The long-term preservation of accurate and authentic digital data: the InterPARES project

Luciana Duranti

School of Library, Archival and Information Studies
The University of British Columbia
Vancouver, British Columbia, Canada
luciana@interchange.ubc.ca

Abstract

This paper discussess the goal, objectives, structure and methodology of InterPARES 2, the second phase of an international multidisciplinary research project on the permanent preservation of the authenticity of electronic records, and presents the research conducted to date.

1. The Problem

Ongoing technological change is causing widespread concern around the world regarding the preservation of the material produced or stored using digital technologies. A portion of our society's recorded memory created and preserved digitally has already been compromised, and there are enormous costs associated with recovering electronic entities that have become inaccessible. While the extent to which valuable digital material has been lost or has become retrievable only at great expense has yet to be adequately quantified, it is already apparent that the threat is real and widespread. Moreover, even if we could ensure the preservation of electronic entities and overcome media fragility and technological obsolescence, preserved materials would be of little value unless we can be sure they are accurate, that is, precise and free of error or distortions, and authentic, that is, that their identity and their interity have not been inadvertently or maliciously compromised, and they are what they purport to be, immune from corruption and tampering. For centuries, our presumption of accuracy and authenticity has been premised on the presence or absence of visible formal elements and on an uninterrupted line of legitimate custody. The use of digital technology has not only reconfigured those formal elements, allowed for the bypassing of production controls, and made of physical custody an elusive concept, but, first and foremost, it has eliminated the original, that is the first complete instantiation of recorded data being communicated either across space (to persons other than the author) or time (saved for later access by the author or legitimate successors).

If digital materials will ever be considered accurate and authentic as those on traditional media, the practices by which they are created, maintained, made accessible and used must be analyzed, and strategies and standards for their preservation must be developed. This is the mission of InterPARES (International research on Permanent Authentic Records in Electronic Systems), a research endeavour that aims to develop the theoretical and methodological knowledge essential to the permanent preservation of authentic materials generated and/or maintained electronically, and, on the basis of this knowledge, to formulate model policies, strategies and standards capable of ensuring that preservation. At the end of its first phase, that ran from 1999 to 2001, it issued, in addition to methods of selection and preservation, a series of authenticity requirements for materials that, although digital, were very similar to their analog counterparts, especially in that they had a fixed form.[1] Increasingly, however, organizations and individuals have been generating materials of a dynamic, experiential, or interactive nature, which will require different, and perhaps data-type specific, authenticity requirements and selection and preservation strategies.

Dynamic materials depend for their content upon data estracted from databases which may have variable instantiations. The challenge they present to those who generate and access them is their lack of fixity, but more serious issues are raised by experiential and interactive objects. Clifford Lynch describes experiential digital objects as objects whose essence goes beyond the bits constituting them to incorporate the behaviour of the rendering system, or at least the interaction between the object and the rendering system. He also maintains that defining the authenticity of such objects is a much more complex problem than with raw data or traditional works, because it is dependent not on the ability to reproduce a copy of the object's original bit-stream, but on the ability to recreate the environment in which that object was experienced, an activity that involves issues of intellectual property, copyright, etc.[2]

An interactive system is one in which each user entry causes a response from or an action by the system. To generate preservable data in such systems, we need to ascertain a) how user input affects the creation and form of digital data; and b) if and when the interactive system and its inherent functionality need to be preserved for those data to remain meaningful and authentic.

Whether dynamic, experiential, and interactive digital objects are indeed to be preserved over the long-term depends of course on their relationship to the activity of their creator and on the value that society attributes to them. Scientific rsearchers have a long history of creating such objects, and clearly the professionals charged with the preservation of the archives containing them may have to face the concrete challenge of preserving views of dynamic systems, maintaining the functionality of interactive data, and recreating the environment of experiential objects. It is important both to know to what extent the requirements, methods and strategies developed by the InterPARES 1 project to preserve authentic electronic material with a fixed form apply to these new situations, and to develop new ones where they do not. These issues are further compounded when individual creators lack the knowledge and tools to generate data that can be preserved over the long term.

For these reasons, it is necessary to develop an understanding of the new digital objects, not only in the later phases of their life cycle, but from the moment of their creation. In fact, it is probably necessary to revisit the concept of recorded data itself, so that both the identification and the protection of these new types will be possible. We have to consider the possibility of substituting the characteristics of stability and fixity with the capacity of the system where the data reside to trace and preserve each change each digital object has undergone. And perhaps we may look at each digital entity as existing in one of two modes, as an entity in becoming, when its process of creation is in course (even if such creation is ongoing), and as a fixed entity at any given time the data

[1] The requirements developed by InterPARES 1 can be found on the project's website at http://www.interpares.org/book/interpares_book_k_app02.pdf

[2] Lynch, Clifford. "Experiential Documents and the Technologies of Remembrance." *I in the Sky: Visions of the Information Future,* edited by Alison Scammell. London: Library Association Publishing, 2000.

is viewed. There is no doubt that knowledge and strategies must be developed that are beneficial for both the creators and preservers of these complex new materials.

