InterPARES 2 Project

International Research on Permanent Authentic Records in Electronic Systems

International Research on Permanent Authentic Records in Electronic Systems (InterPARES) 2: Experiential, Interactive and Dynamic Records

PART SIX

INVESTIGATING THE ROLES AND REQUIREMENTS, MANIFESTATIONS AND MANAGEMENT OF METADATA IN THE CREATION OF RELIABLE AND PRESERVATION OF AUTHENTIC DIGITAL ENTITIES

Description Cross-domain Task Force Report

[including Appendices 17 and 18]

by
Anne Gilliland, University of California, Los Angeles

Co-authors
Lori Lindberg, University of California, Los Angeles
Victoria McCargar, Los Angeles Times
Alison Langmead, University of California, Los Angeles
Tracey P. Lauriault, Carleton University
Monique Leahey-Sugimoto, University of California, Los Angeles
Joanne Evans, University of California, Los Angeles
Joe Tennis, University of Washington
Holly Wang, University of California, Los Angeles

Status: Final (public)

Version: Electronic

Submission Date: February 2007

Publication Date: 2008

Project Unit: Description Cross-domain Task Force

URL: http://www.interpares.org/display_file.cfm?doc=ip2_book_part_6_description_task_force.pdf

How to Cite: Anne Gilliland et al., "Part Six—Investigating the Roles and Requirements, Manifestations and Management of Metadata in the Creation of Reliable and Preservation of Authentic Digital Entities: Description Cross-domain Task Force Report," [electronic version] in *International Research on Permanent Authentic Records in Electronic Systems (InterPARES) 2: Experiential, Interactive and Dynamic Records*, Luciana Duranti and Randy Preston, eds. (Padova, Italy: Associazione Nazionale Archivistica Italiana, 2008). <http://www.interpares.org/display_file.cfm?doc=ip2_book_part_6_description_task_force.pdf>

Table of Contents

Introduction

Metadata that are associated with either an information system or an information object for the purposes of description, administration, legal requirements, technical functionality, use and usage and preservation, play a critical role in ensuring the creation, management, preservation, discovery, use and re-use of trustworthy materials, including records. Recordkeeping[1] metadata, of which one key type is archival description, play a particularly important role in documenting the reliability and authenticity of records and recordkeeping systems as well as the various contexts (legal-administrative, provenancial, procedural, documentary and technical) within which records are created and kept as they move across space and time. In the digital environment, metadata are also the means by which it is possible to identify how record components—those constituent aspects of a digital record that might be managed, stored and used separately by the creator or the preserver—can be reassembled to generate an authentic copy of a record or reformulated per a user's request as a customized output package. Metadata in the sciences also provide essential data quality elements such as accuracy, lineage, reliability, margins of error, limitations and precision, among others, that assist the user to assess whether the dataset in questions is fit for the intended use of the scientist.

Issues relating to the creation, capture, management and preservation of adequate metadata are, therefore, integral to any research study addressing the reliability and authenticity of digital entities created by dynamic, interactive and experiential systems, regardless of the community, sector or institution within which they are being created. The Description Cross-domain Task Force examined the conceptualization, definitions, roles and current functionality of metadata and archival description in terms of requirements generated by InterPARES 1 as well as case study data and models generated during InterPARES 2. Because of the needs to communicate the work of InterPARES in a meaningful way across not only other disciplines, but also different archival traditions; to interface with, evaluate and inform existing standards, practices and other research projects; and to ensure interoperability across the three focus areas of InterPARES 2, the Description Cross-domain also addressed its research goals with reference to wider thinking about and developments in recordkeeping and metadata.

InterPARES 2 addressed not only records but also a range of digital information objects (often referred to as "entities" by InterPARES 2, but not to be confused with the term "entities" as used in metadata and database applications) that are the products and by-products of artistic, scientific and governmental activities that are carried out using dynamic, interactive or experiential digital systems. The nature of these entities was determined through a diplomatic analysis undertaken as part of extensive case studies of digital systems that were conducted by the InterPARES 2 Focus Task Forces. This diplomatic analysis established whether the entities identified during the case studies were records, non-records that nevertheless raised important concerns relating to reliability and authenticity or "potential records." To be determined to be records, the entities had to meet the criteria outlined by archival theory—they had to have a fixed documentary format and stable content. It was not sufficient that they were considered to be or were treated as records by the creator. "Potential records" is a new construct that indicates that a

[1] "Recordkeeping" is used in the archival literature in the context of the records continuum to signify an archival worldview of the integration and continual interactivity of processes and responsibilities related both to records creation and to archival management of those records. However, this is not a universally accepted premise, with the lifecycle model drawing a much clearer demarcation between the management of active records and the preservation of archival records. In the Chain of Preservation activity model developed by InterPARES 2, which is based upon the lifecycle model, "recordkeeping" refers to the phase in the lifecycle that comes between "record creation" and "record preservation."

digital system has the potential to create records upon demand, but does not actually fix and set aside records in the normal course of business. The work of the Description Cross-domain, therefore, addresses the metadata needs for all three categories of entities.

Finally, since "metadata" as a term is used today so ubiquitously and in so many different ways by different communities that it is in peril of losing any specificity, part of the work of the Description Cross-domain sought to name and type categories of metadata. The Description Cross-domain also addressed two areas of increasing importance in the digital environment: incentives for creators to generate appropriate metadata; and management issues associated with the retention, maintenance and eventual disposition of the metadata that aggregate in exponentially increasing amounts around digital entities over time.

Research team

The following is a list of researchers and research assistants who participated in the Description Cross-domain Task Force at some point throughout the Project.

Chairs and Co-chairs:

Terry Eastwood	2005-2006 (Chair)
Anne Gilliland	2001-2005 (Co-chair)
Sue McKemmish	2001-2005 (Co-chair)

Researchers:

Martine Cardin	Lavalle University, Quebec, Canada
Terry Eastwood	The University of British Columbia, Canada
Anne Gilliland	University of California, Los Angeles, USA
Hans Hofman	National Archives of the Netherlands
Richard Marciano	San Diego Supercomputer Center, USA
Victoria McCargar	consultant, Los Angeles Times, USA
Sue McKemmish	Monash University, Melbourne, Australia
Joe Tennis	University of Washington, USA
James Turner	Université de Montréal, Canada
Stefano Vitali	Archivio di Stato di Firenze, Italy

Research Assistants:

Bart Ballaux	The University of British Columbia, Canada
Lauren Cardinal	University at Albany, State University of New York, USA
Chia-Ning Chiang	The University of British Columbia, Canada
Joanne Evans	Monash University, Melbourne, Australia
Michael Garabedian	University of California, Los Angeles, USA
David Gibbs	University of California, Los Angeles, USA
John Juricek	University of California, Los Angeles, USA
Eleanor Kleiber	The University of British Columbia, Canada
Alison Langmead	University of California, Los Angeles, USA
Tracey P. Lauriault	Carleton University, Ottawa, Canada
Monique Leahey-Sugimoto	University of California, Los Angeles, USA
Lori Lindberg	San Jose State University and UCLA, USA
Jennifer Osorio	University of California, Los Angeles, USA

Catherine Miller The University of British Columbia, Canada
Rachel Mills The University of British Columbia, Canada
Shaunna Moore The University of British Columbia, Canada
Randy Preston The University of British Columbia, Canada
Nadav Rouche University of California, Los Angeles, USA
Wendy Sokolon The University of British Columbia, Canada
Emily Staresina University of California, Los Angeles, USA
Stuart Sugarbread University of California, Los Angeles, USA
Shannon Supple University of California, Berkeley, USA
Melissa Taitano University of California, Los Angeles, USA
Holly Wang University of California, Los Angeles, USA
Eun Young University of California, Los Angeles, USA
Yuchai Zhou Carleton University, Ottawa, Canada

Research Questions

Metadata investigations in the digital environment tend to cover a lot of territory, and the scope of the Description Cross-domain as determined in the research proposals funded by the various agencies that supported this work reflect that. The overall work was directed by the questions posed in the Project funded by the Social Sciences and Humanities Research Council (SSHRC) of Canada:

- What is the role of descriptive schemas and instruments[2] in records creation, control, maintenance, appraisal, preservation and use in traditional recordkeeping systems in the three focus areas?
- What is the role of descriptive schemas and instruments in records creation, control, maintenance, appraisal, preservation and use in emerging recordkeeping systems in digital and Web-based environments in the three focus areas? Do new tools need to be developed and, if so, what should they be? If not, should present instruments be broadened, enriched, adapted?
- What is the role of descriptive schemas and instruments in addressing reliability, accuracy and authenticity requirements (including the InterPARES 1 Benchmark and Baseline Authenticity Requirements) concerning the records investigated by InterPARES 2?
- What is the role of descriptive schemas and instruments in archival processes concerned with the long-term preservation of the records in question?
- Do current interoperable frameworks support the interoperability of descriptive schema and instruments across the three focus areas? If not, what kinds of frameworks are needed?
- What are the implications of the answers to the above questions for traditional archival descriptive standards, systems and strategies? Will they need to be modified to enable archival programs to meet new requirements, or will new ones need to be developed? If so, what should they be?
- To what extent do existing descriptive schemas and instruments used in the sectors concerned with the focus areas addressed by this project (for example, the geospatial data community) support and inform requirements such as those developed by InterPARES 1?

[2] This phrase is used throughout to refer to metadata in the broadest sense, as well as to archival description specifically.

Will they need to be modified to enable these sectors to meet these requirements, or will new ones need to be developed? If so, what should they be?

What is the relationship between the role of descriptive schemas and instruments needed by the creator and those required by the preserver to support the archival processes of appraisal, preservation and dissemination? What tools are needed to support the export/import/exchange of descriptive data between systems?

- What is the role of descriptive schemas and instruments in rights management and in identifying and tracking records components, versions, expressions, performances and other manifestations and derivative works?
- Is it important to be able to relate the record of artistic and scientific activity to the associated expression, performance, product, work or other manifestation of it and, if so, in what ways can descriptive activities facilitate it?

Additional research direction came from the projects funded by the United States National Science Foundation (NSF) and the National Historical Publications and Records Commission (NHPRC) that supported the U.S. Team's participation in InterPARES 2. This included formulation and testing of metadata models; and identification of new and existing methodologies and strategies for ensuring that records created using interactive, experiential and dynamic systems can be trusted as to their content (that is, are reliable and accurate) and as records (that is, are authentic) while used by the creator; new and existing methodologies and strategies for selecting records that have to be kept for legal, administrative, social or cultural reasons after they are no longer needed by the creator; new and existing methodologies and strategies for preserving them in authentic form over the long term; and advanced technologies for the implementation of these methodologies in different sectors and disciplinary and socio-cultural contexts. The research was also to develop hypotheses of metadata necessary for prototype systems; and rules for the ongoing description of digital records.

In the course of its work, the Description Cross-domain surfaced and addressed several additional provocative questions:

- Can a vocabulary be created to assist in the identification of different types and functions of metadata?
- What kind of management regime needs to be put in place to ensure the creation and maintenance of trustworthy metadata?
- Can metadata associated with the creation and active use of records ever contribute to archival description, particularly in the capture and elucidation of certain kinds of context and fundamental identification and arrangement information relating to the records?
- Should a metadata specification model generated out of InterPARES 2 support a single or multiple worldviews on the activities, roles, responsibilities and points of engagement with the record (e.g., lifecycle, records continuum and information continuum perspectives)?
- Can metadata-based automated tools support any new kinds of roles and capabilities for the description and use of preserved digital materials?

The latter questions have particular relevance for specifying how the benchmark and baseline requirements developed in InterPARES 1 and discussed further below, are implemented within recordkeeping and archival processes and systems design, as well as for the conceptualization and labelling of the models being developed.

Research Methodology

Multiple, interdependent activities and associated methods were used to generate products and data that could be triangulated to answer the questions outlined above (the researchers primarily engaged in each activity are indicated in parentheses).

- Collecting, compiling and analyzing data on the types and sources of metadata used in real-life dynamic, interactive and experiential systems as identified through case and general studies in arts, science and government settings that were conducted in other InterPARES 2 groups. Method used: *case studies* (focus group case study researchers, UBC project staff, Gilliland).
- Conducting a special investigation to identify state-of-the-art thinking and practice relating to metadata in news archives. Method used: *survey* (McCargar, Supple).
- Developing a database for analyzing warrant (i.e., the mandate from law, professional best practices, professional literature and other social sources) requiring the creation and continued maintenance of description and other metadata supporting the accuracy, reliability, authenticity and preservation of records and other record-like objects. This warrant will be integrated into public recommendations made by the Description Cross-domain and other InterPARES 2 research units with regard to evaluating, extending or revising existing descriptive and metadata schemas; encouraging the creation of meaningful metadata in the arts, science and government; as well as promoting the Metadata Specification Model in systems design. Method used: *literary warrant analysis* (Researchers: Gilliland, Sugarman, Gibbs, Garabedian).
- Developing and compiling a metadata schema registry that unambiguously describes salient features of relevant extant descriptive and other metadata schemas, element sets, standards and application profiles; and identifies existing cross-walks between them. Methods used: *iterative systems design* (Researchers: Gilliland, McKemmish, Hofman, Marciano, Lindberg, Evans, Rouche, Wang, Leahey-Sugimoto, Langmead, Zhou[3]).
- Developing an analytical framework for assessing the extent to which current metadata sets and implementations meet the requirements of the InterPARES benchmark and baseline requirements and/or the ISO Records Management Metadata Standard requirements (subsequently integrated with the registry to create the Metadata and Archival Description and Analysis System (MADRAS); and identifying how such metadata could be extended or modified to meet better recordkeeping requirements. Methods used: *requirements operationalization, warrant analysis, schema analysis, metadata mapping* (Researchers: Gilliland, McKemmish, Hofman, Marciano, Lindberg, Evans, Rouche, Wang, Leahey-Sugimoto, Langmeade, Youn).
- Developing metadata specifications to accompany the activity models constructed by the Modeling Cross-domain. The specifications identify the type, source and application of metadata implicit or explicit in the models and when, how and by whom it should be created.[4] These specifications can also form the basis for developing automated tools (not to be confused with descriptive instruments) that can be used to assist with the creation,

[3] Yuchai Zhou (2005), "Profiling and Visualizing Metadata for Geo-referenced Multimedia Information in a Geospatial Portal: A Case Study for the Cybercartography and the New Economy Project" (Master's thesis, Department of Geography and Environmental Studies, Carleton University, 2005).

[4] The metadata specification model for the Business-driven Recordkeeping Model developed by the Modeling Cross-domain is still to be developed.

capture, management and preservation of essential metadata for active and preserved records. Method used: *modeling and empirical instantiations* (Researchers: Tennis, Eastwood and Preston).

- Interfacing with other relevant research and development activities such as the development of the ISO 23081 Records Management Metadata Standard, the Monash University-based Clever Recordkeeping Metadata Project[5] and the work of the San Diego Supercomputer Center on the development of metadata tools for the automated creation, harvesting and end-user manipulation of metadata (Hofman, Gilliland, McKemmish, Marciano, Evans and Lindberg).

Figure 1 illustrates the relationships between the constituent components and some of the associated activities of the Description Cross-domain. Numbers 1-3 on the flowchart indicate the primary loci of activity and eventual products.

Metadata and Archival Description Registry and Analysis System (MADRAS)[6]

MADRAS was initially envisioned as a metadata registry that could be used by the Description Cross-domain to identify relevant metadata sets and schemas that it wished to evaluate to generate recommendations in response to its research questions. However, it quickly became clear that if the Description Cross-domain was to operate on the assumption that metadata were essential to the creation of reliable and preservation of authentic records in electronic systems of any type, then it also needed to address issues associated with how trustworthy metadata are created and maintained. It was also clear that the Description Cross-domain needed to operationalize the benchmark and baseline requirements generated by InterPARES 1 in terms of how they might be met through metadata and archival description. MADRAS evolved, therefore, beyond being a schema-level (i.e., not a comprehensive element-level) metadata registry, to include an analytical assessment tool that could be used by the researchers to evaluate the current capabilities of registered metadata schemas. With an extension of U.S. research funds until June 2007, it is now envisaged that the beta production version completed in InterPARES 2 and used by the Project's researchers to answer their research questions, will be revised as a full-fledged, publicly available metadata assessment and tracking tool with more sophisticated public interfaces, report formats and privacy controls that will support those who wish to register proprietary or draft schemas.

