Conference Proceedings

Edited by:
Luciana Duranti and
Elizabeth Shaffer

# The Memory of the World in the Digital Age: Digitization and Preservation

An international conference on permanent access to digital documentary heritage

United Nations
Educational, Scientific and
Cultural Organization

Memory of the World
20th Anniversary

Hosted by:

a place of mind
THE UNIVERSITY OF BRITISH COLUMBIA

In collaboration with

UNIVERSITY OF TORONTO

26 to 28 SEPTEMBER 2012
Vancouver, British Columbia, Canada
Sheraton Vancouver Wall Centre

United Nations
Educational, Scientific and
Cultural Organization

Memory of the World
20th Anniversary

Organisation
des Nations Unies
pour l'éducation,
la science et la culture

Mémoire du monde
20e anniversaire

Conference Proceedings

Edited by:
Luciana Duranti and
Elizabeth Shaffer

# The Memory of the World in the Digital Age: Digitization and Preservation

An international conference on permanent access to digital documentary heritage

26 to 28 SEPTEMBER 2012

Vancouver, British Columbia, Canada

Sheraton Vancouver Wall Centre

**UNESCO Memory of the World Programme, Knowledge Societies Division**

This book of Proceedings includes most of the papers and posters presented at the International Conference "The Memory of the World in the Digital Age: Digitization and Preservation" held on 26–28 September 2012 in Vancouver, British Columbia, Canada, by the UNESCO Memory of the World Programme, Knowledge Societies Division, and The University of British Columbia in collaboration with the University of Toronto.

The proceedings have been compiled and formatted with minor editing; papers and posters appear as submitted. The authors are responsible for the choice and the presentation of the facts contained in this publication and for the opinions they express, which are not necessarily those of UNESCO and do not commit the Organization.

The designations employed and the presentation of material throughout this publication do not imply the expression of any opinion whatsoever on the part of UNESCO concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries.

The organizers of this UNESCO Memory of the World Programme Conference would like to sincerely thank everyone who contributed to the Conference in Vancouver and to these proceedings.

Published by UNESCO 2013, with the financial support of the Social Sciences and Humanities Research Council of Canada | Conseil de recherches en sciences humaines du Canada (SSHRC) and the International Research on Permanent Authentic Records in Electronic Systems (InterPARES) Project.

Social Sciences and Humanities Research Council of Canada | Conseil de recherches en sciences humaines du Canada

Canada

**InterPARES Project**

**International Research on Permanent Authentic Records in Electronic Systems**

# Preface

This publication presents the proceedings of the international conference 'Memory of the World in the Digital Age: Digitization and Preservation' which was held in Vancouver, Canada, from 26 to 28 September 2012.

More than 500 experts and other interested persons from all regions of the world participated in this knowledge-sharing and policy-driving event to discuss and exchange opinions on how to protect the world's documentary heritage. Although this heritage is the record of knowledge, its physical carriers are extremely vulnerable and can easily disappear without a trace. Whether recorded on a clay tablet or an electronic tablet, our methods of sharing content and knowledge need to be protected.

It is impossible to exaggerate the importance of documentary heritage in our lives. It governs our actions whether these relate to creating the basis of mutual respect between different civilizations and communities or building knowledge societies. Documentary heritage provides the foundation of peace, our identity and knowledge.

UNESCO's interest in this subject matter is as fundamental as its constitution with its mandate to contribute to building peace through the spread of knowledge from improved access to printed and published materials. These core materials, our documentary heritage, have been preserved in archives, libraries and museums for generations.

But while measures needed to maintain access to print materials are globally understood, the newer challenges related to preserving digital information are not keeping pace with technological development. The need for dedicated hardware and software, associated with their rapid obsolescence, hamper our ability to keep invaluable content accessible. Unless timely migration to newer technologies, operating systems and software platforms is assured, we face the risk developing digital Alzheimer's.

UNESCO's expectation from this Conference was to obtain a better definition of our expected role, and our contribution to setting a global digital agenda. The UNESCO/UBC Vancouver Declaration sets out specific recommendations which we will be implementing and incorporating into our digital strategy. Likewise, we expect that our Member States, professional organizations and private sector bodies will also implement the recommendations addressed to them.