Technological obsolescence, which poses a continual challenge to the accessibility, readability and intelligibility of electronic objects, is of even more concern in the context of scientific activities than in that of administrative activities. Inadequate record-management practices have already precipitated the disappearance of many data sets that depended upon now obsolete software and hardware for their continued existence, including the research material of the US Navy. This has generated enormous difficulties for scientists concerned with the long-term preservation of the unique and authoritative version of their work, requiring them to devote valuable time and resources to preservation efforts and engendering an urgent demand for effective and tested strategies.

To meet these challenges requires an understanding of the nature of the new electronic objects and their creating processes. Research must be done into their characteristics and development, the requirements for their reliability, accuracy, and verifiable authenticity, and methods and strategies for their selection and preservation. To this end, the international team of researchers formed for InterPARES 1, together with additional researchers with discipline-specific knowledge, decided to initiate a second phase of its research, called InterPARES 2.

2. InterPARES 2: Intellectual Framework

InterPARES 2 began in 2002 and its completion is scheduled for the end of 2006. It goal, objectives, structure and methodological principles have been articulated in an intellectual framework on which all co-investigators agreed.

2. 1 *Research goal*

The goal of InterPARES 2 is to ensure that the portion of society’s recorded memory that is digitally produced in interactive, dynamic and experiential systems can be created in accurate and reliable form, and maintained and preserved in authentic form, both in the short and the long term, for the use of those who created it and of society at large, regardless of digital technology obsolescence and media fragility.

2.2 *Research objectives*

– To develop an understanding of interactive, dynamic and experiential systems and of the materials produced and maintained in them, of their process of creation, and of their present and potential use;
– to formulate methods for ensuring that these digital objects are generated and maintained by the creator in such a way that they can be trusted as to their content (that is, are accurate and reliable) and as records (that is, are authentic);
– to formulate methods for selecting among them those that have to be kept after they are no longer needed by the creator because of their larger value to research;

– to develop methods and strategies for keeping the materials selected for continuing preservation in authentic form over the long term;
– to develop processes for analyzing and criteria for evaluating advanced technologies for the implementation of the methods listed above in ways that respect cultural diversity and pluralism; and
– to identify and/or develop specifications for policy, metadata, and automated tools necessary for the creation of an electronic infrastructure capable of supporting the creation of accurate and reliable, and the preservation of authentic digital objects.

2.3 *Guiding methodological principles*

2.3.1. Interdisciplinarity

The project is interdisciplinary in the measure in which its goal and objectives can only be achieved through the contribution of several disciplines and of all categories of stakeholders: individual creators of digital objects, the information technology sector, the archival and conservation professions, etc. are involved in the formulation and selection of case studies, gathering of empirical evidence, and analysis. Such a mode of research ensures that the project's results will find ready acceptance within the targeted communities. The scholars conducting the research come from the following fields: Archival Science, Archaeology, Astronomy, Chemistry, Computer Engineering, Computer Science, Dance, Diplomatics, Film, Geography, History, Information studies, Law, Library Science, Linguistics, Mechanic Engineering, Media Studies, Music, Performance Art, Physics, Photography, Space Sciences, and Theatre. The countries actively involved in the project are: Canada, United States, Australia, Belgium, China, France, Ireland, Italy, Japan, Netherlands, Portugal, Singapore, Spain and the United Kingdom. The Advisory Board also includes an archivist from South Africa.

2.3.2 Transferability

The ultimate goal of the project is archival in nature, in that it is concerned with the development of a trusted system for making and keeping digital entities[3] and of a preservation system that ensures the authenticity of the entities under examination over the long term. This implies that the work carried out throughout the project in the various disciplinary areas must be constantly translated in archival terms and linked to archival concepts, which are the foundation upon which the systems intended to protect the digital entities are designed. However, upon completion of the research, the archival systems need to be made accessible and comprehensible to records creators, organizations and institutions and disciplinary researchers. In other words, the research outcomes must be translated back into the language and concepts of each discipline that need to make use of them. In light of the above, all researchers are committed to learning the key archival concepts that are identified by the archival scholars in the team as constituting the core of the InterPARES 2 research, so that each discipline can identify the corresponding entities within its own body of knowledge.

[3] A trusted system comprises the whole of the rules that control the creation, maintenance, and use of the materials of the creator and that provide a circumstantial probability of the accuracy, reliability and authenticity of the digital objects within the system.