---

[5] See Records Continuum Research Group (1998), "Create Once, Use Many Times - The Clever Use of Metadata in eGovernment and eBusiness Processes in Networked Environments." Available at http://www.sims.monash.edu.au/research/rcrg/research/crm/.

[6] For further details on the development of MADRAS, see Anne J. Gilliland, Nadav Rouche, Joanne Evans and Lori Lindberg (2005), "Towards a Twenty-First Century Metadata Infrastructure Supporting the Creation, Preservation and Use of Trustworthy Records: Developing the InterPARES 2 Metadata Schema Registry," *Archival Science* 4(1): 43–78; Joanne Evans and Nadav Rouche (2004), "Utilizing Systems Development Methods in Archival Systems Research: Building a Metadata Schema Registry," *Archival Science* 4(3-4): 315–334; Joanne Evans and Lori Lindberg, "Describing and Analyzing the Recordkeeping Capabilities of Metadata Sets," in *DC-2004: Proceedings of the International Conference on Dublin Core and Metadata Applications, October 11-14 2004, Shanghai, China* (Shanghai, China: Shanghai Scientific and Technological Literature Publishing House, 2004), 75–80. Online reprint available at http://www.dublincore.go.kr/dcpapers/pdf/2004/Paper_27.pdf; Anne J. Gilliland-Swetland and Sue McKemmish, "A Metadata Schema Registry for the Registration and Analysis of Recordkeeping and Preservation Metadata," in *Proceedings of the Second IS&T Archiving Conference, April 26-29, 2005, Washington, D.C.* (Springfield, VA: Society for Imaging Science and Technology, 2005), 109–112; and Lori Lindberg, Monique Leahey-Sugimoto, Nadav Rouche and Holly Wang, "MADRAS: A Metadata and Archival Description Registration and Analysis System for the Analysis of the Recordkeeping Capabilities of Metadata Sets," in *Proceedings of the Third IS&T Archiving Conference* (Springfield, VA: Society for Imaging Science and Technology, 2006), 216–218.

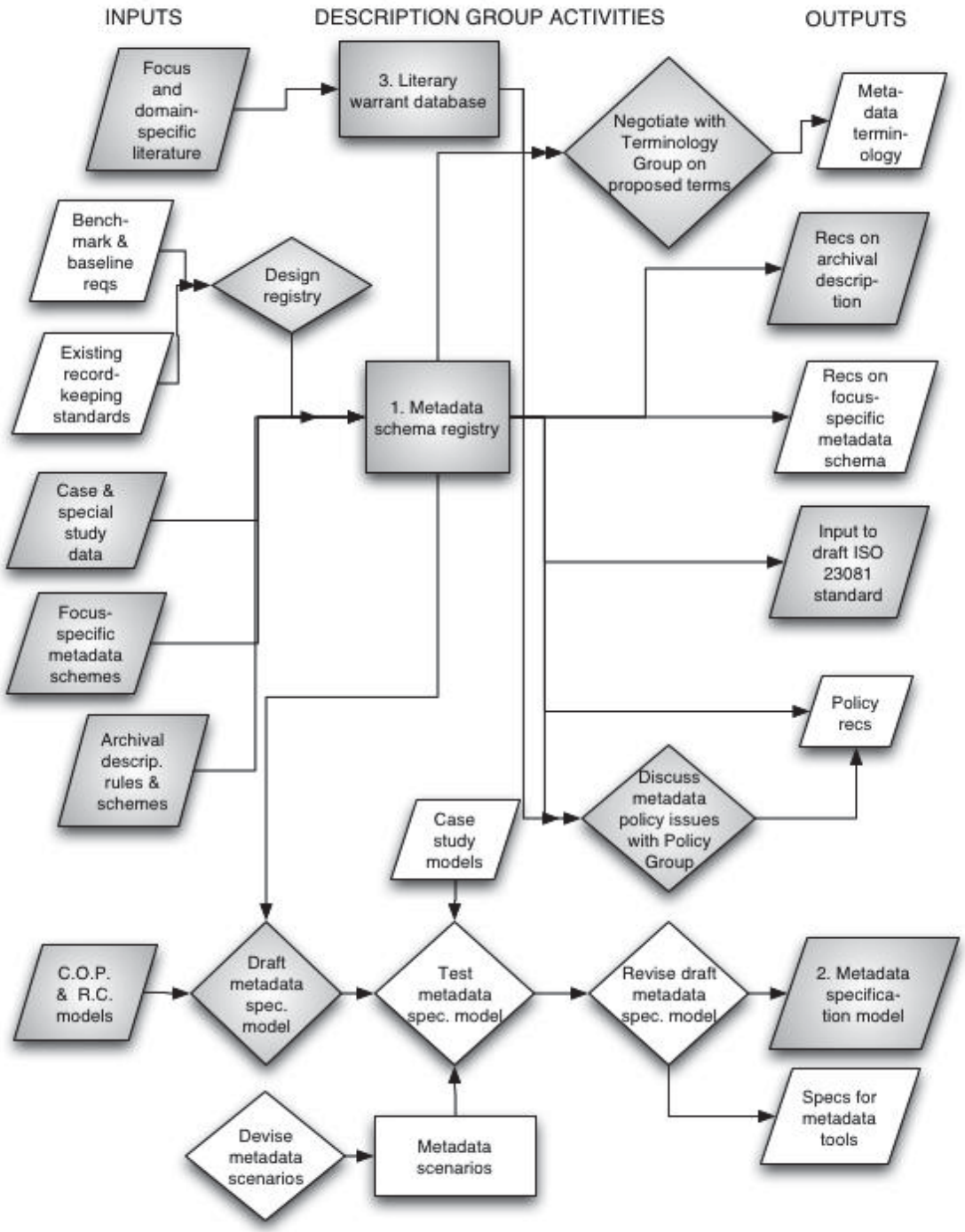

Figure 1. Flowchart of Description Cross-domain Activities

The purpose of MADRAS is fourfold:

1. To support the unambiguous registration of relevant metadata schemas, sets and application profiles;
2. To support the analysis of registered items against requirements derived from the InterPARES 1 benchmark and baseline requirements as well as from the ISO 23081 Records Management Metadata Standard and to make recommendations for how they might be extended or otherwise revised to address the reliability, authenticity and preservation needs of records created within the domain, community or sector to which they pertain.
3. To provide a standardized framework by which any existing or draft metadata schema or set can be assessed for its ability to address the above mentioned requirements and which could be adopted by standards-setting bodies in different areas of practice.
4. To generate analytical data to be provided to the working group (ISO TC46/SC11-WG1) that oversees the development of ISO 23081 for possible incorporation into Part III of that standard.

The inputs for MADRAS development included the following:

a. The benchmark and baseline requirements generated by InterPARES 1.
b. Requirements derived from an analysis of ISO 23081.
c. Requirements derived from analysis of other salient electronic records standards and projects, including the conceptual and relationship models of records in business and socio-legal contexts developed by the SPIRT Recordkeeping Metadata Project and Kate Cumming's "Derivation of the Classification of Recordkeeping Metadata by Purpose Scheme."[7]
d. Metadata schemas and sets identified in the course of the case and general studies undertaken by the focus groups.
e. Other relevant focus-specific metadata schemas and sets identified by the focuses or by the Description Cross-domain (e.g., Geomatics Metadata Standard, ISO 19115).
f. Archival description rules, sets and related practices (e.g., ISAD(G)/ISAAR, EAD/EAC/ DACS, RAD and the Australian Series System).

The current beta environment for MADRAS is implemented using PHP, a server-side scripting language that provides Web development tools for building dynamic Web sites. The back-end Web server is Apache 1.3 and the database server is MySQL 3.22. Both servers are hosted on a machine running the Unix operating system. PHP, Apache and MySQL are all open-source technologies and are used by many database-driven Web applications. Information about the process of building MADRAS has been kept in MADRAS itself using an online note sharing system. The current size of MADRAS is 20 megabytes (without appended documents) with around 100 PHP files. More files will be generated in conjunction with the development of the analysis interface. The researchers expect that MADRAS will grow into a mid-sized application after processing more feedback from InterPARES researchers and adding more data and infrastructure. MADRAS is allowed 50,000 queries per hour from the database server and MySQL 3.22 has a 4-gigabyte limit on table size (limitations are a function of MySQL).

[7] See Kate Cumming, "Purposeful Data: The Roles and Purposes of Recordkeeping Metadata" (Ph.D. dissertation, Monash University, 2005).

MADRAS registry component and the schema registration framework

As Chris Hurley has noted:

> Contextual metadata documents circumstances relevant to the making of the record: who, when, how, why … Efforts now being made to regularize the process whereby knowledge of context is captured as metadata for electronic record-keeping should not blind us to a fundamental truth. Because records themselves are timebound, metadata must be verified within a context which is both current *and* historical. Records cannot remain current unless the metadata is externally validated.[8]

Hurley is arguing that beyond the comprehensive and rigorously delineated metadata and archival description necessary for creating reliable records and maintaining and demonstrating the authenticity of archival records, there is a need for overt integrity control and transparency of those metadata and of archival description. This can only be the case if the metadata themselves are trustworthy and comprehensively managed for as long as they are required. In other words, reliability and authenticity are concerns for recordkeeping metadata as well as for the records and recordkeeping processes to which they relate. Metadata generated and managed by records creators and archival description generated by archivists, must be sufficient, appropriate, understandable and of high quality. MADRAS and the metadata specification model, therefore, are two tools that seek to support a highly reflexive recordkeeping metadata regime that addresses both of these concerns.[9]

The MADRAS registry component was developed with the following primary purposes in mind:

- to describe relevant metadata schemas and their features in a standardized way;
- to provide an overview of existing and emerging schemas;
- to provide an overview of the applicability of the schemas to recordkeeping and archival functions;
- to describe the scope and purpose of the schemas;
- to specify what type(s) of metadata they cover; and
- to identify related schemas (e.g., schemas that control data values, schemas that provide structure for metadata elements).

With one of the expected outcomes of the Description Cross-domain research within InterPARES being a production of "scholarly comparative discussions of existing descriptive standards and an intellectual framework of descriptive standards for the records under examination [within InterPARES 2],"[10] MADRAS was developed to act as a data collection and analysis tool for InterPARES 2 researchers. After developing a framework for the standardized description of metadata schemas, a metadata schema itself was produced in the form of an XML DTD. From the DTD, a prototype database was developed to assist researchers in the refinement of the DTD

[8] Chris Hurley (1995), "Ambient Functions: Abandoned Children to Zoos," *Archivaria* 40 (Fall): 22. Emphasis in original. Online reprint available at http://journals.sfu.ca/archivar/index.php/archivaria/article/viewFile/12095/13080.

[9] Archives have always been metadata-rich environments, although they are not always recognized as such, just as archival description is not always recognized by archivists as the primary means by which they demonstrate the authenticity of their holdings. Archivists must be cognizant that the accession records, finding aids and use records they typically create today are not only part of the archival description for the records to which they relate, but they are also records in their own rights. The scrutiny, therefore, that archivists give to the records and recordkeeping metadata of others to assess and validate their management and reliability, they must also give to their own.

[10] InterPARES 2 Project, "Overview of InterPARES 2 Intellectual Framework," 7. Available at http://www.interpares.org/display_file.cfm?doc=ip2_overview_of_intellectual_framework(20030311).pdf.

design. Now that the production version of the registry is operational, the prototype database has been retired. MADRAS is the result of the lessons learned from the prototype database.

The following outlines the development process and design decisions involved in the building of MADRAS:

- The decision to develop the registry as a way to approach the Description Cross-domain research was based upon the realization that it was impossible to assess all relevant schemas within the time available to the Project and also that any such assessment would date rapidly, given the current pace of schema evolution. Instead, researchers decided to develop a tool that could be used into the future by any party wishing to register and assess a schema they were using or planned to use against InterPARES requirements, as well as discover other schemas, view them and their assessments and learn about them in a consistent, structured environment. This decision is significant because it reflects a pragmatic approach to the political realities of metadata schema creation and use. Schemas have proliferated in many communities and are closely tailored to their specific needs. The Description Cross-domain decided that it was very unlikely that any community would adopt a schema developed by InterPARES in place of or in addition to its own. Instead, the approach adopted demonstrates how interested parties can use their own community or implementation-specific schemas, compare them to others and begin to think about them in the larger world of metadata schema development as well as in different ways. This reflexive thinking was evident in the Description Cross-domain researchers' own discovery processes when they decided additionally to address in the MADRAS analysis component, the requirements contained in the ISO 23081 standard, so that users could both assess their schemas and compare differences between recordkeeping metadata requirements as articulated by the InterPARES Project research and another research collaboration, that within the ISO.
- The first step toward developing a registry was to develop a draft XML Document Type Definition (DTD) that would become the backbone of the registry. XML was chosen because of its platform independence, flexibility at handling hierarchical data and relative ease of migration. In its original conception, MADRAS was to be an integrated description and analysis tool, with all data encoded within the DTD. The researchers decided to move ahead with the analysis component of MADRAS simultaneously using an Excel spreadsheet-based worksheet, allowing for parallel work activities while the DTD was being tested and a prototype registry built. A form of the analysis worksheet was originally intended for integration into the DTD. However, the analysis became such a large and significantly complex component it was not integrated into the DTD and became an independent part of the registry.
- To develop the DTD, the researchers examined how metadata about metadata schemas should be sourced to ensure their reliability and authenticity, for example, through recordkeeping requirements for metadata registries described in the ISO standard for metadata registries.[11] In addition, the researchers were mindful that the description data within the registry component was not required to conduct an in-depth analysis of a schema but rather to extract structured objective information about a schema as it is described in schema documentation; for example, an official name of a schema, its acronym, publisher information, documentation pointers or citation information and copyright statements.

[11] International Organization for Standardization, International Electrotechnical Commission, ISO/IEC 11179: Information Technology—Metadata registries (MDR). Available at http://metadata-stds.org/11179/index.html.

- The registry DTD was developed with a classification hierarchy of elements three levels deep. Level one elements corresponded to major sections of metadata about metadata schemas (hereafter descriptors):
  REGISTRATION
  IDENTIFICATION
  TECHNICAL REQUIREMENTS
  RIGHTS
  PROVENANCE
  DESCRIPTION
  DOCUMENTATION
  RELATIONSHIPS
  NOTE
  Each of the Level one elements (except NOTE) possessed one or more Level two sub-elements and some Level two sub-elements possessed one or more Level three sub-elements.
- Once the registry DTD was developed, the researchers identified multiple key metadata and descriptive schemas and sets (both from the archival field and from those in use in sectors within the three InterPARES 2 focus areas—arts, science and government) and registered them in a prototype database to test and refine MADRAS database. This prototype database, built in Microsoft Access, allowed the Description Cross-domain researchers to view the registry records and test the DTD as an encoding standard for the registration and description of metadata schemas. Visualizing and working with the data input as well as with the descriptors (the DTD elements) in the prototype registry allowed for the rearrangement of DTD elements, identification of mandatory elements and the proposal of possible controlled vocabularies for certain element values, along with the identification of the need for particular element data value encoding schemas such as ISO 8601 for representation of dates and other relevant ISO standards for forms of country names, languages and so on.
- Guidelines for registering and describing schemas were developed and refined as the DTD researchers' experience with the system increased. To test these guidelines and to check for intercoder consistency, graduate students who had not previously been involved in MADRAS development were assigned schemas to register. In addition, students from other research laboratories not familiar with archiving metadata methods also assisted with the registration of their discipline-specific metadata standards. This provided valuable feedback from persons who were viewing the registry for the first time and who were not necessarily from recordkeeping backgrounds.
- Documentation of system functionality and requirements was developed to support the ability to maintain the system and facilitate the eventual transfer of it from UCLA, where it was developed, to a maintenance agency.

