InterPARES

Archival Research into the Preservation of Complex Digital Records

Yvette Hackett
Library and Archives Canada

December 9, 2004
Kyoto, Japan

Today, I will discuss the research being done by the InterPARES project. To put this work into context, it is important to understand that the current project - called InterPARES 2 - is actually the 3rd project in a series of archival investigations of digital materials. Therefore, I will begin with an overview of the findings of the earlier projects, since these results inform the current work.

It is important to emphasize here that InterPARES research is primarily focussed on records originally created in electronic form; in other words, these are document with no original analog version, which are usually paper-based. While the technology of microfilming and digitizing differ, both produce copies, and the issue of how copies of records are to be dealt with were addressed when microfilming was first introduced into libraries, archives and the workplace. While there may be financial benefits to libraries and archives in preserving such digitized copies, they are of secondary importance as long as the original analog record survives.

The role of digitized copies is frequently an "access" role, both in the workplace and in archival institutions. Digitized copies can be made available more easily and to a much wider group of employees or researchers than analog originals which can only be consulted on site, usually by one person at a time.

Some archives do refer to the "preservation" role of digitized copies. It is important to understand that this preservation role is primarily limited to a reduction in the handling of the original once a microfilmed or digitized copy is available. Also, minimizing a record's exposure to high light levels, or to varying temperatures and humidity levels all reduce the damage which an unstable environment can accelerate.

There have been some cases where paper-based originals have been destroyed following microfilming or digitization. The creator of the records may make this decision for operational reasons. Guidelines for trustworthy microfilming or imaging systems already exist[1], and the subsequent reliability of the microfilmed or digitized record will be demonstrated by the creator's continued use of the copy.

Presentation Overview

All 3 projects began at the School of Library, Information and Archival Studies at the University of British Columbia. I emphasize this academic starting point because it has, from the beginning, set the research within a set of assumptions and definitions developed, over several centuries, by the records management and archival disciplines. By enlisting the

---

[1] For example, the Canadian General Standards Board standard CGSB 72.11-93 entitled *Microfilm and Electronic Images as Documentary Evidence*, 1 October 1993. It was amended 1 April 2000.

participation of records creators, archivists and information technologists, the research continually tests traditional knowledge and requirements against current recordkeeping practices, current operational realities in archival institutions, and the ability of the current state of technology to support the needs of the first two groups.

The UBC Project

In 1994, a project entitled "The Preservation of the Integrity of Electronic Records" got underway at University of British Columbia (UBC). It is known variously as The UBC Project, the UBC-MAS project and the UBC-DoD project.

The goal of the first project was "to identify and define conceptually the nature of an electronic record and the conditions necessary to ensure its reliability and authenticity during its active life, based on the concepts and methods of diplomatics and archival science"[2]. In other words, the project focussed on records while still in the hands of the records creator.

Diplomatics

In 1994, diplomatics was still largely unknown in North America. It emerged in 17th century Europe as an analytical technique for determining the authenticity of historical records issued by sovereign authorities. The tenets and methods of diplomatics were first laid out in 1681 by Jean Mabillon, a Benedictine monk. Mabillon examined, among other things, the language of documents, their characteristic parts, their seals, and the systems of chronology used in dating them. Over 4 centuries, diplomatics has been used to help determine a record's authenticity for legal purposes and to assess medieval records as historical sources.

In the early 1990's, Dr. Duranti began to explore the possibility of extending diplomatics to address the growing questions about both ensuring and verifying the reliability and authenticity of contemporary records, including those in electronic form. The results of this study appeared in six articles, originally published in *Archivaria*, the journal of the Association of Canadian Archivists. They have now been published as a book entitled *Diplomatics: New Uses for an Old Science*[3]

The UBC Project attracted the attention of the United States Department of Defense, specifically the Records Management Task Force, who joined the research in 1995 in search of improved methods of records management for both their traditional records, and the growing volume of records being created in electronic form.

Results

-The research concluded that:

- the reliability of electronic records, which is their ability to stand for the facts they are about, is best ensured and maintained by the creating body through control on the form of the records and on their procedure of creation.