Only through collaborative strategic alliances can we overcome the major challenges threatening the preservation of digital information. We believe that the presentations featured in this publication provide the basis for a global commitment to preserving the memory of our world in this digital age.

**Jānis Kārkliņš**
**Assistant Director-General**
**for Communication and Information**

# Data, Documents, and Memory

*A Taxonomy of Sources in Relation to Digital Preservation and Authenticity Metadata*

**Joseph T. Tennis**
*University of Washington Information School*

**Abstract**

*This paper describes efforts to operationalize methods of metadata application to digital records. We summarize theoretical and applied research from the field and from the InterPARES research project to establish a grounding in the issues and suggestions for metadata assignment. Key to our understanding of metadata is 1) its relation to documentation (a more narrative form of description of records), and 2) its permanence in relation to the records themselves. This paper closes by presenting a draft taxonomy of sources useful to decision-making in relation to metadata assignment and documentation creation.*

**Author**

Joseph T. Tennis is an Assistant Professor at the Information School of the University of Washington, a member of the Textual Studies faculty at UW, and an Associate Member of the Peter Wall Institute for Advanced Study at The University of British Columbia. He has been active in the InterPARES research project since 2005, and currently serves as an advisor and researcher on metadata issues. He holds a B.A. in Religious Studies, an M.L.S., an Sp.L.I.S. in Book History, and the Ph.D. in Information Science from the University of Washington. He works in classification theory, the versioning of classification schemes and thesauri (a.k.a. subject ontogeny), and the comparative discursive analysis of metadata creation and evaluation, including archival metadata, both contemporary and historical.

## 1. Introduction

Current archival theory has pointed to the role of both documentation and metadata as key to attesting to the authenticity of body of records. Upon closer inspection, we see that even if this is what theory prescribes, the complexity and variety of this kind of documentation and metadata seems *confusingly rich*. On the one hand we understand metadata to be machine and human readable assertions about resources, in general. In the archival context we constrain that meaning by saying that there are two kinds of metadata, and that those metadata are focused specifically on records and aggregations of records. The two types are intrinsic and extrinsic metadata—those that are permanently linked to the record and those that are not (Gilliland, 2008). Following the work carried out in the InterPARES research project, we find that the complexity in this case lies in asserting what metadata are required to persist along with the record over time (InterPARES). If it is intrinsic to the record, we might assume that it should persist. Those that are extrinsic pose a different methodological problem. How do we assess what stays and what is deleted? And though we assume intrinsic metadata should persist with the record, they are not without their complications. Advancing archival theory problematizes our ability to wed metadata inseparably with digital records, claiming that it is itself a study in addition and deletion, requiring particular methodological commitments. With both intrinsic and extrinsic metadata we must say what stays and what goes.

Documentation is another complex theoretical concern. The complexity lies with the relationship between documentation and digital recordkeeping in particular, and recordkeeping in general. Evolving theory related to authenticity and the conception of the *fonds* finds that the archivist must document

*Proceedings of The Memory of the World in the Digital Age: Digitization and Preservation. An international conference on permanent access to digital documentary heritage, 26-28 September 2012, Vancouver, British Columbia, Canada,* edited by Luciana Duranti and Elizabeth Shaffer (UNESCO 2013).

interventions made by the creator and preserver (e.g., MacNeil 2008 & Millar, 2002). This means that in constructing a coherent picture of a fonds or describing the provenance of a body of records, theory guides the archivist to document his or her decisions made. This documentation must also follow the record through various stages of preservation—persisting along with the records.

The question surfaces: how do we make operational the theory of metadata and documentation in digital records preservation systems, given this level of complexity? The first step that could be taken is to create a taxonomy of these *sources* of information so that we might use that categorization to manage that type over time. This paper takes the first step in that direction by proposing such a draft taxonomy. This taxonomy accounts for the characteristics of metadata and documentation mentioned above, and adds to it from InterPARES research and contemporary theory found in the literature. The result is a rubric that can used to make decisions about how to design out systems that can keep authentic digital records.