The registration process for MADRAS involves manual entry of values into templates. Population of the prototype database demonstrated wide variation in how schemas are published and information about them is presented. In such circumstances, manual processes involving human cognition, collation and data entry appear to be the only viable registration method, since humans are best able to negotiate the situation-specific mappings and cope with gaps and ambiguities in the schemas. Utilizing such an approach, however, also introduces scalability and sustainability issues for MADRAS, given the amount of manual processing required. It points to

the need for standardization in the way metadata standards, schemas, crosswalks and their meta-information are published so that registration can be automated or at least semi-automated. This also raises the question of what meta-information should be made available as part of the publication of metadata standards and for the consumption of what types of agents.

The development of the registry component of MADRAS was an iterative process that was striving for an ideal system. Even through the tremendous intellectual capital invested in the prototype database, the DTD was not able to address fully all of the researchers' questions, and the difficulties in so doing provided valuable insight into some of the metadata issues being addressed. Additional valuable questions were raised during the transformation of the DTD and the prototype into the production version of MADRAS and the building of the technological infrastructure for distributed registration and analysis activity in the Web environment.

Analytical framework

The analytical component of MADRAS was developed through iterative prototyping and warrant analysis over a period of three years. The technique of warrant analysis was employed to determine the criteria against which judgments as to the recordkeeping and archival capabilities of metadata schemas could be made. The process involved studying each warrant for statements made regarding requirements for recordkeeping metadata and turning these into a series of questions. These questions were then compiled into an analysis worksheet using an Excel spreadsheet. Although there was a degree of overlap in these statements, the strategy was to have separate sections for each warrant as part of the data gathering that would feed into the metadata model developments.

A primary set of conditions against which metadata schemas registered in MADRAS are assessed is the benchmark and baseline requirements that were generated out of the InterPARES 1 Project.[12] The benchmark requirements are based on the notion of a trusted recordkeeping system. They include requirements that support the presumption of the authenticity of digital records before they are transferred to the preserver's custody. The baseline requirements are based on the notion of the preserver as trusted custodian and support the production of authentic copies of digital records after they have been transferred to the preserver's custody. These are the only extant sets of requirements that specifically address how creators and archivists can assess the authenticity of records. As noted by Evans and Lindberg:

> The benchmark requirements identify the record attributes (metadata) that need to be 'explicitly expressed and inextricably linked' to a record in order for its identity and integrity to be asserted. The benchmark requirements also identify 'the kinds of procedural controls over the record's creation, handling and maintenance that support a presumption of its integrity.' The role of the benchmark requirements is to act as a tool for preservers to use in assessing the authenticity of electronic records. The higher the number, and the greater the degree to which a system meets these requirements, then the stronger the presumption of the authenticity of the electronic records held within it.

[12] See 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*, Luciana Duranti, ed. (San Miniato, Italy: Archilab, 2005), 204–219. Online reprint available at http://www.interpares.org/book/interpares_book_k_app02.pdf. Abridged versions of the benchmark and baseline requirements are provided in Appendices 21a and 21b, respectively. Available at http://www.interpares.org/display_file.cfm?doc=ip2_book_appendix_21.pdf.

> In contrast, the baseline requirements specify the requirements that must be met in order to produce authentic copies of electronic records from a preservation system. This includes archival descriptive metadata documenting 'the records juridical-administrative, provenancial, procedural and documentary contexts,' and controls over the records transfer and reproduction processes to ensure the maintenance of the records' identity and integrity.[13]

As this excerpt indicates, many of the benchmark requirements could potentially be implemented through metadata and archival description, particularly such aspects as identity, linkages, documentation of documentary forms, juridical requirements, business rules and technical procedures, access privileges, establishment of the authoritative record when multiple copies exist and transfer of relevant documentation; as could almost every aspect of the baseline requirements. The benchmark and baseline requirements, however, had only been expressed conceptually, and in narrative form, by InterPARES 1 and were not operationalized for any kind of technological implementation; for example, as a set of logical propositions or production rules. Nor were the requirements deconstructed in a way that would specify how other processes and metadata might help to meet them. For example, how might the different types of context identified in InterPARES 1 be manifested or documented through metadata? One way of addressing a problem such as this is to decompose archival and recordkeeping notions of "context" into types that can then be associated with specific processes and attributes. InterPARES 1 identified five different types of contexts as being relevant to the maintenance of authentic records over time: juridical-administrative, provenancial, procedural, documentary and technological.[14] Some of these types need to be further decomposed to identify their constituent metadata manifestations.[15]

Accordingly, the development of the analytical framework used in MADRAS sought to operationalize these narrative requirements in terms of how they might be satisfied both through the metadata associated with the active record and recordkeeping system and archival description. The same then had to be done for the ISO 23081 requirements, which were also narratively expressed. Once the framework was drafted, Description Cross-domain researchers analyzed multiple existing schemas, standards and guidelines to assess the extent to which they met the requirements, given their stated scope. Where the analysis indicates that a schema falls short, the output report generated by MADRAS delineates exactly where and how and researchers can then recommend augmentations or modifications to ensure that the schema meets those requirements that fall within its stated scope. MADRAS can also be used to identify potential companion metadata schemas that can be used to address those parts of the requirements that are unaddressed because they are out of scope (e.g., because the schema addresses the creator or the preserver side only, or is content rather than context- or recordkeeping-centric). When the beta system becomes publicly available, anyone will be able to

[13] Evans and Lindberg, "Describing and Analyzing the Recordkeeping Capabilities of Metadata Sets," op. cit.

[14] See Authenticity Task Force (2001), "Appendix 1: Template for Analysis," in Duranti, *Long-term Preservation*, op. cit., 198–203. Online reprint available at http://www.interpares.org/book/interpares_book_j_app01.pdf.

[15] For example, the juridical-administrative type could potentially be decomposed to address specific types of juridical-administrative requirements that manifest themselves directly in emerging metadata initiatives, such as those relating to rights management for records. Digital rights management (DRM) metadata are increasingly being integrated into systems by creators, publishers and information providers, for example, as mechanisms for expressing and automatically enforcing rights and licensing requirements relating to information resources. In an age where records are more and more often the product of private activity or collaboration or of outsourcing relationships between government and the private sector or academic research, collaborative science and industry, such developments not only reflect these changes in records creation but can have significant implications for both researchers and the types of preservation regimes to which the records may be subject.

register and evaluate a current or draft schema or application profile. In this way, the analytical framework can be applied beyond the duration of the InterPARES 2 Project to assess schemas, sets and application profiles as they develop and evolve. This approach also ensures that multiple models for managing records can be supported—both those that seek to apply an end-to-end recordkeeping metadata schema and those where different parties have responsibility for different aspects of recordkeeping and archival preservation.

To draw on as many perspectives as possible and to try to identify where there might be consensus or divergence about relevant recordkeeping requirements (especially where there might appear to be differing viewpoints emerging from the lifecycle and records continuum perspectives), several other prominent standards, guidelines and requirements were also consulted, including ISO 15489 Information and Documentation—Records Management (2001), the U.S. Department of Defense's Design Criteria Standard for Electronic Records Management Software Applications (DoD 5015.2-STD, 2002), and the European Union's Model Requirements for the Management of Electronic Records (MoReq) that specifies requirements for Electronic Records Management Systems (ERMS).

The decomposed requirements were conceived and expressed in the analytical framework in the form of evaluative questions, with the questions designed primarily to elicit a positive or negative response. For positive responses, a schema's element or elements that satisfied a particular question could be noted. The original Excel spreadsheet was organized to systematically describe schemas and assess them over seven sections: (1) General; (2) Recordkeeping General; (3) ISO 23081; (4) InterPARES benchmark requirements; (5) InterPARES baseline requirements; (6) Classification of Purpose of Recordkeeping Metadata; and (7) General Comments.

The questions were then coded to specific sections of these two instruments so that an actual analysis could be performed.[16] The structure of the worksheet, the nature of the individual questions and the analysis process as a whole was defined and refined through iteration and testing. The questions were applied to a sample of schemas to determine their feasibility, granularity and usefulness as well as the meaning of the response. Schemas included in the sample were selected on the basis of being able to help in determining whether the analysis could make distinctions between recordkeeping and non-recordkeeping schemas, between "single" and "multi-entity" schemas and between schemas operating in different dimensions.

The first attempt to organize the analysis questions was based on a view of what metadata are supposed to do (for example, describe record content, context and structure and then recordkeeping activities). However, to facilitate user comprehension, it was eventually decided to separate the questions by the different recordkeeping entities suggested by the instruments: Record, Agent, Mandate, Business Process and Recordkeeping. To do this, the researchers employed an iterative development process, focused on refining and arranging questions. They paid careful attention to the ways in which each instrument used its own terminology and brought that forward into the analysis questions.

The initial statement of requirements was progressively refined through the development of a prototype database and its population, with a sample of metadata schemas. This process helped to ensure that a flexible descriptive schema was developed that could cope with the diversity of metadata schema publication and documentation practices. It also enabled the testing of the feasibility and applicability of the proposed elements and determination of the sources of metadata values.

[16] See Appendix 17.

The researchers decided that the first iteration of the system would be for InterPARES' researchers themselves and then the system should be revised for future use by other interested stakeholders (i.e., records keepers, archivists, etc.). The analysis worksheet underwent a number of versions and changes through the initial testing and validating that resulted in a final accounting of four major versions of the worksheet with smaller subversions (4.1, 4.2 and so forth). First, the analysis was mocked-up in Excel. Later, during the design development phase, FileMaker Pro was used to work up a model for the display of information in MADRAS that was eventually recreated in the actual MADRAS system.

Challenges encountered in the development of the analytical framework ranged in complexity. Often, it was necessary to return to first principles. For example, during the process of creating MADRAS, researchers needed to come to an agreement (or not) on the operational meaning of the word "record." What did they consider to be a record? A relationship? Along the same lines, researchers needed to consider what the base unit of analysis should be (in other words, to what level of granularity should the analysis proceed?). In the end, the decision was made that the system would proceed to the element and not to the sub-element level.

The researchers experimented with developing various versions of a decision tree. Lacking consensus, they decided not to use any of the versions in the current production version, but did agree to revisit the use of a decision tree in a later version. The process did, however, help with the decision to push certain questions to the registry and table relationships in the analysis until it was decided whether or not a relationship should be elevated to its own entity. Some of the other activities involved in the framework development included the following:

- mapping between related InterPARES and ISO requirements;
- developing controlled vocabularies for classifying the purpose of schema and standards, and for types of metadata specified in schema and standards (drawing on ISO 23081, the SPIRT Recordkeeping Metadata Research Project outcomes and the Records Continuum and InterPARES Models).; and
- exploring the boundaries between and around records and related metadata, and noting that some metadata relate to the content, structure (documentary form) and business context of the record (concerned with the nature of the business transaction captured in the record), and that some metadata relate to the recordkeeping processes that manage the record.

The analysis worksheet stayed fairly stable until the spring of 2005, when the shift from the manual worksheet-based analysis to an automated version of the analysis began. The automation of the analysis process, a goal of the MADRAS tool development, surfaced a number of procedural and technical considerations, not the least of which was the time spent on manual analysis and the time spent to teach new analysts how to do the work. Research team members observed that the analysis reference instruments had a number of areas of overlap and that as a result similar questions that sought similar answers were asked over more than one section of the spreadsheet. The decision was made to map each of the reference instruments against one another to take advantage of commonalities amongst the instruments. This decreased the amount of repetitive work, as well as verified for the researchers that the research findings across the different projects producing the reference instruments came to some common conclusions. For example, when considering the *Classification of Purpose of Recordkeeping Metadata* schema developed by Kate Cumming, the researchers looked very carefully at her classification schema and where it might be expressed or assumed as the basis for requirements expressed in the remaining analysis reference instruments. Cumming concludes that all recordkeeping metadata are created to satisfy one of seven particular purposes:

1. unique identification;
2. authentication of records;
3. persistence of records content, structure and context: by fixing their content, ensuring that their structure can be re-presented, and maintaining sufficient organizational and functional context to preserve their meaning over time and beyond their context of creation;
4. administering terms and conditions of access and disposal;
5. tracking and documenting use history, including recordkeeping and archiving processes;
6. enabling discovery, retrieval and delivery for authorized users; and
7. restricting unauthorized use.[17]

It was determined that these purposes were all articulated in the warrants in one way or another and did not need explicit consideration as a separate grouping of questions in the analysis. The mapping of the reference instruments decreased the number of questions asked in the analysis, making the process more efficient and less time-consuming. In addition, it allowed the analysts to be able to look at the data produced in new ways and apply findings more broadly.

Automating the process of analysis also required re-thinking how consistency could be ensured across different analysts. The researchers were trying to automate a system that relied on an unknown: the extent of the human analyst's knowledge; and this raised interesting issues. The original method of analysis using Excel spreadsheets had demonstrated that analysis could vary considerably according to the knowledge and experience of the analyst. The researchers had to assume certain pre-existing knowledge on the part of the user of the system. It was decided that users would most likely be experienced records keepers or those familiar with archival terminology.

During the automation process, the strengths and weaknesses of the original analysis spreadsheets were assessed to clarify and bolster the effectiveness of MADRAS. This surfaced several issues with the original spreadsheets including that:

- The original worksheet facilitated documenting, rather than analyzing, a metadata schema. (Solution: focus on analyzing rather than on documenting the schema.)
- The original worksheet was repetitious. Information documented in one section was repeated in another. (Solution: eliminate redundancy.)
- The original worksheet and evaluation instruments had confusing language. (Solution: simplify and add documentation. For example, a definition file was created that strives to provide a single definition of terminology to assure analyst consistency.)
- The original worksheet was in a format that did not transfer easily to a database/online worksheet. (Solution: creation of an environment that was flexible enough to experiment with—a FileMaker prototype was created as a design sandbox.)
- The criteria for ranking schemas and evaluating answers were not clear. (Solution: create a system where as much ambiguity as possible could be eliminated.)
- The original analysis process did not allow for the discovery of other relevant types of metadata that might be present in a schema but not in any of the analysis instruments. (It was not possible to address this as the analysis was so strongly focused on the InterPARES and ISO instruments.)

[17] See Cumming, "Purposeful Data," op. cit.

- Although the original analysis process asked for repeatability and the obligation value for each element, the Excel worksheet did not ask for this information. (Solution: The researchers separated out the repeatability (or lack thereof) of fields as well as whether a field is mandatory into the element registration process.)

As the design process continued, the researchers conducted a series of user tests, which generated quite a bit of feedback used to improve the design of the system. They also focused on the creation of a tool where users answer questions about a schema and indicate precisely what elements the schema uses to fulfil a specific requirement.

The researchers attempted to confront the issue of how one separates what is explicitly stated in schema documentation and what is implicit, since they wished to create a tool that would test for the *explicit* nature of the metadata. This issue arose from the following section of ISO 23081:

> Records management has always involved the management of metadata. However, the digital environment requires a different expression of traditional requirements and different mechanisms for identifying, capturing, attributing and using metadata. In the digital environment, authoritative records are those accompanied by metadata defining their critical characteristics. These characteristics must be explicitly documented rather than being implicit, as in some paper-based processes.[18]

ISO interactions

Hans Hofman, National Archives of the Netherlands, served as both a member of the Description Cross-domain and as a member of TC46 SC11 WG01, the Technical Committee overseeing ISO 23081 development. He provided input to and feedback on the development of the registry and the analytical framework from the ISO perspective. One of the MADRAS developers, Lori Lindberg, also travelled to Paris to present the MADRAS work and get feedback directly from the Technical Committee. The feedback from that presentation was that the framework was too "record-centric," and so the researchers revised the framework somewhat to be more entity-focused.