---

[2] MacNeil, Heather, "*Trusting Records: Legal, Historical, and Diplomatic Perspectives* (Dordrecht: Kluwer, 2000), 90. For a detailed description of the UBC-MAS Project and its findings, see Luciana Duranti and Heather MacNeil, "The Protection of the Integrity of Electronic Records: An Overview of the UBC-MAS Research Project," Archivaria 42 (fall 1996): 46-67.

[3] Duranti, Luciana, *Diplomatics: New Uses for an Old Science* (Lanham, Maryland: Scarecrow Press in association with the Society of American Archivists and the Association of Canadian Archivists, 1998).

- but that their authenticity, which is the trustworthiness of a record as a record, is best ensured and maintained by the body who has custody of the records through control of the process of transmission and of preservation through time.

These conclusions largely confirmed the continued validity of traditional records management practices, while translating them into an electronic work environment. In practical terms, this translation resulted in a U.S. Department of Defense standard - DoD 5015.2, the Design Criteria Standard for Electronic Records Management Software Applications. This software specification describes the essential characteristics and functionality for a software application designed to control digital objects. In its simplest form, such an application must control 3 things:

- metadata (frequently stored in a database structure) to record information about the document or digital object
- the document or digital object itself
- a link between the two

InterPARES (1999-2001)

Following the success of the first project, it was logical to apply the research methods used by it to the development of methods for the long term preservation of records created in such software applications. This time, the perspective on the records was no longer that of the creator or records manager, but that of the "preserver", a generic InterPARES term used to describe any person or institution responsible for the long-term preservation of authentic records.[4]

Originally referred to as UBC Part 2, project members selected a new name which reflected the international nature of the undertaking. The acronym InterPARES stands for "International Research on Permanent, Authentic Records in Electronic Systems".

Organizational Structure

The research was distributed among 4 tasks forces, each responsible for a specific conceptual area:

- Authenticity
- Appraisal
- Preservation, and
- Strategies

The final reports of each of these Task Forces are available at the InterPARES web site. I will give the URL at the end of my presentation. Each task force would examine the impact electronic records might have on traditional archival principles and procedures. The Authenticity Task Force depended primarily on case studies of systems currently operating in primarily large institutions in both the public and private sectors in a number of countries. The Appraisal and Preservation Task Forces used a type of modelling (IDEF modelling) to define archival functions and then test them against the systems examined in the case studies. The Strategies Task Force used the findings of the other three task forces to develop an intellectual framework for policy development.

Appraisal Task Force

The primary objective of the Appraisal Task Force was to "determine whether the

---

[4] InterPARES Project, "The InterPARES Glossary," available on the InterPARES 2 web site at http://www.interpares.org/book/interpares_book_q_gloss.pdf.

evaluation of electronic records should be based on theoretical criteria different from those for traditional records, and how digital technologies affect the methodology of appraisal"[5]. The Task Force was also interested in how the concept of Authenticity being developed by the Authenticity Task Force would exhibit itself within the archival appraisal function.

The Task Force concluded that:

- appraisal of electronic records should begin when records are still active and continue until the formal transfer of the selected records to the custody of the preserver;
- the medium of records affects the process of appraisal, but not the fundamental task of assigning value; and that
- information compiled during appraisal must be "packaged" and preserved as it is crucial to enable the eventual transfer of records, their preservation and their description.

Preservation Task Force

The Preservation Task Force focussed on what "preservation" means for electronic records and whether it is different than in a paper-based environment.

In its final report, the Preservation Task Force concluded that "it is not possible to preserve an electronic record: it is only possible to preserve the ability to reproduce the record"[6]. As a result of this finding, it extended the concept of the preservation process to include access to the records, because this process of reproduction takes on additional degrees of difficulty in a digital environment.

Their analysis also separates an electronic record into intellectual and digital components (the former being the formal elements of a record, the latter being units of storage within a record, which do not necessarily coincide.

Finally, they suggest replacing the traditional expression "unbroken chain of custody" with "continuing chain of preservation", a more complex concept which attempts to represent the preservation of authentic electronic records as a continuous process that begins with records creation and acknowledges that "the entire process of preservation must be thoroughly documented as a primary means for protecting and assessing authenticity over the long term".