## 2. Metadata vs. Documentation

In its most common definition, metadata is data about data. However, this definition is not adequate to distinguish metadata from documentation. And this is a distinction, in the context of archival metadata theory and practice that we want to establish and maintain. For our purposes, *general* metadata is human- and machine-readable assertions about a resource, where resource is the World Wide Web Consortium's (w3c) term. Resource to the w3c is anything with an identity. We have scoped resource in our context to be records, and our assertions are the various things that can be said about records for the purpose of authenticity, preservation, and retrieval. We have scoped metadata thusly based on InterPARES research (InterPARES). Specifically, we have drawn on the InterPARES Benchmark Requirements Supporting the Presumption of Authenticity of Electronic Records (InterPARES 1: Authenticity Task Force, 2005) the Baseline Requirements Supporting the Production of Authentic Copies of Electronic Records (InterPARES 1: Authenticity Task Force, 2005), and the Chain of Preservation (COP) model (Preston, 2009; Duranti and Preston, 2008).

Examples of metadata drawn from these sources are the names of persons concurring in the formation of the record. InterPARES has identified five persons that can be identified with the generation of a digital record: author, addressee, writer, originator, and creator. In each of these cases we can fill in the blank:

"The addressee of the document is *x*"

Documentation has been a concept closely associated with all stages of the lifecycle of records, but has become an even richer concept in the digital environment and in relation to contemporary discussions of metadata. In discussions of archival appraisal and description we have evolving theory of how documentation is required for the presumption of authenticity and to reveal the details of the agents and actions that helped create any given fonds.

In the context of archival description, we have accounts from MacNeil about the differences between metadata and archival description (1995). In this paper MacNeil distinguishes archival description from metadata by claiming the latter is the view from the ground, while archival description is like a view from an airplane. What description provides is an overview of the whole body, history, and scope of the body of records. Metadata, on the other hand, serve as the raw material for archival description. They are what the archivist uses to construct their bird's-eye-view.

Given this perspective on metadata, we can see that much of it should be discarded at the point of archival description. That is, once the body of records has crossed the threshold of preservation, bringing with it all attached metadata, the archivist paints a picture of the body of records from all available evidence and then discards the evidence, including much of the metadata. For example, before records move to the preserver, for recordkeeping purposes we might trace which records had been destroyed. In our Chain of Preservation Model (Preston, 2009), this is activity A3.4.3. However, once the body of records crosses the threshold of preservation it may not be the decision of the creator to keep this information, and in this case, it would not be part of the preserver's work to keep track of it. Of course, one can imagine the opposite, but we can take this as our example. In this case, the creator of the records only wants a statement of what is in the fonds, not what has been destroyed from the fonds. The metadata associated with this decision to destroy records is no longer kept after it is entrusted to the preserver.

The fonds, as a unit of analysis in archival work, has undergone some close inspection in the literature. Both MacNeil (2008) and Millar (2002) have reexamined our assumptions about the way we talk about, and hence document, the creator and their body of records. MacNeil's concern with authenticity, arrangement, and archival description drew her to research intentions of both authors of scholarly texts and archivists. Seeing through this comparison that:

> the theory of final intentions is underpinned by a particular ideology concerning the nature of artistic creation, i.e., the author as solitary genius. The principle of original order, for its part, is underpinned by particular ideologies concerning the nature of historical inquiry . Lehmann's articulation of the Prussian principle of original order in the latter part of the nineteenth century, for example, resonated with the ideology of "scientific" history." (MacNeil, 2008, p. 13)

The upshot of this work is that we must document our own actions as archivists in representing creator's intentions. Thus, it is not by metadata alone that we can best represent the fonds, its history, and its accumulation into its current state. MacNeil calls the rearrangement of records by different custodians the records' archivalterity.

Millar in a not dissimilar vein, separates *respect des fonds* and provenance as two distinct concepts with which archivists must reckon. Her case study is the Hudson Bay Company's records. They are spread out over different archival institutions, and they serve as a lesson in provenance. This concept, provenance, would in Millar's formulation, encompass three distinct histories: 1) creator history, 2) records history (or recordkeeping history), and 3) custodial history. These would constitute what she sees as the only useful guiding principle, especially when compared to the unrealizable concept of the fonds. She wants archivists to work with a new concept she calls *respect de provenance*, and they would do that by writing out the three different histories.