MADRAS has been developed and constructed by researchers with varying knowledge of records and recordkeeping and drawn from disparate recordkeeping philosophies. Enduring challenges include how to accommodate the various audiences and communities that may utilize MADRAS and providing a transparency of the analysis process to accommodate those without a recordkeeping background who are concerned about these issues but are relatively unfamiliar with recordkeeping theory, processes and terminology. Another, more significant, challenge is how to construct and present questions that address the complexity of the metadata model behind ISO 23081 and the conceptual entities incorporated within the standard in a user-friendly manner. As the metadata counterpart to ISO 15489, the international records management standard, ISO 23081 is in itself quite detailed and complex, with multiple types of metadata accruing at various layers and at different times within a recordkeeping system. With ISO 23081 incorporating the significant findings about the authenticity of records developed within the InterPARES Project as well as the conceptual recordkeeping model behind the Australian Recordkeeping Metadata Standard, itself the basis for ISO 15489, this assessment tool must

[18] International Organization for Standardization, ISO 23081-1:2006 - Information and documentation—Records management processes—Metadata for records—Part 1: Principles, 2. It is also worth noting, however, that the question may be less a matter of the latency or explicitness of schema documentation than the shortcomings of using tools, such as plain XML, that do not allow for the specification of semantic constraints of entities and the relationship(s) between entities.

accommodate both of the major models of records management currently in use in the archives and records management communities—the lifecycle model as reflected in the InterPARES research and the continuum model developed in Australia.[19]

Data and data analysis

A list was generated of major metadata schemas and sets that are in use in the archival field as well as in the areas covered by each InterPARES 2 focus areas. These include:

- *ANZLIC Metadata Guidelines: Core Metadata Elements for Geographic Data in Australia and New Zealand*—defines metadata elements that describe characteristics of spatial datasets maintained in the public and private sectors;[20]
- *Arizona Electronic Recordkeeping Systems (ERS) Guidelines - IV Functional Requirements for Recordkeeping Systems*—describes specifications for recordkeeping functionality that should be incorporated into any digital information system to ensure it can produce records that are accepted as evidence, well managed and preserved, and that benefits are appropriate to the costs;
- *Australian Recordkeeping Metadata Schema (RKMS)*—defines a highly structured set of metadata elements that conforms to a data model based on that developed for the Resource Description Framework (RDF) and that is designed to be extensible and can inherit metadata elements from other schemas;[21]
- *CURL Exemplars in Digital Archives (CEDARS) Metadata for Digital Preservation*—defines a metadata specification based on the Open Archival Information System (OAIS) model that is designed to be used within the Cedars demonstrator services and as a contribution to international efforts at standardization on preservation metadata;[22]
- *Digital Rights Expression Languages (DREL)*,[23] *Online Information Exchange (ONIX) Metadata Specification*[24]—an international standard for representing and communicating book industry product information in electronic form;
- *eXtensible rights Markup Language (XrML)*—a general-purpose, XML-based specification grammar for expressing rights and conditions associated with digital content, services or any digital resource;[25]
- *Global Information Locator Service (GILS)*—an open standard for searching basic information descriptions based on the ISO 23950 search standard;[26]
- *ISO 19115:2003 Geographic Information—Metadata (geomatics metadata standard)*—defines the metadata elements required for describing geographic information and services, including the identification, the extent, the quality, the spatial and temporal schema, spatial reference and distribution of digital geographic data;[27]
- *ISO 82045-2 Document Management—Part 2: Metadata elements and information reference model*—provides a comprehensive set of standardized metadata elements for

[19] See Lindberg et al., "MADRAS," op. cit.
[20] Available at http://www.anzlic.org.au/download.html?oid=2358011755.
[21] Available at http://www.sims.monash.edu.au/research/rcrg/research/spirt/deliver/index.html.
[22] Available at http://www.leeds.ac.uk/cedars/MD-STR~5.pdf.
[23] Available at http://www.jisc.ac.uk/uploaded_documents/TSW0603.pdf.
[24] Available at http://www.editeur.org/onix.html.
[25] Available at http://www.xrml.org/.
[26] Available at http://www.gils.net/index.html.
[27] Available at http://www.iso.org/iso/en/CatalogueDetailPage.CatalogueDetail?CSNUMBER=26020.

document management, including data exchange and implementation of a document management system;[28]

- *MAchine-Readable Cataloging (MARC)*—defines a data format by which computers exchange, use and interpret bibliographic information;[29]
- *Metadata Encoding and Transmission Standard (METS)*—a standard for encoding descriptive, administrative and structural metadata regarding objects within a digital library, expressed using XML;[30]
- *Minnesota Recordkeeping Metadata Standard*—defines metadata elements developed to facilitate records management by government entities at any level of government;[31]
- *New South Wales Recordkeeping Metadata Standard (NRKMS)*—describes metadata that can be used by NSW public sector bodies to meet the business, accountability and archival requirements for records; based on the principles of AS 4390: 1996 (the Australian standard for records management);[32]
- *NISO Z39.87-2002 AIM 20-2002 Data Dictionary—Technical Metadata for Still Images*,[33] *Metadata for Images in XML (NISO MIX) Schema*[34]—defines a set of metadata elements for raster digital images to enable users to develop, exchange and interpret digital image files;
- *NLA Pandora AGLS Metadata Element Set*—metadata elements designed to improve the visibility, accessibility and interoperability of online information and services;[35]
- *Open Digital Rights Language (ODRL)*—a proposed language for the Digital Rights Management (DRM) community for the standardization of expressing rights information over content;[36]
- *PREMIS Data Dictionary for Preservation Data*—defines and describes an implementable set of core preservation metadata with broad applicability to digital preservation repositories;[37]
- *Preservation Metadata - Networked European Deposit Library (NEDLIB) Metadata for Long Term Preservation*—defines preservation metadata elements for a deposit system for electronic publications largely based on the OAIS model;[38]
- *Preservation of Electronic Records in a Records Management Application (PERM) Preservation Attributes*—designed for managing a persistent archives of electronic records created by desktop applications through use of an XML Archiving & Packaging Tool (XAPT);[39]
- *Record Keeping Metadata Requirements for the Government of Canada*—defines metadata elements that identify the type of information Departments are required to

[28] Available at http://www.iso.org/iso/en/CatalogueDetailPage.CatalogueDetail?CSNUMBER=34513.
[29] Available at http://www.loc.gov/marc/.
[30] Available at http://www.loc.gov/standards/mets/.
[31] Available at http://www.mnhs.org/preserve/records/metadatastandard.html.
[32] Available at http://www.records.nsw.gov.au/recordkeeping/nsw_recordkeeping_metadata_standard_4614.asp.
[33] Available at http://www.niso.org/standards/standard_detail.cfm?std_id=731.
[34] Available at http://www.loc.gov/standards/mix/.
[35] Available at http://ww.naa.gov.au/recordkeeping/gov_online/agls/metadata_element_set.html.
[36] Available at http://odrl.net/1.1/ODRL-11.pdf.
[37] Available at http://www.oclc.org/research/projects/pmwg/premis-final.pdf. Several InterPARES 2 researchers, in particular Victoria McCargar, were involved with the development of the PREMIS metadata set, which occurred concurrent with the work of InterPARES 2.
[38] Available at http://nedlib.kb.nl/results/NEDLIBmetadata.pdf.
[39] Available at http://www.npaci.edu/online/v6.2/perm.html.

capture to describe the identity, authenticity, content, context, structure and management requirements of records created in the context of a business activity;[40]

- *Recordkeeping Metadata Standard for Commonwealth Agencies*—describes the metadata that the National Archives of Australia recommends should be captured in the recordkeeping systems used by Commonwealth government agencies;[41]
- *South Australian Recordkeeping Metadata Standard (SARKMS)*—a technical standard outlining the basic core set of metadata elements required to capture and maintain recordkeeping metadata to assist in meeting the requirements of providing adequate records management;[42] and
- *Victorian Electronic Records Strategy (VERS) Metadata Schema*—defines metadata elements for a system that will capture, archive, manage and provide access to reliable and authentic electronic records.[43]

Schemas were initially selected based on the following processes and criteria:

- Schema documentation was reviewed and checked for relevance to recordkeeping and/or to see if it would be appropriate to analyze.
- Schemas that did not have sufficient documentation were removed from the list.
- Any schema that was listed as a "crosswalk" was removed since the system was not designed to analyze crosswalks (although the existence of crosswalks related to schemas was noted in the analysis).
- Because of time constraints, schemas that had a very large number of elements were given a lower priority.

The researchers decided that it would be too time-consuming to enter all of the elements of an individual schema manually. For some schemas (such as VERS) that have a large number of schemas *and* that have elements categorized according to a schema, the researchers decided that to enter just the name of the element container and specify the element that satisfied the condition in a note field. In future, an "import" function might be added to collect this data automatically from electronic versions of the schemas instead of having to do it all manually.

From this initial selection, schemas were prioritized based on their type. Schemas typed as those intended for recordkeeping purposes were given high priority. These generally were schemas for either local governments (for example, Minnesota) or for national recordkeeping purposes (for example, the Australian RKMS). Since there were such a high number of schemas for government recordkeeping, the researchers also tried to prioritize by sector. Schemas relating to InterPARES focus areas, such as the arts or geospatial applications, were given a higher priority than others.

One thing the researchers noticed during the prioritization process was that *all* of the schemas were from English-speaking countries, apart from one that was developed in China. It would be interesting to try to find more schemas developed by non-Anglo communities and try to analyze those. The researchers also noted that among the selected schemas, there was not a wide variety by domain/sector registered in the system. Schemas for the legal or medical fields were not represented, for example. It would be useful to get a sampling of these schemas for comparison, especially to see if other relevant metadata were revealed.

Other considerations included weighing what might be gained from analyzing schemas that were not developed specifically for recordkeeping purposes. How do they differ? Are the

[40] Available at http://www.imforumgi.gc.ca/documents/2001/meta/meta00_e.asp.
[41] Available at http://www.naa.gov.au/recordkeeping/control/rkms/contents.html.
[42] Available at http://www.archives.sa.gov.au/files/management_standard_metadata.pdf.
[43] Available at http://www.prov.vic.gov.au/vers/standard/.

schemas that were not designed for recordkeeping purposes all necessarily deficient when assessed against recordkeeping requirements? Again, might they include elements not previously considered that might be useful for recordkeeping purposes?

Testing, cross-validation and revision of the analytical framework (also referred to as the Schema and Analysis and Evaluation Instrument) were conducted by three different analysts who encoded selected archival schemas, some examples of key metadata schemas from related information fields (for example, Dublin Core) and schemas from scientific and artistic domains independently.

Findings about MADRAS tools and instruments

Upon proceeding with analysis of selected schemas, the researchers were somewhat surprised by the spotty nature of schema documentation. Since a schema is analyzed based upon its documentation, it is vital that this information be clear and concise, but often the researchers found it to be insufficient/deficient. Insufficient schema documentation led to the realization that the analysis questions needed to be refined to make sure that they were focused on eliciting responses about what a given schema is intended to do as opposed what that schema "can be made" to do. This in turn led to the realization that very few schemas can be analyzed accurately independent of their implementation.

Although it was agreed that the analysis undertaken within MADRAS should proceed only to the element level, while actually answering the analysis questions, the researchers found that they spent 85% of their time pouring over the definitions of sub-elements. Although this does not necessarily suggest taking analysis down to the sub-element level as a rule, it must be acknowledged that the real meat of a metadata schema does not tend to live at the element level, especially when one is being asked to describe records in the intricate manner proposed by the InterPARES and ISO 23081 instruments.

Because the language used in ISO 23081 and the InterPARES benchmark and baseline requirements differs, it was a challenge to clarify the meaning and intention in each of the documents and then to unify them. This proved to be difficult because the focus of the instruments is quite different. InterPARES focuses on *domain-independent* digital records, while the ISO standard focuses on records in all media made in the *course of business*. In addition, there are times when InterPARES and ISO 23081display such different approaches to a particular recordkeeping problem that the MADRAS analysis questions—in trying to satisfy both "masters"—become confusing. For example: in addressing MADRAS Question 206, "Chronological Date" as opposed to "Creation Date," InterPARES, drawing upon its diplomatics lineage, lists four date types in benchmark requirement A.1.a.iii: "Chronological," "Received," "Archival" and "Transmission." ISO documentation is concerned only with "Creation." Therefore, the picklist for this question, which has to combine the language from both sets of requirements, demonstrates how the combination of two different instruments can cause confusion. In this case, the differences in approach to dates appears to spring from the fact that the InterPARES requirements only admit those dates to which the record keeper can directly attest (i.e., one can identify the date written on a document (Chronological Date) but cannot actually be sure that this was its *creation* date), while ISO appears to believe that the record keeper will be able to identify an authentic creation date.

At other times, ISO 23081 seems overly vague:

> **Example 1**: MADRAS Questions 214 & 215: "Technical characteristics and dependencies of a record" v. "Technical requirements to render or reproduce record"
>
> The ISO documentation makes this distinction, but does not fully explain what makes one different from the other. The researchers assume that "characteristics and dependencies" is mainly about format, while "requirements to render or reproduce" is more about the entire technical environment needed, but it is unclear.

> **Example 2**: Questions 504 & 507: "Rules that regulate record management" v. "Rules that regulate records management operations"
>
> The ISO Standard is ambiguous. 9.3.1b (which stands behind question 504) states, "capture the business rules or other system controls that regulate record creation and management," while 9.3.1d (which stands behind 506) states, "capture the business rules or other system controls that regulate records management operations." How does the "record creation and management" from question 9.3.1b differ from the "records management operations" of 9.3.1d? The researchers assume that 9.3.1b is about creation, access and use while 9.3.1d is about activities performed only by records managers, such as preservation actions. Furthermore, since these instruments also largely directed how the researchers crafted the system, some of the concepts in ISO 23081 posed particular challenges. The standard describes that the researchers need to capture information "at record capture" and "after record capture." This is not a distinction made in the InterPARES requirements. To incorporate the concept into the analysis tool, the researchers considered metadata about a record's "content, context and structure" to be the metadata created "at record capture." Any other metadata that the researchers describe are, thus, by definition "after record capture." This amounts to isolating the metadata that deal directly with recordkeeping/administration, which appears to be in the spirit of ISO 23081.

Findings about the schemas

As mentioned above, the instantiations provided an interesting commentary on the status of metadata schema publication and documentation practices. It raised issues about persistent identification (for example, stability of URLs for schema documentation), standards for schema documentation and standards for their description addressing lack of and inconsistency in metadata to describe schema documentation.

As noted earlier, analyzing every schema identified as relevant was beyond the scope of this Project. However, the researchers did analyze enough from different sectors and of different types to be able to make the following observations:

- Almost no schema analyzed, with the exception of New South Wales, met all the requirements that were relevant to the schema's stated scope. In general, those schemas that are not designed for recordkeeping prove to be less compliant than the others. It is also often the case that the schemas—no matter the domain—fall short in being able to describe how a record/agent/mandate/business process changes "over time."[44]

[44] Schemas designed for managing geomatics data may provide an exception to this general observation, since time is a key element of any geographic feature.

- Some schemas were never intended to satisfy the kinds of requirements identified in the analytical framework, but nevertheless address some of them.
- Many record creation or preservation implementations may need to employ more than one schema simultaneously or sequentially to document all relevant aspects of their activities (this is even more likely to be the case where a records continuum approach is being used).
- Even if a schema were to meet all the requirements, this is unlikely to be the case in specific implementations/application profiles. The process of completing these selective analyses has demonstrated that many metadata schemas cannot effectively be separated from their implementation. Since it was decided that implementation issues would not be considered during analysis, many schemas appear to fall short in certain areas, and one might even fairly say that some of the analysis questions are poorly answered because of this distinction. For example, the ANZLIC standard (Metadata for Spatial Data Directories in Australia and New Zealand) requires (and the eGMS suggests) an implementation concomitant with the schema that notes which encoding schemas are being used within the implemented XML/HTML tags, not within the metadata elements proper. It must be remembered, however, that not only do existing metadata schemas predominantly not meet the necessary recordkeeping requirements, but actual implementations of specific metadata schemas often only use selective metadata elements and often not in standard ways.