Authenticity Task Force

The work of the Authenticity Task Force has, I believe, attracted the most attention and I will spend a little more time discussing it. Their findings also form the basis of the "authenticity" research in InterPARES 2.

Conceptual Framework for Authenticity

In traditional archival theory and jurisprudence, if records are relied upon by their creator in the usual and ordinary course of business, they are presumed authentic. But in electronic systems, the presumption of authenticity must be supported by evidence.

There are 6 key definitions in InterPARES research.

Document/Record

While a document is simply "recorded information", a record is "any document created

---

[5] InterPARES Project, "Research Plan," available on the InterPARES 1 web site, at http://www.interpares.org/researchplan.htm.

[6] InterPARES Project, "Preservation Task Force Report," p. 6, available on the InterPARES 1 web site, at http://www.interpares.org/book/interpares_book_f_part3.pdf.

(and by this I mean, made or received and set aside for further action or reference) by a physical or juridical person in the course of a practical activity as an instrument and by-product of it.[7]"

Reliability/Authenticity

To be considered reliable, a record must be able to stand for the facts it is about (such as a marriage licence, as a record of the exchange of vows at a ceremony; or a land title as proof of ownership of a certain property).

A record's authenticity relates only to evidence that it is what it purports to be and has not been tampered with or otherwise corrupted. A record can therefore be "unreliable", but authentic. It was created in the ordinary course of business, but misrepresents the facts - but this is an issue for the records' creator.

Authenticity is an absolute concept. A record cannot be "sort of" or "more or less" authentic. But the "presumption" of authenticity moves on a sliding scale. A weak presumption of authenticity would not prevent an archives from acquiring something, particularly if it was unique, but the authenticity issue would be addressed in the archival description. If, following the assessment of authenticity, the presumption is weak, the Authenticity Task Force identified other means of verification, such as comparison to other existing copies of the record, textual analysis, or a study of audit trails).

Identify/Integrity

Authenticity is rooted in the identity and integrity of the record. These two characteristics are defined as follows:

Identity is established by those attributes of a record that uniquely characterise it and distinguish it from other records, while integrity involves the intact articulation of the record's content and its required elements of form.

Findings of the Authenticity Task Force

Having established this framework, and tested its assumptions in the case studies and in traditional archival science, the Authenticity Task Force developed 2 sets of Requirements for Authenticity.

- The first are the Benchmark Requirements Supporting the Presumption of Authenticity of electronic records. These relate to the environment in which the records are created. They are not mandatory, but rather represent a goal for records' creators which can be adopted to the degree that their operations require evidence of reliability and authenticity.
- Second are the Baseline Requirements for the Production of Authentic Copies. These apply to the archival institution (or "preserver"). They are mandatory.

Benchmark Requirement

There are 8 Benchmark Requirements in all. The Identity of a record is ensured by the recording of up to 11 elements of metadata, including significant names, dates, classification numbers, identification of attachments, among others. Integrity requires an additional 4 data elements.

These elements must be expressed in the record, or be inextricably linked to the record, or be consistently recorded in the context of records' creation. This concept has been implemented through the "profile" in electronic records management systems which are

---

[7] InterPARES Project, "The InterPARES Glossary," available on the InterPARES 2 web site at http://www.interpares.org/book/interpares_book_q_gloss.pdf.

compliant with DoD 5015.2.

The balance of the Benchmark Requirements are procedural. They involve the existence of access privileges, of measures for avoiding loss or change due to technological failure or obsolescence, the determination of record forms, etc. The case studies showed that while some of the controls tended to be at least partially implemented within automated systems, the others, if they were present, relied primarily on traditional procedures outside the system. For example, access is frequently controlled by passwords, but also in some circumstances by physical security. Regular backup procedures are probably the most familiar method used to protect against the loss and corruption of records, while we guard against technological obsolescence by regularly upgrading hardware and software and migrating applications to new platforms.