Both Millar's histories and MacNeil's creator's intentions, and subsequent change of records arrangement by the chain of custodians, require something more than metadata can offer. They require documentation. We have found the same need for documentation in the InterPARES research project. In the process of drafting a metadata application profile that is consistent with diplomatic assumptions about records, in accordance with the findings of the Benchmark and Baseline Requirements (InterPARES 1: Authenticity Task Force, 2005), established by InterPARES 1, and based on the Chain of Preservation model (COP model) (Preston, 2009; Duranti and Preston, 2008), we found that metadata alone could not maintain presumption of authenticity in digital records systems through time.

## 3. The Chain of Preservation

The lifecycle of a body of records has been represented in ideal form in the Chain of Preservation model (COP model) (Preston, 2009; Duranti and Preston, 2008). Through this model we have begun to enumerate the metadata required for the presumption of authenticity (Tennis and Rogers, 2012). We call our metadata the IPAM, which stands for InterPARES Authenticity Metadata. There are a total of 428 assertions made about records and their context in the IPAM. We have categorized them into 12 categories. They are given below.

- AT – attachments: Signals those items attached to the record—indication of attachments is necessary for the integrity of the record.
- AU – authentication: Those elements that indicate the identity of the persons involved in the creation of the record.
- B – archival bond: Those elements that illuminate the connection of the record to other records to which it relates, and its context, whether it is preserved or destroyed.
- D – date: Points in time in the life cycle of the record(s) that need to be documented.
- DO – external documentation: Links to information that governs preservation, transfer, and access to the record(s) over time.
- F – form: The rules of representation that determine the appearance of an entity and convey its meaning.
- H – handling: Representation of the office or officer formally competent and/or responsible for carrying out the action to which the record(s) relates or for the matter to which the record(s) pertains.
- L – location: Indications of where the record(s) are stored, backed up, duplicated.
- P – persons: Identification of individuals or legally defined entities who are the subject of rights and duties and are recognized by the juridical system as capable of or having the potential for acting legally with regard to the record(s)
- R – rights and access: restrictions or privileges that apply to the record(s).
- S – subject: The action or matter to which the record(s) pertain.
- T – technology: The carrier(s) of the form and content of the record.

Of these, documentation (DO) rivals persons (P) as the most frequent assertion made. We have established at least 46 links to external documentation as required for the presumption of authenticity of digital records. For example in the context of records creation we need to indicate which records were transferred (DO0), whether the records were modified (DO1), and whether the records were backed up (DO3). To assert digital records transfer, modification, or backup, we need links to external documentation. The other documentation deals with corrections to records, updates to records, access to records, etc. They deal with the integrity of the records, the systems in which they are kept, and serve as an attestation of what kind of interventions effect the form and content of the records as they move from creation to preservation.

## 4. Taxonomy of Sources

If we build directly out of Millar, MacNeil, and InterPARES we can see a categorization of metadata and documentation surface. There are two categories of metadata and three categories of documentation. The two categories of metadata are Identity Metadata and Integrity Metadata. Identity metadata comprise:

Table 1. Table of Identity Metadata.

| | |
|---|---|
| D00 | the date of *document* creation |
| D01 | chronological date (and possibly time) of compilation and capture; |
| F01 | documentary form—that is, whether the document is a report, a letter, a contract, etc.; and |
| T01 | digital presentation—that is, file format, wrapper, encoding, etc. |
| D02 | chronological date (and possibly time) of transmission from the originator; |
| D03 | chronological date (and possibly time) of receipt and capture; |
| F01 | documentary form—that is, whether the document is a report, a letter, a contract, etc.; and |
| T01 | digital presentation—that is, file format, wrapper, encoding, etc. |
| | |
| P02 | - *author(s)*—that is, the physical or juridical person(s) responsible for issuing the document; |
| AU04 | - *subscription*—that is, the name of the author or writer appearing at the bottom of the document; and |
| AU05 | - *qualification of signature*—that is, the mention of the title, capacity and/or address of the person or persons signing the document; |
| AT01 | - indication of any attachments—that is, mention of autonomous digital objects linked inextricably to the document. |
| P03 | - *writer(s)*—that is, the physical person(s) or position(s) responsible for articulating the content of the document; |
| P04 | - *addressee(s)*—that is, the physical or juridical person(s) for whom the document is intended; |
| P05 | - the physical person(s), position(s) or office(s) responsible for the electronic account or technical environment where the document is generated and/or from which the document is transmitted;[1] |
| P06 | - *receiver(s) or recipient(s)*—that is, the physical or juridical person(s) to whom the document may be copied or blind copied for information purposes; |
| S01 | - name of the action or matter—that is, the subject line(s) and/or the title at the top of the document; |
| AU01 | - indication of the presence of a digital signature; |
| AU02 | - *corroboration*—that is, an explicit mention of the means used to validate the document; |
| AU03 | - *attestation*—that is, the validation of the document by those who took part in the issuing of it, and by witnesses to the action or to the 'signing' of the document; |
| | |
| B01 | classification code; and |
| B04 | planned disposition (if not evident in the classification code). |
| B02 | registration number. |
| P01 | The physical or juridical person who makes, receives or accumulates records by reason of its mandate/mission, functions or activities and who generates the highest-level aggregation in which the records belong (that is, the fonds). Syn.: creator. |
| R01 | indication of copyright or other intellectual rights; |
| H01 | name of handling office (if not evident in the classification code); |
| H02 | name of office of primary responsibility (if not evident in the classification code and records retention schedule); |
| R02 | access restriction code (if not evident in the classification code); |
| R03 | access privileges code (if not evident in the classification code); |