The researchers have also identified that there are two major element/sub-element relationships:

- For a number of schemas (for example, the RKMS/Minnesota group and CDWA), the upper-level elements are only "envelopes" for a series of sub-elements. That is to say, the elements take no data values themselves, but serve as a type of header for the sub-elements, and it is these sub-elements that are actually assigned data values.
- For others (such as eGMS), the elements do take data values, and the sub-elements are actually "refinements" to those values.

Another finding is that some tools, especially those outside the more traditional recordkeeping/archival domain, do not fall neatly into some of MADRAS' classifications. How can the researchers modify MADRAS to account for this?

> **Example:** CEDARS Preservation Metadata
>
> Element obligation value is not designated as "Mandatory," "Optional" or "Conditional." Rather, the coding is based on the level of specificity indicated by the element (i.e., the extent to which it may be usefully applied across a wide range of digital materials). Values used in coding include "less significant," "very significant" and "significant."

In the above example, therefore, the element coding is assigned based on the types of objects rather than on the function/purpose of the metadata. So what does this mean? It means that it is difficult to compare element obligation encoding values between schemas since the reason the coding is being applied may differ from schema to schema. In other words, the researchers would be comparing apples to oranges. Moreover, the "significance" value is a subjective coding.

Because the MADRAS questions are so heavily weighted towards business process-specific recordkeeping issues, some non-recordkeeping schemas are not fully appreciated for what they can do. Not surprisingly, and perhaps also not a problem for the purpose of MADRAS, the analytical tool has difficulty evaluating aspects of a metadata schema that address aspects such as depth of description or monetary value that are emphasized by schemas in non-recordkeeping

domains (for example, CDWA and ANZLIC). Related to this, as might be expected, granularity of content description required to meet the user needs of those specific communities tends to be higher in non-recordkeeping schemas, while the recordkeeping schemas focus more on context description. Finally, it is noted that some non-archival specialists who register their discipline-specific schemas into MADRAS may have difficulty in doing so, as the MADRAS tool is specifically designed to meet the needs of archivists and is therefore expressed using archival terminology that may be unfamiliar to practitioners in other disciplines.[45]

To discuss these issues a little further, ANZLIC is an example of a complex metadata schema that was not designed specifically for recordkeeping purposes; it is all about describing the *content* of a dataset accurately, and, insofar as this is the case, is more or less doomed to perform poorly in a MADRAS analysis that privileges tracking contextual information over time. Of the thirty-six elements analyzed in MADRAS, twelve of them are used to describe the contact information of the custodian of, or source contact for, the dataset. Ten other elements describe the contents of the dataset (for example, its date, physical location, extent and keywords). Beyond content description, seven elements are used to describe the reliability and quality of the dataset (for example, lineage, positional accuracy and update frequency) and three elements describe access information including both format and rights. The remaining four elements consist of the title, a unique identifier, a date for the metadata record and a spot for "anything else." There is nothing in this schema that allows the user to see how this dataset has been used or housed over time.

On a final note, it seems almost impossible for any single-object schema to measure up to ISO 23081's requirement that a recordkeeping system not only track which mandates/agents/business processes are related to which record, but also track the set of mandates/agents/business processes themselves. In fact, what ISO is describing is the complete recordkeeping system, but most schemas are just meant for the record-centric portion of that system. Ultimately, this would be an implementation issue, because most metadata schemas do not assume that they are the only schema being used in a system. One way to address the issue might be to track the mandates separately, manually inserting the appropriate code or link within the system using the schema at hand.

MADRAS products

MADRAS, as an automated tool that facilitates schema analysis as well as serves as a registry of existing and evolving schemas; the analytical framework as a stand-alone tool that is to be incorporated into ISO 23081 but that can be used independent of both MADRAS and the ISO standard to assess current and draft schemas and application profiles; and the evaluative reports on the schemas analyzed by InterPARES researchers all constitute products of this research.

For each schema or set registered, a set of evaluative reports can be generated that: (a) indicate whether the schema meets all, some or none of the InterPARES benchmark and baseline requirements or ISO 23081 metadata requirements (recognizing that users may be interested in addressing either or both sets of requirements), (b) pinpoint in what ways, if any, the schema falls short and (c) provide guidance as to how the schema could be modified or augmented to meet all the relevant requirements.

[45] This was the case, for example, when a geomatics student worked with researchers at UCLA during the Excel spreadsheet development phase of the tool. The student was a metadata expert in the field of geomatics but required very intensive support from the UCLA researchers to fill in the required fields. In the end, a UCLA researcher had to register the schema and results from that process are pending.

One additional product that is still in process is the doctoral dissertation of Lori Lindberg, which is examining the implications of this analysis for ISAD(G), ISAAR, EAD and EAC and making specific recommendations for extensions to those descriptive standards and establishment of a new framework within which future development should take place.

Warrant Database

Scope and rationale

Description Cross-domain researchers made a decision early in the InterPARES 2 Project that developing an entire new metadata schema to address InterPARES requirements was neither practical nor likely to be adopted either within the recordkeeping and archives community or those communities within the various focus areas of the Project. There were several factors behind this decision: the difficulty in developing an all-encompassing schema that would work in so many different settings, issues of how to ensure that the schema would be able to continue to evolve after the end of the Project and difficulties in persuading communities (including archival communities) that had already invested in their own metadata frameworks, to adopt one developed by InterPARES. Instead, it was decided that the researchers would develop a way of assessing those schemas already developed by different communities against InterPARES requirements and provide them with feedback about how they could be extended or modified to address recordkeeping issues. The researchers then discussed how they could develop persuasive arguments that might lead those communities to respond to the Description Cross-domain's recommendations. The researchers decided that they needed to understand better what the communities were already saying about metadata and associated issues such as trust, reliability, authenticity, status as original, accuracy, ownership and custodianship, moral rights and preservation; which individuals were regarded as authoritative on these issues; and to what internal or external mandates they might likely respond. Armed with this knowledge, the researchers felt that they would be in a position to address the relevant communities in terms of their own concerns and mandates, if they existed, rather than appearing to impose InterPARES' upon them.

The literary warrant database was built using the method developed by Wendy Duff as part of the Pittsburgh Electronic Records Project. This involved identifying a warrant for a particular course of action based upon such things as legal or other juridical mandate, professional best practices, professional literature and other social sources.[46] In this case, the researchers were particularly interested in identifying literature and other sources that discussed the need for the creation and continued maintenance of description and other metadata supporting the accuracy, reliability, authenticity and preservation of records and other record-like objects.

Working with input from researchers from other InterPARES groups, the researchers conducted a literature review across each focus area to identify how different communities currently perceive and discuss the need for, and role of, metadata in ensuring the creation and preservation of reliable and authentic materials. The researchers designed and set up the Web-

[46] See Wendy M. Duff, "The Influence of Warrant on the Acceptance and Credibility of the Functional Requirements for Recordkeeping" (Ph.D. dissertation, University of Pittsburgh,1996); Wendy M. Duff (1997), "Warrant and the Definition of Electronic Records: Questions Arising from the Pittsburgh Project," *Archives and Museum Informatics* 11(3-4): 223–231, Available at http://tort.library.utoronto.ca:8080/bitstream/1778/4315/1/Warrent.pdf; and Wendy M. Duff (1997), "Compiling Warrant in Support of the Functional Requirements for Recordkeeping," *Bulletin of the American Society for Information Science* 23(5): 12–13. Available at http://doi.wiley.com/10.1002/bult.60.

based database to capture standardized literary warrant analyses. The software chosen allowed researchers to input remotely into a single database, but little effort was spent on developing a public interface since initially the tool was developed solely to support the researchers. Guidelines were developed for using database and analyzing warrant, and researchers from the Description and other InterPARES groups were trained in their use so that they could input materials they encountered during their research activities. Description Cross-domain researchers then analyzed materials for which records had been in the database, thus populating the database. In 2005, it was decided that the warrant analysis database might be a useful product for the public also, and the data it contained was transferred from UCLA to the University of British Columbia and loaded into a new database with a public interface.[47]

Results of warrant analysis

The database now contains 177 records that include not only bibliographic information but also summaries of the major arguments used in support of metadata concerns within different communities that can be referenced when developing presentations, publications and other InterPARES 2 products aimed at those communities. In this database are identified the warrants that fed into the analysis framework and that, additionally, have influenced each other.

News Archives Survey

Although a series of diverse case studies were conducted by InterPARES focus groups that included the gathering of data about metadata on behalf of the Description Cross-domain (discussed below), the Description Cross-domain was presented in 2005 with a unique opportunity to study contemporary thought and practice in a professional area that has changed both rapidly and radically with the development of online interactive, multimedia technologies—the news industry and its archives. Researchers decided that a survey of perceptions and practices relating to metadata in this industry would provide important insight into how one specific community is addressing metadata and preservation issues more broadly.

Conducting the survey

In recent years there has been a growing awareness that historic news archives in electronic formats are at risk.[48] In the popular media, printed newspapers are frequently described as a threatened species in the digital world, and Wall Street has responded accordingly by undervaluing media properties across the board. Efficiencies gained through automation have wiped out traditional "morgues" with their paper clippings and film negatives, and there are fewer archivists to tend to their born-digital avatars. Even microfilm, that reliable, long-lived preservation medium, is under serious threat from publishers who no longer see the need for it amid a nightly river of page PDFs extracted from sophisticated pagination systems.[49]

[47] Available at http://www.interpares.org/ip2/ip2_warrant_db.cfm.
[48] Victoria McCargar (2005), "Following the Trail of the Disappearing Data," *The Seybold Report* 4(21): 7–14.
[49] Bernard F. Reilly, Jr., "Knowledge Biodiversity: The Perilous Economic of World News Heritage Materials," in *Proceedings of the ACRL Twelfth National Conference, April 7-10, 2005, Minneapolis, MN* (Minneapolis, MN: Association of College and Research Libraries, 2005), 238–243. Available at http://www.ala.org/ala/acrl/acrlevents/reilly05.pdf.

In spite of the myriads of information channels available in the Digital Age, newspapers are still cited by historians as the most often used and most important resources in their research.[50] But even as the Library of Congress, with its National Digital Newspaper Project, pursues filming and digitizing 19th century editions, tomorrow night's all-digital output is every bit as threatened as a crumbling volume of newsprint, because the industry and profession are unprepared to handle it. Moreover, news is increasingly being created and transmitted to the newspapers from reporters in the field using online transmission of digital text, photographs and video.

Victoria McCargar, an InterPARES researcher and leading authority on electronic news archives, with the assistance of Shannon Supple, at the time a graduate researcher at UCLA, created a survey instrument to benchmark current trends in digital preservation among news archivists.[51] After receiving the appropriate permissions for human-subjects testing through UCLA, the survey was uploaded to a professional interface at the SurveyMonkey Web site in August, 2005. The invitation to participate in the survey was communicated through a popular and very active listserv mounted by the News Division of the Special Libraries Association, which numbers more than 650 news librarians and archivists. The survey was available to participants through the end of October, 2005.

The survey consisted of eighty questions divided into the following categories:

- Institutional environment
- Professionalism
- Budget
- Use of archives
- Policy
- Technology
- Metadata
- Digital preservation
- Copyright

Additional sections allowed for comments and for survey-takers to volunteer contact information if they were willing to participate in follow-up data-gathering. The survey instrument was designed in its initial questions to discover areas in common among organizations, such as which departments have responsibility for archival systems and how archival systems are budgeted. Later questions homed in on issues specific to digital preservation.

Data analysis was begun in February with the goal of making a "first-cut" presentation at the Special Library Association's 2006 annual conference in Baltimore.[52]

Discussion of the survey results

Despite the advances in digital preservation research in the last ten years, there is still a remarkably low level of awareness of the risks to cultural heritage material in the private sector, which falls outside the domains of academic libraries, archives and government. One of the

[50] Helen R. Tibbo (2003), "Primarily History in America: How U.S. Historians Search for Primary Materials at the Dawn of the Digital Age," *American Archivist* 66(1): 9–50.
[51] This "News Archive Survey Instrument" is available on the InterPARES Web site and on the DVD accompanying this book.
[52] See Victoria McCargar (2006), "You Can Kiss Your Assets Goodbye: The State of News Archives at the Dawn of the Digital Era," paper presented at the Special Libraries Association Annual Conference 2006, June 10-15, 2006, Baltimore, MD, USA. PowerPoint slides available at http://www.ibiblio.org/slanews/conferences/sla2006/presentations/assets.pdf. Audio available at http://www.ibiblio.org/slanews/conferences/sla2006/audio/vicky.mp3.

challenges in mounting a preservation survey of news archives was the lack of basic understanding of the issues among potential participants. The goal of the survey instrument was to capture as much data as the researchers could from each participant before she or he reached questions that could not be answered without a fuller understanding of the complexities of these issues. In fact, of the seventy-seven participants who started the survey, only twenty-eight—fewer than half—completed it. Those who did, however, helped paint a picture of a great volume of historic, cultural heritage material at risk.

A somewhat more subtle function of the survey was to try to educate survey-takers about digital preservation on a basic level. The question, "How knowledgeable is your staff about digital preservation?" revealed a low level of understanding; 55% of respondents answered "Low," and almost a fourth stated they had no idea what level of understanding prevailed. Only 15% indicated they had some knowledge, and only two respondents indicated that they had a "high" level of understanding. Questions like this are useful for establishing a benchmark for gauging increasing awareness.

One of the most interesting—and unsettling—questions addressed instances of actual loss: "In any of your previous preservation activities (including upgrading software, moving to a new storage medium, moving to a new software product), did you experience any loss of data or metadata, or otherwise compromise the archives?" Of the twenty-eight responses, only five answered that they had not. Twenty-one of the remaining twenty-three reported some form of loss, ranging from minor (a few corrupt images on CD-ROMs) to the serious (the loss of controlled vocabulary terms for certain objects) to the disastrous (loss of an entire collection of thousands of photographs). The two instances of "don't know" were telling insofar as they point to an archives environment where data validation is not routine. Indeed, these instances of loss seem to have been uncovered by accident, in the course of a system upgrade or on the fly. If losses are not detected quickly, the chance of retrieving an intact original from backup is lost.[53] Moreover, this lack of routine bit-level validation has implications for data authenticity even in the short term, as will be noted below.

Some of the other results of interest were:

- A low level of commitment by management to archival policy. Only 33% of responding newspapers enjoyed "very committed" oversight. In a future survey it would be worthwhile to explore the extent to which this is a result of revenue interests (mounting the Web sites via archival data feeds) or a commitment to preservation for its own sake.
- The concept of *authenticity* in the digital environment is still rooted in the old model of microfilm as juridical version. To the extent to which news archivists answered that authenticity was a consideration in their archives—about half indicated that it was "important" or "very important—*authenticity* refers to how closely the material in the database reflects what was printed on paper. Bit-level authentication of individual files in the digital preservation sense is an unknown concept. Saying that, larger newspapers do recognize the legal implications of having an "authentic" representation of a printed article or photograph, and some, such as the *Atlanta Journal-Constitution*, have a notary public on the newsroom staff who can validate printed copies from microfilm to fulfil a legal request, either one arising from the newspaper's own activities or those between third parties.[54]

[53] Victoria McCargar (2006), "The Heart of Darkness: A Foray into Aging JPEGs," *The Seybold Report* 5(22): 9–12.
[54] Personal conversation between Victoria McCargar and Virginia Everett, news director of the *Atlanta Journal-Constitution,* May 9, 2006, in Atlanta, GA, USA.