Among the remaining requirements is the need for procedures to authenticate a document. Authentication is a sort of sub-category of authenticity in that it is a statement that a record is what it purports to be at a given moment in time - for example, a method to declare a certified true copy for use in court.

Baseline Requirements

The Baseline Requirements for the Production of Authentic Copies are intended for the preserver and as such are more stringent.

They include controls over records transfer, maintenance and reproduction expressed, for example, as unbroken custody of the records, and adequate security for records in the archival repository; adequate documentation of the impact of the reproduction process on the form, content, structure, accessibility and use of the records in archival custody; and finally, information about significant changes made to the records since their creation, whether performed by the creator prior to transfer, or by the preserver.

In addition to these 3 Baseline Requirements, most of the Benchmark requirements "for the creator" also apply to the "preserver".

InterPARES 2

In the last year of InterPARES 1, participants began turning their attention to non-textual records. Research into the problems being experienced by digital composers illustrated that non-textual electronic records shared many of the same longevity problems being experienced by textual electronic records, such as media obsolescence, lack of backward compatibility and proprietary software formats. But this artistic environment also differed in several ways: it operated on a much different concept of authenticity and authentic performance; it depended on more complex hardware and software relationships extending to MIDI interfaces and synthesizers; and it included "interactive" pieces where a computer improvises a musical accompaniment using algorithmic software, to a performance by a musician or dancer.

Scope of Research

While InterPARES 2 started with a growing interest in music and the arts, the full scope of the project developed into a much larger undertaking which can be divided into 3 sets of threes.

First, the project will examine documents created in the course of artistic activities, scientific activities and government on-line. The scientific community has dealt for many years with the concept of accuracy, particularly as it applies to machine-readable data. Government on-line will examine government's current push to load data, information and records of all kinds to web sites, and to interact with citizens in this environment. As for artistic activities, the project will examine as many varied artistic forms as possible, within the limitations of time and of the expertise attracted to the project.

Secondly, the project is concentrating on documents which it describes as "dynamic, interactive and experiential". Developing adequate definitions for each of these terms is part of research itself.

Next, the project will study these fields from 3 separate perspectives:

1. creation and maintenance
2. preservation and access
3. accuracy, reliability and authenticity

As such, the research agenda encompasses the scope of both the original UBC project and InterPARES 1. It has moved beyond the limitations of the "record" to include a wide range of textual and non-textual documents.

Finally, there are a number of cross-domain areas which will integrate the case study results from the point of view of Archival Description, IDEF Modelling, Policy and Terminology.

General Studies

A number of General Studies are underway. I will briefly describe three of them.

MUSTICA presented an opportunity to become part of a research initiative with two major French research institutes - the Institut national de l'Audiovisuel (INA) and the Institut de recherche et coordination Acoustique/Musique (IRCAM). The project will attempt to identify the various generations of digital files generated during the artistic creation and performance processes, and confirm which ones are necessary for long-term preservation and access. It will also be able to tap into and analyze significant long-term experience at IRCAM with metadata, its creation and subsequent utility in providing long-term access. Finally, the project will attempt to construct a typology of digital music files.

The study of Persistent Archives Based on Data Grids focusses on the San Diego Supercomputer Center's project to develop a prototype for a persistent archives based upon data grid technology designed for the National Archives and Records Administration (NARA) in the United States. The study examines the minimal capabilities required within grid technology to ensure the preservation of digital records.

Another General Study that is currently underway is a survey of file formats and encoding languages that are used for non-textual materials. File formats and encoding language are being analysed to determine structure and properties, as well equivalence classes which would help identify potential conversion tools.

Case Studies

The project has undertaken a number of case studies. A decision was made, early in the project to create a core case-study questionnaire designed to elicit all the information required by the various research groups.

At this mid-point in the project, 22 case studies have been approved. Twelve have been completed and their final reports should be posted to the InterPARES web site between now and February 2005.

Case Studies - Artistic

In the Artistic area of research, it is difficult to be absolutely precise about which artistic disciplines these case studies cover because, as is the case with so many things in the early 21st century, boundaries are blurring. But there is performance art, theatre, dance, moving images, installation art, music, and on-line publication. In several cases, we are studying different manifestations or perspectives on the same discipline.