| | |
|---|---|
| B03 | vital record code (if not evident in the classification code); and |
| | |
| AN01 | priority of transmission; (urgent, etc.) |
| D04 | transmission date, time and/or place; |
| SS01 | actions taken; |
| D05 | dates and times of further action or transmission; and |
| AT02 | information on any attachments—that is, mention of autonomous items that were linked inextricably to the document prior to its transmission for the document to accomplish its purpose. |
| | |
| F02 | draft or version number; |
| D06 | archival or filing date—that is, the date on which a record is officially incorporated into the creator's records; |
| AT03 | indication of any annotations[5] or new attachments (e.g., records profiles); |
| R02 | access restriction code (if applicable and if not evident in the classification code)—that is, indication of the person, position or office authorized to read the record; |
| R03 | access privileges code (if applicable and if not evident in the classification code)—that is, indication of the person, position or office authorized to annotate the record, delete it, or remove it from the system; |
| B03 | vital record code (if applicable and if not evident in the classification code)—that is, indication of the degree of importance of the record to continue the activity for which it was created or the business of the person/office that created it;[6] and |
| B04 | planned disposition (if not evident in the classification code)—for example, removal from the live system to storage outside the system, transfer to the care of a trusted custodian, or scheduled deletion. |
| B01 | expression of archival bond (e.g., via classification code, file identifier, record item identifier, dossier identifier, etc.); |
| P01 | name of the creator—that is, the name of the physical or juridical person in whose archival fonds the record exists; |
| R01 | indication of copyright or other intellectual rights (if applicable);[2] |
| B05 | indication, as applicable, of the existence and location of duplicate records, whether inside or outside the record-making or recordkeeping systems and, in instances where duplicate records exist, which is the authoritative copy—that is, the instantiation of a record that is considered by the creator to be its official record and is usually subject to procedural controls that are not required for other instantiations;[3] |
| H01 | name of the handling office (if not evident in the classification code)—that is, the person or office using the record to carry out business; |
| H02 | name of the office of primary responsibility (if not evident in the classification code or the records retention schedule)—that is, the office given the formal competence for maintaining the authoritative version or copy of records belonging to a given class within a classification scheme;[4] |
| T02 | indication of any technical changes to the records—for example, change of encoding, wrapper or format, upgrading from one version to another of an application, or conversion of several linked digital components to one component only—by embedding directly in the record digital components that were previously only linked to the record, such as audio, video, graphic or text elements like fonts; |

Identity Metadata are permanent and fixed to the records of the creator. The majority of the other metadata is Integrity metadata—that is it is metadata that accounts for the handling of records in digital systems from this point of creation through to the point of permanent preservation. By definition integrity metadata can be compiled as reports in external documentation. Thus, though the system may generate metadata, this metadata is then compiled into documentation and the metadata discarded as no longer necessary. Integrity metadata are further erased as they are reported in Creation, Recordkeeping, and Preservation Documentation.