- A lack of dedicated funding. About 20% of responding news libraries indicated that they had a budget earmarked specifically for preservation, and another 10% had a separate preservation budget. However, it is highly unlikely that this funding factors in digital preservation; it is almost certainly dedicated to *digitization* projects to unlock the commercial value of historic photography, and, ironically, sets up a new preservation problem for the collection of newly scanned JPEGs.
- A lack of control over the technology environments in which news archivists operate. Only 13% stated that the archivists were responsible for software and 5% for hardware support. In both cases, the responsibility fell to the information technology department and/or the newspaper's vendors. In some instances the photography department was the responsible group.[55] All of these point to a situation where those best equipped to deal with digital preservation—information professionals—are not the major stakeholders in the archives.
- Metadata standards are soft or nonexistent. The reigning schema, IPTC, is widely used (it is the basis for most commercial systems), but of the 58% of respondents who said they use it, up to two-thirds reported that the schema is "somewhat to highly customized" in their archives. The remaining respondents indicated no standard schema or did not know whether one was in place. Schemas associated with digital preservation like PREMIS and MIX (and their envelope METS) are unknown in news libraries.
- There is a proliferation of file formats such as digital video, information graphics, GPS databases and the Web pages in many of the archives as the impact of multimedia publishing matures. However, few controls are in place. More than three-quarters (79%) of responding news libraries reported no policy for handling digital materials over the long term. Of the 21% that have such a policy, only 12% attempted to address problematic, fragile formats, and none of the archivists reported regular reviews to address technological change.
- Similarly, most newspapers do not attempt to capture metadata about these formats, which is considered critically important information in the PREMIS schema. Fewer than 40% of survey respondents indicated that they attempt to catalogue hardware and software metadata in their archives, while only 15% record the operating system and 7% record the necessary peripherals even though all of these elements are specified in PREMIS.[56] These numbers cannot be extrapolated across all news archives since only twenty-six respondents of the original seventy-seven were still participating at this point in the survey and probably represent just the small portion of the community that actually understand digital preservation issues.
- The one area of digital preservation metadata where newspapers are arguably quite thorough is copyright. The U.S. Supreme Court decision in *Tasini*[57] led to the removal of entire sections of many publications, and, in the interim, most papers have better controls in place to identify authorship, ownership and certain aspects of provenance. However, news archivists are much less informed about legal issues relating to preservation of copyrighted material in their digital archives, including reformatting, migrating or

[55] Photographers' archiving practices are highly idiosyncratic; see Jessica Bushey and Marta Braun (2006), "InterPARES 2 Project - General Study 07 Final Report: Survey of Recordkeeping Practices of Photographers using Digital Technology." Available at http://www.interpares.org/display_file.cfm?doc=ip2_gs07_final_report.pdf.

[56] McCargar was a member of the PREMIS Working Group in 2004-05 and catalogued a typical newspaper complex/compound object using the draft schema: http://www.oclc.org/research/projects/pmwg/premis-examples.pdf.

[57] *New York Times Co., Inc., et al. v. Tasini et al.* (00-201) 533 U.S. 483 (2001) 206 F.3d 161, affirmed. Available at http://supct.law.cornell.edu/supct/html/00-201.ZS.html.

normalization. 60% of respondents answered "don't know" when asked about what actions they are legally allowed to take. The remaining respondents who did indicate an awareness of legal issues, were, in many cases, misinformed. Working in units of for-profit institutions, news archivists face proscriptions on preservation activities that are not encountered by nonprofit and public repositories; this is an evolving situation as the Library of Congress tackles revisions to the Digital Millennium Copyright Act, the so-called Section 108 Study Group.[58]

Conclusions

In aggregate, the data describes a wealth of historic material in risky, proprietary formats and an important segment of the archivist profession that is ill-equipped to handle them.

Measuring awareness and institutional change over the next few years is important to understanding whether news properties, left to their own devices, will be capable of sustaining this content into the future. News librarians and archivists—practitioners often wear both hats—are well aware that they are responsible for their publications' writing of daily history. The opportunity to comment at the end of the survey questions afforded a few participants a chance to vent their frustration: "The researchers are so busy creating digital archives the researchers are not paying attention to the problems the researchers will leave behind," and "The archival aspect of a newsroom library is often considered an ancillary function of the newsgathering operation, not a key strategic priority for the company." Newspapers, increasingly pressed to boost revenue as advertising shrinks, have hard priorities that may not coincide with preservation; as one survey respondent put it, "In pursuit of the bottom line, management seems to feel that it is more important to spend money than getting the paper out today than it is to archive for the future."

Benchmarking news archives at this juncture will help digital preservationists monitor what might be identified as an impending crisis. But those hoping for solutions to arrive from stronger standards and best practices may be in for a long wait; pursuing a third-party repository model may be a more promising avenue.[59]

Metadata Specification Model

The premise underlying the work of the Description Cross-domain is that detailed trustworthy metadata are key to ensuring the creation of reliable, and preservation of authentic, records and other entities in electronic systems. This argues for an end-to-end metadata management regime that addresses which metadata need to be created and/or carried forward in time, for what purposes, by whom, and how they are to be preserved and validated. Bound up with this, however, are difficult issues associated with how to create rich metadata in a resource-efficient manner as well as how to manage and continue to ensure the trustworthiness of the volume of metadata one ends up accumulating over time (including metadata associated with the preservation, reproduction and dissemination aspects of the archival function). This raises interesting questions such as whether certain metadata can be efficiently segregated and

[58] See http://www.loc.gov/section108/. McCargar contributed a public comment on behalf of news archivists (see Victoria McCargar and Peter F. Johnson (2006), "Comments to Section 108 Study Group: News Archives," submitted April 28, 2006. Available at http://www.loc.gov/section108/docs/McCargar.pdf).

[59] McCargar is consulting on a project to develop an audit instrument for a trusted news repository at the Center for Research Libraries; for a brief overview, see Center for Research Libraries, "Auditing and Certification of Digital Archives." Available at http://www.crl.edu/content.asp?l1=13&l2=58&l3=162.

eliminated after validation, certification and summarization by a preserver. Without addressing this question, preservers will ultimately end up managing more metadata than the entities to which they refer.

One goal of the metadata specification model was to identify an overall set of metadata requirements that specify what metadata need to be created, from which sources, how and by whom, at which points within both the Chain of Preservation (lifecycle) and the Business-driven Recordkeeping (records continuum) models being developed by the IP2 Modeling Cross-domain and retention periods for such metadata. This metadata specification model could then form the basis for developing specifications for automated tools that can be used to assist with the creation, capture, management and preservation of essential metadata for active and preserved records. A second goal was to develop an economical and consistent way of talking about different classes of metadata to facilitate systems design, task allocation and management, as well as automated metadata creation.

Actions taken and products created

Description Cross-domain researchers had to wait until work was sufficiently advanced on the InterPARES 2 activity models to begin work on the development of metadata specification models for the Chain of Preservation and Business-driven Recordkeeping models. Because the former was the more complete toward the end of the Project, the researchers were able to develop a metadata specification model for it.[60] In the metadata specification model for the Chain of Preservation model, the following definition was used for "metadata:" *a machine or human-readable assertion about a resource relating to records and their resources*. Descriptive metadata are defined as those categories of metadata carried forward to be used as evidence for archival description. One hundred thirty-seven (137) different metadata assertions were identified (i.e., different instances of types of metadata), sixteen types of assertions were identified. Two cut across all stages of the lifecycle, one cut across two stages, and the other fifteen were evidenced only in one stage. The resulting model is still a theoretical model that is awaiting validation through instantiation—both through walkthroughs based on the case studies conducted by the InterPARES 2 focus groups[61] of specific implementations and by actual system building. When researchers start to work on the development of the metadata specification model for the Business-driven Recordkeeping model, it is anticipated that the researchers will encounter some of the same issues as were encountered in developing MADRAS in that the records continuum has a very different set of entity foci to the records-centric notion underlying the lifecycle. Other work that is continuing includes the development of attribute pairs for the metadata identified in these models which would designate the values different assertions should take; the development of a typology of classes or categories of metadata and, potentially, the mapping of both metadata specification models onto the OAIS model.

[60] For a general overview of the metadata elements identified in relation to record-making, recordkeeping and preservation activities, see the narrative discussion of the Chain of Preservation Model in the Modeling Cross-domain Task Force Report. For a more detailed description of the Chain of Preservation metadata specification model, see Joseph T. Tennis (2008), "Metadata in the Chain of Preservation Model," *Archivaria* (in press).

[61] For discussions of two walkthroughs done of earlier drafts of the Chain of Preservation Model, see William Underwood, Kevin Glick and Mark Wolfe (2007), "InterPARES 2 Project - General Study 12 Final Report: Validation of the InterPARES 2 Project Chain of Preservation Model Using Case Study Data." Available at http://www.interpares.org/display_file.cfm?doc=ip2_gs12_final_report.pdf; and Randy Preston (2004), "InterPARES 2 Project - Modeling Cross-domain: Walkthrough of the Manage Chain of Preservation Model Using Case Study 14 Data," draft report. Available at http://www.interpares.org/display_file.cfm?doc=ip2_cs14_COP_model_walkthrough.pdf.

Case and General Studies Data Analysis

Actions taken

One major component of the work of InterPARES 2 was a series of case and general studies undertaken by the arts, science and government focus task forces, examining dynamic, interactive and experiential environments in each focus.[62] These studies examined many facets of these environments and included several questions in their protocols that potentially addressed issues of concern to the Description Cross-domain:

- How are the digital entities identified (e.g., is there a [persistent] unique identifier)?
- From what application do the record system(s) inherit or capture all digital entities and the related metadata (e.g., e-mail, tracking systems, workflow system, office system, databases, etc.)?
- Does the recordkeeping system provide ready access to all relevant digital entities and related metadata?
- Does the recordkeeping system document all actions/transactions that take place in the system re: the digital entities? If so, what are the metadata captured?
- What descriptive or other metadata schemas or standards are currently being used in the creation, maintenance and use of the recordkeeping system or environment being studied?
- What is the source of these metadata (institutional convention, professional body, international standard, individual practice, etc.)?

Description Cross-domain researchers recognized that metadata issues could also surface in a more general manner in the course of the case study.

Case study data analysis

The Description Cross-domain researchers sought to identify, through the answers to the above metadata-specific questions and the data collected overall in the case studies, which, if any, metadata schemas and sets were currently being implemented; whether these schemas and sets were home-grown for this particular creator, required by the software implementation used, native to the creator's sector or discipline and/or a recognized industry or national/international standards; whether or not any metadata used addressed recordkeeping concerns, either fully or in part; and the extent to which real-world implementation of metadata measured up to or surpassed the ideal of the metadata requirements delineated in the Analytical Framework. The following discussion is drawn from the data and reports generated by the case study researchers.[63]

In the arts focus, where it is usually the product of the artistic activity that is the object of concern rather than a record that is the by-product of that activity, only two case studies uncovered use of metadata standards, and none of these were standards developed specifically for recordkeeping, archival or preservation functions. Case study 09(03),[64] the Commercial Film Studio component of the multi-component Digital Moving Images case study, cites use of several common bibliographic description and resource discovery metadata schemas—

[62] For a detailed synopsis of the InterPARES 2 case studies, see the section in the Domain 1 Task Force Report titled "Characterization of the Case Studies." Available at http://www.interpares.org/display_file.cfm?doc=ip2_book_part_2_domain1_task_force.pdf.
[63] See Appendix 18 for a comprehensive summary of the case study data relating to metadata in each of the focus groups.
[64] See James Turner, et al. (2004), "InterPARES 2 Project - Case Study 09(3) Final Report: Digital Moving Images - Commercial Film Studio." Available at http://www.interpares.org/display_file.cfm?doc=ip2_cs09-3_final_report.pdf.

Categories for the Description of Works of Art (CDWA), the Dublin Core (DC), the Thesaurus for Graphic Materials I: Subject Terms (TGMI), the Thesaurus for Graphic Materials II: Genre and Physical Characteristics Terms (TGM II) and the Anglo-American Cataloguing Rules, (AACR). Case study 09(04) (WGBH)[65] also cites use of Dublin Core and LCSH, as well as an industry schema, the Public Broadcasting Core (PBCore). Case study 03, the *HorizonZero* case study,[66] uses the CanCore standard, which is derived from the Dublin Core metadata set and is based on and fully compatible with the IEEE Learning Object Metadata (LOM) standard and the IMS Learning Resource Metadata specification. In terms of overall metadata implementation, none of the arts focus case studies indicated conscious attempts to apply metadata, beyond a few efforts to establish file naming conventions, largely for retrieval purposes, some version control and, in some cases, rudimentary tracking of file check-in or -out or file archiving.

In the science focus, records need to be not only reliable and authentic, but *accurate*. Data quality parameters are essential in the sciences. Relevant and sufficient metadata, therefore, need to be created to document data lineage,[67] especially in situations where datasets, models, software applications, datasets or multimedia objects have been acquired from elsewhere, which is often the case particularly in data portals. Scientists will not trust data they access from other sources without metadata that clearly indicate the reliability, authenticity and accuracy of those data.

Several science focus case studies exhibit the use of metadata schemas that, although not originating in the domains of recordkeeping, preservation or archives, nevertheless do address to varying degrees the requirements for the long-term management and preservation of authentic digital entities. Although this is encouraging, it does, in so far as each scientific area continues to define its own metadata standards, nevertheless raise concerns about interoperability and extensibility issues related to collaboration and recordkeeping.

Case study 08, the NASA case study,[68] applies naming conventions and incorporates the metadata elements contained in the Planetary Science Data Dictionary, which defines rules for constructing Data Element and Data Object names within the Planetary Data System (PDS), which are NASA institutional and data type specific Planetary Science Metadata. The case study also includes metadata that are associated with a data product (mainly relating to data processing, although there is a required version element and there are also optional data type and description elements, including mission, instrument and instrument type). When restricted areas are accessed, the system logs the user ID, date, time and operation performed.

---

[65] See Mary Ide (2005), "InterPARES 2 Project - Case Study 09(4) Final Report: Digital Moving Images -WGBH Boston." Available at http://www.interpares.org/display_file.cfm?doc=ip2_cs09-4_final_report.pdf.

[66] See Brent Lee (2004), "InterPARES 2 Project - Case Study 03 Final Report: *HorizonZero*/Zero Horizon Online Magazine and Media Database." Available at http://www.interpares.org/display_file.cfm?doc=ip2_cs03_final_report.pdf.

[67] In the field of geomatics, lineage means the history of the dataset, the dataset's pedigree as it changes form, its lifecycle from collection to acquisition, through all the dataset's stages of conversion, correction and transformations, its parentage. Specifically lineage contains information that describes the source of the observations, data acquisition and compilation methodologies, conversions, transformations, analyses and derivations to which the data have been subjected, and the assumptions and criteria applied at any stage of their life as well as any biases. In fact, lineage is normally the first part of a quality statement since most other data quality elements are affected by lineage. Data producers have documented procedures and quality requirements they have to meet, and lineage is a kind of audit trail to attest to the fact that the producer has met their standards. For the user, lineage provides a dataset its pedigree, to decide on its fitness for use. The "ultimate purpose of lineage is to preserve for future generations the valuable historical data resource. The key to our understanding of the Earth system may lie in the data collected by past generations" (Derek G. Clarke and David M. Clark, "Chapter 2: Lineage," in *Elements of Spatial Data Quality*, Stephen C. Guptill and Joel L. Morrison, eds. (Oxford: Elsevier Science, 1995), 13–30). Lineage can also be found in a dataset's associated publications, reports, and technical notes (see Tracey P. Lauriault et al. (2007), "Today's Data are Part of Tomorrow's Research: Archival Issues in the Sciences," *Archivaria* 64 (Fall): 123–179).

[68] See William Underwood (2005), "InterPARES 2 Project - Case Study 08 Final Report: Mars Global Surveyor Data Records in the Planetary Data System." Available at http://www.interpares.org/display_file.cfm?doc=ip2_cs08_final_report.pdf.

Case study 14, the Archaeological Records in a Geographic Information System case study,[69] indicated the potential through the software used, ArcCatalogue, to create, manage and edit metadata in XML, based on the Federal Geographic Data Committee (FGDC) Content Standards for Digital Geospatial Metadata or the ISO 19115 Geographic Metadata Standard. Such metadata would indicate from what source (for example, publication, repository, Web site or database) the data were retrieved. However, the process for creating and maintaining the digital entities is ad hoc even though the Geographic Information System (GIS) dynamically links geospatial metadata and descriptive attribute data from a wide variety of sources. File naming conventions are used for digital entities and certain aggregations of files can take on an associative identity of their own. Time tagging of georeferenced information is part of the documentation of the processes of creating digital maps, models and georeferenced visualizations. No formal recordkeeping system external to the application being used is applied, and heavy reliance is placed on the creator in terms of getting access to the files.