2.3.3. Open inquiry

InterPARES 2 espouses no epistemological perspective or intellectual definitions *a priori*. Instead, researchers in each working group identify the perspective(s), research design, and methods that they believe to be most appropriate to their inquiry. The reason for this openness is that InterPARES 2 is conceived to work as a "layered knowledge" environment, in the sense that some of the research work will build upon knowledge developed in the course of InterPARES 1, some will take knowledge of similar issues developed in other areas of endeavour and bring it to bear on creation and preservation of digital materials, some will reconcile knowledge about records and their attributes, elements, characteristics, behaviour and qualities existing in various disciplines and develop it for archival purposes, and some will explore new issues and study entities never examined before and develop entirely new knowledge.

2.3.4. Multi-method design

As stated, each research activity is carried out using the methodology and the tools that the dedicated investigating team considers the most appropriate for it. Examples of the methods used are surveys, case studies, modeling, prototyping, diplomatic and archival analysis, and text analysis.

The research is guided by detailed research questions that specifically address the records creation process in each of the examined areas of endeavour, and the characteristics, structure and interrelationships of the resulting materials; the issues related to the development of a chain of preservation for those materials that begins with creation and includes appraisal, description, and reproduction as authenticating procedures; the meaning of the concepts of accuracy, reliability and authenticity in the various disciplines; the policies, strategies and standards in each area of activity covered by the research; the descriptive schemas necessary to the identification, use and preservation of the materials produced by each activity throughout their life-cycle; and the models that more appropriately represent the digital object that is investigated and the processes of its creation, maintenance, use, selection and preservation.

3. Research Progress

The need to concentrate the initial part of the research on gathering an understanding of the process of creation in interactive, dynamic and experiential digital environments has been especially encouraged and supported by the participant stakeholders. The researchers have carried out case studies and general studies. The case studies were identified according to the specific kind of activity that generated the material, and conducted by individual teams assembled in an interdisciplinary way for the purpose of investigating the entire life cycle of the digital objects that were examined. Each team comprised at least a scholar of the activity under investigation, a technology expert, an archival scientist, and a student research assistant. Depending on the complexity of the case study, additional experts and students might belong in the team. The general studies were developed to address issues relevant to each of the three types of activities producing records, but not specific to any given case. Examples of the case studies undertaken in the scientific focus are:

-- Archaeological Records in a Geographical Information System: Research in the American Southwest. This case study focuses on records from geographic information systems that were created by the Centre of Desert Archaeology in Tucson, Arizona. Specifically, how these records were created, what their disposition is, what happens after disposition, and their corresponding authenticity, reliability and accuracy while undergoing these processes.

-- Preservation and Authentication of Electronic Engineering and Manufacturing Records. The records which are examined in this case study have been created in computer-assisted engineering, computer-assisted design and industrial automation systems. The focus of this case study is on examining the ability of complex engineering records to stand for the solid objects modeled in the records, and the ability of the manufacturing records to represent the processes required to produce such solid objects.

Examples of general studies are:

-- Persistent Archives Based on Data Grids. This study focuses on the San Diego Supercomputer Centre's project to develop a prototype for a persistent archives based upon data grid technology for the National Archives and Records Administration (NARA). This study examines the minimal capabilities needed within grid technology for preservation of governmental records, focusing on activities related to the preservation of NARA's selected digital holdings.

-- A survey of the e-science literature for file formats and encoding languages that are used for non-textual scientific data, information and records. File formats and encoding languages are also analyzed to determine data, information and/or record structure and other properties related to the concepts of accuracy, reliability and authenticity of the digital objects in question. In addition, the study will determine equivalence classes of file formats and encoding languages and identify conversion tools that can be used for migration.

Several other outcomes have been produced by various project's research units which are concerned with large issues regarding all the disciplinary and professional areas covered by the research, such as metadata or policy. The products of these research activities will be soon posted on the InterPARES web site at www.interpares.org.

4. Conclusion

The InterPARES 2 Project has already produced a large quantity of the material on the basis of which it will develop the project's deliverables, that is, among other things, guidelines for records creators outlining methods for the reliable production and maintenance of data that can be authentically preserved; prototypes of appraisal and preservation systems, and guidelines for records preservers; frameworks for developing policies, strategies and standards, and for the development of descriptive standards for the digital objects under examination; registries of metadata schemas; and literature and terminology databases. However, as Project Director, I recognize that the most desirable outcome of this project has already been achieved: the harmonious collaboration of scholars and professionals from such a large variety of disciplines, backgrounds and cultures towards the long-term preservation of their digital memory is the invaluable product of InterPARES that I watch in awe and cherish every day as the work progresses.