues.

Although many authors stress that the problems are not only technical and that
the other issues (organizational, legal, context, etc.) are just as important, at present,
there has been very little progress in this area.

4.1. International research projects
A few research projects are focusing on the long-term preservation of electronic
information, although they are generally dealing with problem solving rather than
systematical knowledge development. The more important projects include:

- InterPARES\(^1\), CEDARS\(^2\), CAMILEON\(^3\), , the Digital Archives Research Project
  at Edinburgh University Library and the UK National Archives in Great Britain
- NEDLIB\(^4\) in the Netherlands
- The EU project MINERVA\(^5\)
- PANDORA\(^6\) and VERS\(^7\) in Australia
- OAIS\(^8\), ERA\(^9\) at the US National Archives and Records Administration
  (NARA) and San Diego Supercomputer Centre at the University of California
  San Diego in US

---

1 http://www.interpares.org/
2 http://www.leeds.ac.uk/cedars/
3 http://www.si.umich.edu/CAMILEON
4 http://www.kb.nl/coop/nedlib/
5 http://www.minerva-europe.org/
8 http://ssdooo.gsfc.nasa.gov/nost/isoas/
9 http://www.archives.gov/era/
4.1.1. **InterPARES**
The most prominent and far-reaching research project is the International Research on Permanent Authentic Records in Electronic Systems (InterPARES), which aims at developing the theoretical and methodological knowledge essential to the long-term preservation of authentic records created and/or maintained in digital form. This knowledge should provide the basis from which to formulate model policies, strategies, and standards capable of ensuring the longevity of such material and the ability of its users to trust its authenticity. InterPARES is now in its second phase, InterPARES II, which in addition to dealing with issues of authenticity, delves into the issues of reliability and accuracy from the perspective of the entire life-cycle of records, from creation to permanent preservation. It focuses on records produced in complex digital environments in the course of artistic, scientific and e-government activities.

4.1.2. **CEDARS**
The main objective of the CEDARS (Curl Exemplars in Digital ARchiveS) project was 'to address strategic, methodological and practical issues and provide guidance in best practice for digital preservation. The project ended in March 2002.

4.1.3. **CAMILEON**
The CAMILEON (Creative Archiving at Michigan & Leeds: Emulating the Old on the New) project is a joint undertaking between the Universities of Michigan (USA) and Leeds (UK) and is funded by JISC and NSF. The project aims to explore the options for long-term retention of the original functionality and ‘look and feel’ of digital objects, to investigate technology emulation as a long-term strategy for long-term preservation and access to digital objects, and to consider where and how emulation fits into a suite of digital preservation strategies.

4.1.4. **NEDLIB**
NEDLIB (Networked European Deposit Library) is a collaborative project of European national libraries. It aims to construct the basic infrastructure upon which a networked European deposit library can be built. The objectives of Nedlib concur with the mission of national deposit libraries to ensure that electronic publications of the present can be used now and in the future.

4.1.5. **MINERVA**
The MINERVA (Ministerial NEtwoRk for Valorising Activities in digitisation ) Web Preservation Project was established to initiate a broad program to collect and preserve primary source materials. A multi-disciplinary team of Library staff representing cataloging, legal, public services, and technology services is studying methods to evaluate, select, collect, catalog, provide access to, and preserve these materials for future generations of researchers.

4.1.6. **PANDORA**
The overall goal of the PANDORA (Preserving and Accessing Networked Documentary Resources of Australia) project is to develop and implement procedures for the capture, archiving, and provision of long term access to online electronic Australian publications selected for national preservation.
4.1.7. **VERS**

VERS (Victorian Electronic Records Strategy) offers a solution to the problem of capturing, managing and preserving electronic records. VERS is a framework of standards, guidance, training, consultancy and implementation projects, which is centered around the goal of reliably and authentically archiving electronic records.

4.1.8. **ERA**

The Electronic Records Archives (ERA) aims to be a comprehensive, systematic, and dynamic means for preserving virtually any kind of electronic record, free from dependence on any specific hardware or software.

4.2. **Standardization**

Standardization is an important issue for long-term preservation. There are some standards developed for records management.

- ISO 15489-1 is a standard for Information and Documentation – Records management. This standard serves as a checklist in the document management process.
- ISO 15489-2 is an implementation guide for ISO 15489-1. ISO 15489-2 is setting out a process called DIRS (Developing and Implementing Records Systems) that uses a systems management and project management process.
- ISAD(G): General International Standard Archival Description
- ISAAR(CPF): International Standard Archival Authority Record for Corporate Bodies, Persons, and Families
- The Reference Model for an Open Archival Information System (OAIS) is a conceptual framework for an archival system dedicated to preserving and maintaining access to digital information over the long term.

Many archivists see XML as the standard that will solve the format problem. However, XML cannot deal with the whole range of information types that exist at present and the great variety that will certainly emerge in the future.

4.3. **Swedish Research Projects**

Two research projects were initiated at approximately the same time in Sweden. The projects have slightly different aims. As the projects supplement each other, collaboration has been established.

4.3.1. **The LDB project**

The aim of the LDB (Long-term digital preservation) project is the development of methods and techniques for long-term digital preservation for archives in Sweden, and to answer the question concerning how digital information can be prepared and transferred to archival institutes. The project is a collaborative effort between the Technical University of Luleå and the Swedish National Archive. The project includes systems development as well as research. Deliveries of electronic records from authorities to the Swedish National Archive are used as research objects and a new model based on the OAIS reference model will be elaborated.

The major element of the LDB project concerns finding technical solutions and the focus is on both the organizational issues and metadata.

---

10 http://ldb.project.ltu.se/~Projekt_LDB
4.3.2. Archives for the Future

The purpose of the project is to systematically study the electronic records management in a number of Swedish governmental and business organizations. Examples of research fields included in this project are the investigation and comparison of the need for electronic records and archives management from organizational, informational, economic, and jurisdictional demands; to map and compare the existing strategies of the organizations for the management of electronic records; and to identify differences between the information structures, which are based upon models from systems design, and those that are based upon archival theory. One current study concerns the processes by which the electronic information is created or changed and whose long-term aim is to develop a metadata model. Another current study aims to identify the needs for long-term preservation in public electronic services.

This project is a collaborative effort between the Archives and Information Science and Informatics departments at Mid Sweden University. A broad range of companies and organizations (IT companies, forest industry companies, municipalities, county councils, governmental and civil service departments, governmental companies, etc) participate in a reference group.

The Archives for the Future project adopts a pro-active position and attempts to deal with the long-term preservation issues for electronic records at the point of creation. Thus metadata concerning the process(-es) when the records are created must be captured and all the quality aspects of the provenance principle must be acquired. The project will develop a model concerning not only how to deal with this aspect but in addition will study the organizational and legal aspects of pro-active digital records preservation.

5. Concluding remarks

There has been a massive increase in the volume of information being produced and is therefore subject to storage for either the short or long-term. Furthermore, the increasing development of public electronic services, adds information to the already immense volumes. Many problems still remain to be solved before secure and efficient long-term preservation exists. It appears that although many of the challenges associated with digital preservation are strategic, organizational, and structural and not merely technical, present research is concentrated on the solution of technical issues. Digital preservation requires, in addition to the technological development, elaborate strategies, new workflows and organizational structures, standards and common metadata, new specific competences as well as close cooperation across different professions from traditional preservation management to computing science. Thus, it is time for informatics and computer science researchers to address the issue of electronic information preservation more convincingly. The above issues are all interconnected and mean that a radically different approach is required in order to manage digital information.

References


http://gathering.itm.mh.se/framtidensarkiv/