Performance Art is the focus of two case studies - Arbo Cyber, théâtre (?) and Stelarc,

though on closer examination both will also provide information on web sites, which will link with case studies in the e-government area.

Danube Exodus, and Unstable and Variable Artistic Techniques deal with installation art with digital components. Music is the focus of Obsessed Again and a significant aspect of Waking Dream. The Electronic Café began experimenting with interactive art in the 1970's, originally using analog video formats.

Case Studies - Scientific

Two of the Scientific case studies have a particular focus on geomatics - the Cybercartographic Atlas of Antarctica, and Archeological Records in a Geographical Information System : Research in the American Southwest.

This second case study focussed on the Center for Desert Archeology, a small organization which aggregates and analyzes existing archeological research related to the American southwest.

The Cybercartographic Atlas of Antarctica is an unusual case study in that it will be on-going for almost the entire duration of the InterPARES project itself. The creation of a web site containing an "atlas" is itself an experiment in the creation of an interactive environment. The case study will compare the methods available to control traditional forms of electronic records with those currently being developed for web site content management. Once completed, the cybercartographic atlas will contain material which is constantly updated, such as live video feeds from cameras installed in Antarctica. The investigators will examine the importance of "point-in-time" preservation of web sites, as well as the possibility of recording a user's interaction with dynamic web site content.

Case Studies - E-Government

As with the Artistic area, the Government On-line area is depending primarily on case studies of a variety of systems. Many, but not all, are web-based. Some, such as the VanMap case study will be comparable to the two GIS case studies in the Scientific Focus.

The e-government case studies completed to date have already identified two terminology issues. First, there is confusion between the concepts of "authenticity" and "authentication", a tendency also noted in the InterPARES 1 research. Second, the term "reliability" is primarily understood as a measure of service delivery - the amount of down-time a service suffered - rather than as a concept which applies to the content of a record.

Surveys

Surveys are actually a sub-type of General Study. They use a web-based questionnaire which asks fewer questions about recordmaking and recordkeeping practices, but elicits responses from a much larger group of respondents using digital technology. To date, we have focussed on music composers, users of Geographic Information Systems such as archeologists and geographers, and photographers.

The Composers' Survey gathered information about the use of digital technology, composers' intentions and strategies for maintaining records, and the forms that their interactive and dynamic records might take. With a response rate near 33% of the 500 composers contacted, the results show a profession well-established in what can already be termed a "traditional" digital era, with a small minority already moving forward into interactive and web-based environments. The scope of the preservation problem is illustrated in the statistic that 47% of respondents have already lost valuable files through hardware or software obsolescence. 76% of respondents use commercial, off-the-shelf products, a number that offers reassurance by limiting the scope of the needed preservation strategies.

The concept of what actually constitutes the "work" or the "oeuvre" varies widely in this community, ranging from the score to the performance, with a significant segment in this survey insisting that the "work" does not exist. Clearly, for this group, preservation will mean different things to different people. This mirrors the experiences of museum curators who are participating in the Variable Media Initiative, another external research project with which the research team concerned with the Artistic area has established contact. When dealing with installation art with digital components, they have found a wide divergence among artists in their choices of what, if anything, needs to be preserved to correctly represent their artistic intent.

These survey findings reflect to a large extent the findings of a number of other Artistic case studies, with one exception - there is no mention of hardware dependency problems, though these have been a significant aspect of the study of digital composition and performance in "Obsessed Again", in the performance piece "Waking Dream" where it was affecting a visual component of the work. One of the first suggestions for a preservation strategy in this field is to attempt to end this hardware dependency problem by moving the functionality to the software, which will reduce the scope of the problem.

In the short term, the success of the Composers' Survey has prompted the development of a targeting digital photographers, which has been available on-line since September 13.

A third survey, of GIS software users, reflected many aspects of composers' survey and the case studies. Many of the archeologists who responded to the survey work alone or in very small groups. With little need to coordinate with others, recordkeeping tends to be individualistic and undocumented; security tends to be limited to passwords; and occasional back-up to CD-R or some other digital carrier is the primary mechanism for protecting the integrity of the data. One interesting observation is that the organization and documentation of their electronic records currently occurs after-the-fact, rather than during the records' creation process.