Case study 19, Preservation and Authentication of Electronic Engineering and Manufacturing Records case study,[70] conducted by the U.S. National Archives and Records Administration and the San Diego Super Computer Center, examined an engineering experiment to test an XML-based archival format for digital model (CAD) records of machined piece-parts used in high-tolerance manufacturing. The intent of the experiment was to preserve not only the geometric specifications of the model but also its semantically encoded metadata, allowing for their examination by reasoning programs for authentication prior to operationalization in computer-aided manufacturing. The experiment used a domain-specific metadata schema, STEP, the Standard for the Exchange of Product Model Data, a comprehensive ISO standard (ISO 10303) that describes how to represent and exchange CAD digital model information. STEP was extended by adding to it metadata elements in the form of logical expressions that enable reasoning over the topological features of the solid model and its functional context. This logical format, including the data model, was transferred into what was termed a new logical preservation format using the OWL Web Ontology Language, an open-source, public domain XML specification of the World Wide Web Consortium (W3C). OWL is a semantic XML format formally recommended by W3C in 2004 for use as a language to represent machine interpretable content when the content needs to be processed by applications rather than just structured for presentation to humans.[71] The case study report outlines specific metadata schemas delineated in the ANSI Y 14.5 Dimensioning and Tolerancing standard, and the use of a metadata cataloguing system (MCAT). Corporate standards were also used for solid-model and drawing metadata.

Case study 26, the MOST Satellite Mission case study,[72] indicates that MOST researchers chose file formats based upon astronomical best practice and then the metadata created were derived from that file format. Digital entities are identified by unique names and an additional set of unique identifiers. The metadata refer to information such as orbital parameters, telemetry information and target image information. The report notes that some of the metadata fields in

[69] See Richard Pearce-Moses, Erin O'Meara and Randy Preston (2004), "InterPARES 2 Project - Case Study 14 Final Report: Archaeological Records in a Geographical Information System: Research in the American Southwest." Available at http://www.interpares.org/display_file.cfm?doc=ip2_cs14_final_report.pdf.

[70] See Kenneth Hawkins (2006), "InterPARES 2 Project - Case Study 19 Final Report: Preservation and Authentication of Electronic Engineering and Manufacturing Records." Available at http://www.interpares.org/display_file.cfm?doc=ip2_cs19_final_report.pdf.

[71] See http://www.w3.org/TR/owl-features/.

[72] See Bart Ballaux (2005), "InterPARES 2 Project - Case Study 26 Final Report: MOST Satellite Mission - Preservation of Space Telescope Data." Available at http://www.interpares.org/display_file.cfm?doc=ip2_cs26_final_report.pdf.

the FITS files are mandatory, due to the file format being used. In general, no metadata standards are used and the MOST researchers have created their own scheme of important descriptive fields that meet the needs of their particular research project. No formal capture system is in place and access is dependent upon the capabilities of Windows Explorer.

No ability to determine whether a file had been altered, how, when and by whom, such as an audit trail, was identified as being built into any system examined in the science focus case studies with the exception of case study 06, the Cybercartographic Atlas of Antarctica case study,[73] which contains a more detailed account of metadata considerations than any of the other science focus case studies. Case study 06 reports that an Authors' Toolkit will eventually allow changes to associated metadata to be tracked. Case study 06 also conforms to ISO 19115 Geomatics Standards and the case study outlines important metadata elements that should be present and where these should be located (although these metadata are primarily cartographic or relate to the nature or behaviour of multimedia objects contained in the Atlas). Because the Atlas acquires data and their associated metadata from other organizations, it incorporates the metadata that accompanies those datasets. Any digital object being incorporated by the Atlas has been peer reviewed and must be described by the creator using the project's metadata standards. Each object is assessed against the Elements of Spatial Data Quality, including lineage, positional accuracy, attribute thematic accuracy, completeness, logical inconsistency, semantic accuracy and temporal information. The case study reports that authenticity in geography is measured in standard metadata as data lineage. Quality measures are dependent on the type of data and their function (for example, the acceptable margin of error for the precise location and size of a particular ice flow to inform tourist ships is smaller than fish counts to inform fisheries and ecological modeling). Other metadata requirements that are followed include the FGDC and/or British Antarctic Survey Directory Interchange Format (DIF) and OGC interoperability specifications.

The Atlas also tracks provenance and rights metadata associated with multimedia objects that have been incorporated into its content, primarily through a citation, caption or link to a bibliography. Linkages between information objects, their functionality and their associated metadata within content modules are described within an XML document. However, there are no unique or persistent identifiers and there is no formal ID lookup system. All versions of any software code used are tracked using Subversion, a source repository system. Documents that accompany the case study include Elements of Geospatial Data Quality, a Multimedia Metadata Discussion document and a List of Standards Adhered to on the Project.[74]

In the government focus, case study 05, the Archives of Ontario Web Exhibits case study,[75] reports that:

> There is no one recordkeeping system for records generated in the creation of exhibits. Different contributors (most notably the curator, webmaster, scanning technician and manager) each create and maintain their own records of this process. The Web site component files exist on both the development and production servers only. Thus there is no common classification scheme or file naming convention.[76]

[73] See Tracey P. Lauriault and Yvette Hackett (2005), "InterPARES 2 Project - Case Study 06 Final Report: Cybercartographic Atlas of Antarctica." Available at http://www.interpares.org/display_file.cfm?doc=ip2_cs06_final_report.pdf.
[74] Additional discussion of metadata aspects of the Cybercartographic Atlas of Antarctica can be found in Zhou, "Profiling and Visualizing Metadata for Multimedia Information in a Geospatial Portal," op. cit.
[75] See Jim Suderman et al. (2004), "InterPARES 2 Project - Case Study 05 Final Report: Archives of Ontario Web Exhibits." Available at http://www.interpares.org/display_file.cfm?doc=ip2_cs05_final_report.pdf.
[76] Ibid., 11.

As a result, "[r]ecordkeeping throughout the creation process of a Web exhibit is ad hoc and at the discretion of the participating individuals."[77] Since there is no recordkeeping system for the exhibits themselves, "these Web sites and their contents need to be seen within the context of a corporate Web site which has some aspects of a recordkeeping system."[78] Maintenance of the exhibits primarily involves making revisions to them. However, changes made to the exhibits are not documented. Web logging software documents aspects of all interactions with the institution's Web site, but related metadata are not readily accessible, even if it has been captured.

Case study 18, the Alsace-Moselle Land Registry (AMALFI) case study,[79] examined how Web-based applications enable the creation and management of the ordinances and access to the content of the land registry. No metadata schemas or standards are used, and metadata are not discussed in the final case study report. However, every inscription in the database is numbered with a persistent, unique identifier and dated and there are naming conventions for the other entities. Ordinances are also numbered and dated. Each scanned image of the registers is numbered according to the system already in place for numbering individual pages of the registers. Each inscription in the land registry is also connected to a physical file, the annex, by means of a reference number.

The database aggregates the data according to the main categories: parcels, persons, rights and obligations. The database has been organized following a data model closely mapped on the organization of a single inscription (*feuillet*) within the paper register. That is, the main entity is the inscription, of which there is one for each landowner. Each inscription may hold multiple land parcels and multiple inscriptions within the administrative scope of a land registry office. That is, a single inscription contains information relative to all the properties of a single person within a given administrative territory (usually a commune or part of one).

The system keeps track of changes to the digital records. The relevant fields of the database are updated with the information contained in the ordinance once the latter is signed by a judge. Also, since each land parcel listed in the registry also references an entry in the cadastral survey, any change to the cadastral survey must first be reflected in the land registry. However, scanned images of the paper registry and digitally signed ordinances are never modified.

Case study 20, the Revenue On-line Service (ROS) of Ireland case study,[80] reports on a system whereby taxpayers can pay their taxes online. The system keeps track of changes, which are noted and logged with a time/date stamp and the name of the employee making the change. Metadata issues are only touched upon in the case study report, but appear to be addressed by several "in-house" concepts, including by the capture of what is known as the "security wrapper." The "security wrapper" is "the entire transaction dataset received from the customer by ROS. This includes the transaction element; i.e., tax return and payment instruction, as well as the 'security packaging' element; i.e., digital signature, date/time stamp etc."[81] Another possible use of metadata may be in data transfer, although this is unclear. The report does state that "metadata related to the expired certificates, *in addition to the security wrapper*, is maintained

[77] Ibid., 30.
[78] Ibid., 11.
[79] See Jean-François Blanchette, François Banat-Berger and Geneviève Shepherd (2004), "InterPARES 2 Project - Case Study 18 Final Report: Computerization of Alsace-Moselle's Land Registry." Available at http://www.interpares.org/display_file.cfm?doc=ip2_cs18_final_report.pdf.
[80] See John McDonough, Ken Hannigan and Tom Quinlan (2005), "InterPARES 2 Project - Case Study 20 Final Report: Revenue On-Line Service (ROS)." Available at http://www.interpares.org/display_file.cfm?doc=ip2_cs20_final_report.pdf.
[81] Ibid., 27.

within ROS."[82] Although an Irish Public Service Metadata standard exists, it is not used with ROS. Nonetheless, "[t]wenty-two schemas for the tax forms available via ROS are publicly available in XML-DTDs for inclusion in ROS-compatible software developed by third parties. Each schema includes a DTD and element definitions and explanations."[83] The standards for these schema are based on institutional practice. "From an operational perspective, form element selection and management are, in large measure, based on data flow and format requirements of the ITP and related back-end systems applications."[84]

Case study 21, the Singapore Supreme Court Electronic Filing System (EFS) case study,[85] reports that the Bankruptcy Section of the Supreme Court has created an internal procedure manual and workflow chart on the process of filing bankruptcy petitions in accordance with juridical requirements. In addition, the Bankruptcy Act (Commencement) Notification of 1995 details the necessary documentary forms of records related to bankruptcy proceedings. There is also a prescribed documentary template allowing law firms to enter information on their cases. A unique, persistent identifier—the file reference number—is assigned to each case. The digital certificates issued and managed by the system have a unique Certificate Control Number. The naming conventions of the records created under EFS are clearly stated under the Bankruptcy Rules and Act as well as in the registry's internal workflow. To organize records, there exists a uniform classification scheme comprising all Supreme Court cases. "The internal business processes and the juridical regulations laid down by the courts govern the organization of the digital entities of the EFS."[86] To make organization easier and more intuitive in the electronic system, "[t]he file classification of bankruptcy records in EFS mirrors its previous paper based filing system, with some modifications. In the traditional paper environment, the record profile of the case file comprised the case number and the name of the debtor. However, the EFS bankruptcy case file comprises not only [these elements], but also the name of petitioner, case status (pending or concluded) and the bankruptcy status (bankruptcy order, adjourned or withdrawn)."[87]

The term "metadata" is rarely used in the final EFS case study report. Instead, it is sometimes referred to as a "prescribed set of information" or, most frequently, a "documentary template," which acts "as the record profile." "The front-end module allows law firms to enter relevant metadata elements using a prescribed documentary template that are in HTML pages and to attach the corresponding supporting records, which are in PDF."[88] Metadata elements that the law firm must enter include the firm's file reference number; party details, including the party type (i.e., whether the firm is representing the creditor or debtor), the name of the parties, addresses of the parties and the name of the solicitor. "The fields of the documentary template are controlled, to ensure consistency and accuracy of information and this explains why there is a drop down menu for some of the data elements."[89] The schemas for the documentary templates are based on the workflow and juridical requirements of the court.

With regards to capturing the digital entities, it is assumed that when the final report speaks of submissions by law firms to the Court, that this process is equated with the EFS capturing the

[82] Ibid., 55. Emphasis added.
[83] Ibid., 68.
[84] Ibid., 69.
[85] See Elaine Goh (2005), "InterPARES 2 Project - Case Study 21 Final Report: The Electronic Filing System (EFS) of the Supreme Court of Singapore." Available at http://www.interpares.org/display_file.cfm?doc=ip2_cs21_final_report.pdf.
[86] Ibid., 19.
[87] Ibid., 10.
[88] Ibid., 20.
[89] Ibid., 30.

submissions/records in question. The final case study report notes that "the EFS captures both the metadata of the record and the actual record itself."[90] It is assumed that the "record" referred to here is the PDF document, while the metadata are what are entered via the Web-based application using what is referred to in the final report as a "prescribed documentary template"[91] coded in HTML.

There is a tracking function for actions/transactions made in the system. In addition, the EFS maintains a transaction log, a financial audit log and a violation log. The transaction log maintains all changes to the digital entities in the system such as changes to documentary templates and deletion of records and annotations. The financial audit log maintains changes made to the payment of fees made to the court, while the violation log keeps all changes to the digital entities in the system, such as changes to templates and deletion of documents. The violation log also keeps track of unsuccessful (and potentially malicious) attempts to use functions.

The Supreme Court Registry is responsible for the processing, registration and custody of records. The court's workflow and record keeping systems operate together using Visual Basic, Oracle database and FileNet document management systems. The court uses FileNet, a document management system that indexes and stores the PDF files sent by the law firms. "The court's application system manages all incoming submissions by the law firms as well as outgoing replies by the court."[92] The system includes the "record register," which is essentially an index of documents within the case file. In the traditional paper environment, the register includes the record profile of the various types of documents related to the case, the document number and date the documents were filed. In EFS, the record register exists in the form of a sub-directory. Compared to the paper based system, the EFS record register has an additional record profile: the originator of the document (the person who created the record).

In case study 24, the VanMap case study, which looks at a Web-based map system for the City of Vancouver, the HTML and CFML pages and embedded GIF images are identified by unique URLs.[93] The data fields, layers and groups are also identified by field names, layer names and group names, respectively. Metadata is assigned based on what the VanMap Team thinks would be most useful for users. Metadata generated automatically upon creation of the data have not yet been investigated. "Fortunately, the VanMap Web site includes data sheets listing, at varying levels of detail, the types of data, their origins and the means by which they are included in VanMap."[94] The homepage of the staff edition Web site includes links to the data sheets. The data sheets, which can also be reached from the VanMap toolbar, contain information about layers, layer groups, reports and functionalities. Links to the departments responsible for the data are also provided. There is as yet no classification scheme applied to the City's electronic records. In fact, the classification scheme to be applied to paper records is still under development.

Case study 25, the Legacoop of Bologna Web site, reports that document creation and maintenance procedures regarding the Web site are not documented.[95] The case study reports

[90] Ibid., 23.
[91] Ibid., 20.
[92] Ibid.
[93] See Evelyn McLellan (2005), "InterPARES 2 Project - Case Study 24 Final Report: City of Vancouver Geographic Information System (VanMap)." Available at http://www.interpares.org/display_file.cfm?doc=ip2_cs24_final_report.pdf.
[94] Ibid., 5.
[95] See Mariella Guercio (2004), "InterPARES 2 Project - Case Study 25 Final Report: Legacoop of Bologna Web Site." Available at http://www.interpares.org/display_file.cfm?doc=ip2_cs25_final_report.pdf.

that "Internal controls, in terms of the content of what is published, [are] not performed except from what [is] prescribed by professional deontology of [those] responsible for any publication. No other defined audits or controls over documentary production are performed with reference to the digital resources."[96] There is a naming convention for the Web site system that appears to generate a unique identifier. Files are assigned names automatically through the editorial system, according to precise rules. "[F]iles are assigned a progressive identifier according to the category of information they belong to, a process that is transparent to users."[97] The identifier is an incremental number that functions as the primary key in the database system. For traditional (paper) records, the originals are filed in folders organized on a very simple classification scheme. There are some metadata recorded in the creator's registry system for traditional records only. The application provides a profile of the registered incoming and outgoing documents. The following parameters are registered: classification code, recipients, object, date and type of document.