The three categories of documentation follow the stages in the COP model. Creation documentation includes the transfer of records from the context of creation to the recordkeeping system. Recordkeeping Documentation is the outcome of MacNeil's archivalterity, that is, the acts of continuous and discontinuous change that transform the meaning and authenticity of a fonds as it is transmitted over time and space (MacNeil, 2008 p.14). This kind of documentation also reflects the *custodial bond* "meaning the relations that exist between a body of records and the various custodial authorities that interact with the records over time, including archivists and archival institutions," (MacNeil, 2008 p. 14). Any transfer, modification, correction, updates, refreshing that happens to records as they are kept is also reflected, in summary form, in this kind of documentation.

The final documentation is Preservation Documentation. We hypothesize that this category of documentation is relevant from a contingent definition of preserver. The trusted third party and ultimate keeper of the body of records. Ultimate here meaning *the current keeper considered the final keeper*. Once records move (and they do move) we move this documentation into to recordkeeping documentation. Preservation documentation consists of authentication reports, preservation feasibility reports, disposition reports, state-of-records reports (documenting technological carriers and documentary form of records as they cross the threshold of preservation). Preservation documentation also includes Millar's creator history (documenting functional changes, and name the potentially diverse set persons involved in the creation of the fonds etc.), recordkeeping history (which would bring forward all the relevant integrity metadata), and custodial history (which would add to the transfer reports a narrative of context about where records were found, how and why they moved, and attempt to make clear the decisions of previous archivists).

## 5. Toward Operationalized Theory

To consider a taxonomy of sources is to take a step toward operationalizing theory. I have made a bold statement in outlining what metadata I think should be kept permanently and which can be transformed into documentation that follows the records. To tell the story of digital records is a complex task. What contemporary theory of archives tells us is that we must make clear our interventions, narrate our roles and how we see the roles and actions of others. This supplements our understanding of archival description, makes clear the role of temporary and permanent metadata, and makes robust our systems of memory.

## References

Duranti, L. and R. Preston, eds. *International Research on Permanent Authentic Records in Electronic Systems (InterPARES) 2: Experiential, Interactive, and Dynamic Records* Padova: Associazione

Nazionale Archivistica Italiana, 2008. Online reprint available at http://www.interpares.org/ip2/book.cfm.

Gilliland, A. "Setting the stage." In *Introduction to Metadata*, edited by M. Baca, 1-19. Los Angeles: Getty Publications, 2008. Online edition available at http://www.getty.edu/research/publications/electronic_publications/intrometadata/index.html.

InterPARES. www.interpares.org.

InterPARES 1: Authenticity Task Force. "Appendix 2: Requirements for Assessing and Maintaining the Authenticity of Electronic Records." In *The Long-term Preservation of Authentic Electronic Records: Findings of the InterPARES Project*, edited by Luciana Duranti, 204-219. San Miniato, Italy: Archilab, 2005. Online reprint available at http://www.interpares.org/book/interpares_book_k_app02.pdf.

MacNeil, H. "Archivalterity: Rethinking Original Order." *Achivaria* 66 (2008): 1-24.

MacNeil, H. "Metadata strategies and archival description: comparing apples and oranges." *Archivaria* 39 (1995): 22-32.

Millar, L. "The Death of the Fonds and the Resurrection of Provenance: Archival Context in Space and Time." *Archivaria* 53 (2002): 1-15.

Preston, Randy. "InterPARES 2 Chain of Preservation (COP) Model Metadata (Draft)." 2009. http://www.interpares.org/rws/display_file.cfm?doc=IP2-cop-model_metadata_v1.0.doc.

Tennis, J. T., and C. Rogers. "Authenticity Metadata and the IPAM: Progress toward the InterPARES Application Profile." In *Proceedings of the International Conference on Dublin Core and Metadata Applications 2012*, 38-45. Kuching/Sarawak, Malaysia, 2012. http://dcevents.dublincore.org/IntConf/dc-2012/paper/view/109.

UNESCO. *Universal Declaration on Cultural Diversity*. 2001. Accessed 1 February 2009. http://www.un-documents.net/udcd.htm.

UNESCO. *UNESCO World Report: Investing in Cultural Diversity and Intercultural Dialogue*. Paris: UNESCO Publishing, 2009.

Wikipedia. "Optical character recognition." Accessed 1 September 2012). http://en.wikipedia.org/wiki/Optical_character_recognition.