Preliminary Findings

Though a great deal of work remains to be done, some preliminary findings are already suggesting themselves, especially in the Artistic area which has completed the largest number of case studies.

In the Artistic area, we are seeing large differences - in attitude, in procedures and in concerns - between individuals, or small groups of artists and the business/entertainment environment. For the most part, the early adopters of any specific digital technology are not part of a mainstream industry; they are individuals working on the cutting edge and exploring new forms of expression. They do not use an archivist, they don't maintain recordkeeping systems, and they tend not to write down their procedures because they are very small operations. This exact situation was also evident at the Center for Desert Archeology, and their use of geographic information system software.

In the corporate environment, such as the Computer-Based Animation company, there are significant financial interests to protect. They adopt digital technology once it is reasonably well-established and a favourable cost-benefit analysis can be done. The digital technology continues to co-exist with traditional records management practices which were already in place, such as printing to paper, or generating analog audio or video recordings.

However, these traditional record forms are incapable of capturing truly interactive or experiential aspects of digital objects, meaning some digital solutions do have to be found. Conversion to more stable analog forms is inadequate.

Another area of striking similarity between the small artistic groups and the small scientific groups is that both have introduced a rudimentary classification system to help organize their records. Both group records on a project basis.

I have already mentioned the continuing debate between the "work" and the performance in the Artistic area. We are uncovering conflicting information about the need to preserve the "means of production" vs. the record of performance. The first approach allows artists to re-use the work, and potentially to continually change the work, with or without the preservation of earlier versions. The corporate case studies show however that, where there are financial considerations, well-established and trusted traditional record-making practices are used.

The digital formats also seem to be encouraging collaborations, leading to the same files being stored on a number of different personal computers. Redundant storage practices can improve the chance of long-term survival, but create potential proof of ownership problems.

Both the survey and the case studies are showing little interest, on the part of artists, in "authenticity" yet we are getting a fairly consistent set of responses to our questions. First, they consider that the artist is the arbiter of the authenticity of their work during their life time. Second, the artist is more concerned with the preservation of their "intent" rather than with the specific way in which they chose to manifest it as any particular point in time. This suggests the need for additional metadata to more fully capture this "intent" - their definition of what is important to preserve in each work - a finding which suggests a growing link between the case studies and the Description research in particular, in the second half of this project.

In the Scientific area, the "accuracy" of data, which is of little interest to artists, is of paramount importance. Geographers often have the benefit of a concept called "ground truth", the ability to re-verify the data against the geographic location it describes. "Authenticity" was primarily associated with provenance. The degree of trust in a geographic data set is tightly linked to knowledge of who collected the data. As with the recording of artistic intent, the need to document the lineage or provenance of data sets requires "identify" metadata quite similar to that which was identified in InterPARES 1.

There are still 2 and a half years remaining in the InterPARES 2 project. During that time, a great deal of comparative analysis will be done among the case studies of the 3 subject areas of research. This will lead to the identification of additional similarities and differences among the various types of records being studied, as well as the findings of the InterPARES 1 project.

Web Site Address

All the reports and appendices I have described in this presentation are available at, or accessible through, the InterPARES web site. InterPARES 2 products are being posted there as they become available.

Yvette Hackett works at Library and Archives Canada (LAC), an institution which was created in May 21, 2004 by the merger of the National Library and the National Archives of Canada. Prior to that, she worked as an archivist for almost 20 years, beginning with moving image and sound records. She has since worked with textual and non-textual digital records in both the federal government and private-sector areas of the archives.

In recent years, her work has increasingly focussed on international research into the acquisition and preservation of digital records. Yvette was a member of the Authenticity Task Force of the first InterPARES project. In InterPARES 2, she serves as the chair of the Focus Group on Artistic Activities. In 2004, she began to represent LAC on the International Internet Preservation Consortium, which focuses on the long-term preservation of web sites.