General study data analysis

General study 10, Preservation Practices of Scientific Data Portals, involved a survey of the Web sites of thirty-two (32) broadly defined data services, archives, repositories or catalogues in the sciences.[98] The generic term "portal" refers to these services. The primary purpose of the survey was to collect information about the actual practices, standards, and protocols currently used by these portals and their users to ensure access, accuracy, reliability and authenticity of the data. The choice of the portals considered was based on recommendations from InterPARES 2 researchers who were familiar with and used these in their own research work. The portals selected pertained to different communities of practice in sciences such as health, astronomy, biology, engineering, statistics, genetics, geosciences and ecology, to name a few. This research was not intended to be exhaustive but is an overview that discusses the preservation structures in place, or lack thereof, in the examples surveyed.[99]

Most, but not all, of the data portals include metadata, some are very minimalist and include only header files (e.g., IP2SF4, Cambridge Crystallographic Data Centre), others refer to associated peer review articles, some were designed specifically for that particular data set while others adhere to the metadata standards of their discipline (e.g., IP2SF15, Canadian Geospatial Data Infrastructure), Access Portal or institutions (e.g., IP2SF10, World Data Center for Solar Terrestrial Physics).

General study 10 data portal IP2SF14, the Canadian Institute for Health Information (CIHI), includes fifteen databases and registries of health and healthcare information relating to healthcare services across Canada. At the development stage of each database or system, Steering Committee, Advisory Committee or Expert Working Groups are always established and are responsible for instructing and advising on the overall design and data quality. CIHI's Data Quality Strategy provides a common strategy for assessing data quality across all CIHI databases and registries. The Framework is built upon five criteria of quality, each of which has multiple

[96] Ibid., 6.
[97] Ibid., 7.
[98] See Tracey P. Lauriault and Barbara L. Craig (2007), "InterPARES 2 Project - General Study 10 Final Report: Preservation Practices of Scientific Data Portals." Available at http://www.interpares.org/display_file.cfm?doc=ip2_gs10_final_report.pdf.
[99] For a brief overview of the general study 10 research, see Tracey P. Lauriault and Barbara Craig (2006), "Do Data Access Portals, Repositories, and Catalogues, Preserve or Archive Geospatial and Science Data?" paper presented at the GeoTech 2006 Conference, 18-21 June 2006, Ottawa, ON, Canada.

dimensions: *accuracy* (how well information within a database reflects what was supposed to be collected—this includes documentation of all data processes), *comparability* (the extent to which a database can be properly integrated within the entire health system at CIHI—this includes identifying how conversions might pose problems for the data as well as maintaining accessible documentation on historical changes to the database), *timeliness* (how easily the storage and documentation of data allows one to understand how timely data or reports are), *usability* (how easily data may be understood and accessed) and *relevance* (incorporates all of the other dimensions to some degree but focuses specifically on value and adaptability). Data elements are developed for each individual database and these serve as metadata that describe information at its lowest (i.e., field) and most concrete level. A number of classification systems are used in collecting and analyzing CIHI information, including ICD-10-CA, Enhanced Canadian version of the 10th revision of the International Statistical Classification of Diseases and Related Health Problems; CCI, Canadian Classification of Health Interventions; and ICF, International Classification of Functioning, Disability and Health. No rights management metadata were identified during the case study.

General study IP2SF19, the National Virtual Observatory, includes extensive discussion of metadata about data collections and services that describe data and computational facilities and their locations and then how to use them. Metadata that describe resources include identity metadata (supplies a name and identifier for the resource), curation metadata (describes who supports the resource and availability information such as version and release date as well as the resource's provenance) and content metadata (describes aspects such as data type, sky coverage and spectral coverage, as well as rights management). This last can be applied to resources at various levels of granularity. Metadata relating to such aspects as calibration, consistency, and level of documentation also provide the basis for data quality assessment. Data quality is both subjective and quantitative, and data collections may have no single data quality metric. Although the completeness and consistency of the resource metadata itself may be a reasonable indicator of the associated resource, this is at best a qualitative measure. Each contributing institution uses its own metadata standards or guidelines.

Conclusions drawn from the case and general studies

Prior to examining the case study data and reports, the Description Cross-domain researchers had anticipated that they would see more, and more rigorous, metadata implementation in the governmental and scientific rather than the artistic sector, where legal requirements for recordkeeping or domain-specific data quality standards, measures and assurances respectively frequently provide warrant for the creation of such metadata. Overall, this did prove to be the case, although it would be useful to conduct further case studies in more areas of the arts and sciences to assess the extent to which these case studies are typical of the wider domains covered by those foci.

Although several of the science focus case studies indicate a strong awareness of the need for metadata and the role they can play in ensuring the accuracy and long-term usability of digital materials that is absent from the arts focus case studies,[100] overall, neither focus exhibited any

[100] In particular, the findings of the science focus case studies 06 and 19 show that highly specialized metadata related to a specific domain, discipline or business activity needs to be captured, maintained and preserved to ensure the preservation of authentic, reliable and trustworthy digital records. The records preserver in case study 19 went considerably further, using international metadata standards for resource description and then extending them using semantic metadata expressly designed to enable powerful new means of preserving authentic digital records independent of proprietary software and hardware.

real consciousness of the overall role of *recordkeeping* as opposed to resource identification or discovery or other types of metadata in its activities. The science focus case studies indicate that a rich level of metadata are created or could be created and that there is clearly an overall concern with metadata quality elements as well as for the importance of standardized naming conventions and version control. This is understandable, since metadata are essential for the dissemination of scientific data. In fact, without metadata that support effective data linkage, quality assessment and dataset authentication, scientific data sets have little, if any, long-term value. In addition to the authenticity of datasets, which is linked to a clear lineage recorded in the accumulating metadata surrounding scientific data, the value of data quality and lineage metadata in the sciences is considered axiomatic in that datasets, and the databases with which they interface, have little to no value unless the auxiliary information required to understand and use them correctly—i.e., the metadata—is included in, or inextricably linked to, the datasets.[101] However, scientific metadata standards need to explicitly address archival and preservation requirements as well as data quality and lineage requirements. An additional concern that was raised in the science focus is that despite element-rich, complex metadata schemas being developed in areas such as the geospatial domain, there is, in many cases, little incentive and/or few resources available actually to create metadata.[102] If archival researchers wish to influence these communities and persuade them to add even more elements to their schemas, then the researchers must be able not only to persuade them that it is in their own best interests, but also to help them create such metadata automatically and transparently. Case study 06 appears to be working in this direction, providing creators with an XML-based Authors' Toolkit.

In the arts focus case studies, whatever metadata-related practices there are tend to be idiosyncratic, ad hoc and at the discretion of individuals working with the system. Any metadata standards being implemented have been developed for resource description, discovery and use purposes and not with a view to ensuring the long-term preservation of authentic materials.

In the government focus, there was clearly more concern for evidential requirements and several of the case studies also raised the question of interfacing between digital and paper systems, although the metadata structures in both are generally not as integrated as is the business process. Although many metadata standards do currently exist within different government jurisdictions, the case studies did not reveal that those were being implemented in most of the systems examined.

Overall Results

The work of the InterPARES 2 Description Cross-domain represents the most sophisticated and comprehensive analysis undertaken to date of the requirements and real-life context for metadata that relate to the establishment of reliability and authenticity, as well as the long-term preservation and potential re-usability of digital materials.

There are two particularly noteworthy products or outcomes of this research. The first is the development of actual tools and specifications that can help individuals and institutions from a range of sectors and interests generate and preserve their digital assets in more thoughtful and effective ways. For example, whether those materials be records or other kinds of digital objects,

[101] National Research Council, Commission on Physical Sciences, Mathematics, and Applications, *Preserving Scientific Data on Our Physical Universe: A New Strategy for Archiving the Nation's Scientific Information Resources* (Washington, DC: National Academies Press, 1995), 31.
[102] Collaborative scientific projects are, however, an exception.

MADRAS can be used to identify ways in which they can be created and maintained in ways that will support their intellectual and physical integrity in and over time (although obviously the imperative is stronger for records associated with high degrees of risk or liability than it is for low risk records or non-record materials). Moreover, the development of the metadata specification model, which aligns closely with the OAIS model, will assist systems developers, as well as creators, managers and cataloguers of digital materials, in coping with what to date has been a highly intractable problem—the high costs (in terms of money, time, expertise and storage) of creating and managing optimal amounts of metadata to ensure maximum integrity and usability of the digital materials to which the metadata relate. The model provides a basis for developing automated tools that can systematically create, gather and manage various types of metadata, as well as identifying more closely what needs to be manually created and also what can be summarized and discarded at certain points.

The second noteworthy outcome, and one of the most interesting aspects of this multi-faceted work, is documenting the many levels upon which metadata work and need to work. The development of MADRAS established an ideal against which existing or draft metadata schemas and sets can be assessed.[103] The assessment conducted by InterPARES 2 researchers of selected schemas indicated that even recordkeeping or archival schemas fall short of that ideal, and non-recordkeeping schemas, as might be expected, fall much further short. However, that analysis also pointed up how the schemas are themselves, within the communities that generated them, ideals and that application profiles vary considerably from implementation to implementation, often stripping down a schema to what are considered to be "essential" elements or the elements that a given system is able to support or the creating institution or individual is able to afford or has sufficient expertise to create. Finally, coming a long way behind all of these considerations, are the actual implementations examined in the focus group case and general studies and the news archive survey, where there was scant evidence (with only a few notable exceptions in the science and government areas), especially in the arts focus, of any attempt to implement recordkeeping metadata at all. Although the trend in information management is toward the creation of leaner metadata, the researchers believe that it is important to contemplate how to change the dynamics of metadata depreciation and minimalization so that they work more in favour of the complexities of recordkeeping and preservation—educating communities and individuals more thoroughly about the role rich and rigorous metadata play in addressing needs that they may not even recognize until it is too late to do anything about it; and developing more specifications that could be built into off-the-shelf as well as customized software.

One major question surfaced by the Description Cross-domain's work arises not only with the differing scopes and viewpoints of the metadata schemas that have been registered and analyzed by the metadata schema registry, but also in the development of the analytical approach embedded in MADRAS and in the metadata specification models—Should these tools support a single or multiple worldviews on the activities, roles, responsibilities and points of engagement with the record? One of the great contributions and benefits of the InterPARES research over the past several years has been that it has brought together archival researchers not only from academe and practice, but also from very different archival traditions. This, however, has also led to moments of confusion and even contention as the divergent underlying perspectives and practices emerge and must be disambiguated and addressed if they are to be operationalized as

[103] Although it should be noted that as it is, it is difficult to perform a sophisticated interpretation of the analysis when the researchers are holding up all of these very different schemas emanating from very different domains, to a single standard set of questions born of a compromise made from two very different warrants.

tools. The Description Cross-domain researchers found themselves faced with two alternatives—one being the development of research products that tolerate and support more than one approach, the other being to attempt to reconcile approaches that appear at first, and maybe even at second glance, to be irreconcilable.

The Description Cross-domain attempted to straddle both of these alternatives. However, having made a conscious decision to assess the metadata implications of both of the dominant existing models, the relative extensiveness of the Business-driven Recordkeeping Model, with the dimensionality afforded by its four axes of identity, evidentiality, transactionality and recordkeeping entity,[104] necessitated that the Description Cross-domain take a more complex view of metadata and archival description than might have been needed if it had looked only at supporting a Lifecycle Model.

The activity models developed in InterPARES 1 were based on a lifecycle view and presumed a custodial approach to the preservation of archival records. The benchmark and baseline requirements identified responsibilities and capabilities for both the *creator* and the *preserver* but were still predicated upon the physical transfer of records into an archival repository. However, the Description Cross-domain has also had to address the fact that while these two theoretical models currently exist (and it is, of course, quite possible, that further models might emerge in the future), many different kinds of implementations also exist. Some of these implementations adhere to the traditional lifecycle view, but increasingly continuum thinking is influencing practices not only in Australia, but also in Northern Europe and the United States. What is more, archivists and other records keepers who are grappling with the challenges of electronic records, are developing their own hybrids of both approaches. In this context, it should be noted that although, historically, they have been linked closely together, conceptually it is not required that custodialism and non-custodialism be tied to adherence to the lifecycle and continuum worldviews, respectively. It is also important to bear in mind that the world outside of archival science does not use these models, at least not conceived of in these terms, but communities other than archival communities are also targeted user groups for the metadata schema registry and analytical framework and their needs must also be addressed.[105]

These results indicate several deficiencies and challenges in the current state of metadata for these purposes, deficiencies and challenges that need to be tackled by a variety of parties. For the archival and recordkeeping professions, there is a need to acknowledge and address the fact that there are both two worldviews (lifecycle and continuum) that necessitate different metadata specifications, and that the field has also developed two rich sets of metadata requirements that work from very different perspectives and have different degrees of focus and scope (InterPARES and ISO 23089). The work of the Description Cross-domain has also pointed out areas where each set is hard to operationalize in schema and system design. For developers of metadata schemas, MADRAS registration has found all to vary widely in terms of their documentation or meta-information, and some effort to standardize such materials would assist in long-term registration and management of metadata schemas. Moreover, schema developers need to recognize that proprietariness of schemas and their documentation works against schema

[104] Frank Upward (1996), "Structuring the Records Continuum, Part One: Post-custodial Principles and Properties," *Archives and Manuscripts* 24(2): 268–285. Online reprint version available at http://www.sims.monash.edu.au/research/rcrg/publications/recordscontinuum/fupp1.html; Frank Upward (1997), "Structuring the Records Continuum, Part Two: Structuration Theory and Recordkeeping," *Archives and Manuscripts* 25(1): 10–35. Online reprint version available at http://www.sims.monash.edu.au/research/rcrg/publications/recordscontinuum/fupp2.html.

[105] The Open Archival Information System (OAIS) Reference Model is a good example of a high-level model that, at first glance, seems to be a re-expression of a lifecycle model, but upon further scrutiny could equally well support a continuum approach.

registration and long-term management and preservation. Also, MADRAS analysis indicates that almost all schemas, including descriptive schemas developed by archivists in different national and international contexts, also fall short in addressing the needs of electronic records. For systems designers and the builders of automated tools, the metadata modeling demonstrates how more comprehensive metadata could be efficiently created and effectively managed across the life of the system and the records it contains. Finally, for funders of activities that create electronic systems containing records or with the potential to create records, this research raises the question as to whether a metadata creation, management and preservation regime should be required for those systems.

Areas for Future Research and Development

Several areas for future research and development emerged from the work of the Description Cross-domain. Two potential research questions are discussed below:

> *Can metadata associated with the creation and active use of records ever contribute to archival description, particularly in the capture and elucidation of certain kinds of context and fundamental identification and arrangement information relating to the records?*

One aspect of an integrated metadata creation and management regime that makes some in the archival community nervous is the notion, also raised by projects such as the Archivists' Workbench,[106] that certain types of metadata, created while the records to which they relate are active, could be captured or analyzed automatically and used to partially automate or even to replace archival description. As identified by InterPARES 1, records have many types of interacting contexts that need to be documented. Often with electronic records, because of their virtual nature and also their complexity, it can be more difficult to identify these contexts than it might be with traditional records. However, often it is the case that the system within which the record has been created or maintained has in place metadata mechanisms, or could be designed to have them, that document some of the context in which archivists are interested (albeit that these are generally created contemporaneous with the record and lack the hindsight and birds-eye view of the archivist).

Indeed, what is distinctive about recordkeeping metadata is the range of ways in which they can automatically capture salient contexts of records as they move through time, space, systems and types of use and user. For example, metadata can provide detailed descriptions of business processes and logs or audit trails of any changes made to records and associated dates. They can also describe the functionality of the original technical environment and enable users to distinguish the authoritative record from drafts and derivative versions. Metadata can also link separately stored data or record content to the appropriate documentary form to facilitate creating an imitative authentic copy of the original (an approach akin to that being used with the Persistent Archives Technology).

[106] This project involves the development of a prototype, infrastructure-independent management tool for software-dependent records in the form of a software application called the "Archivist's Workbench." For more information, see San Diego Supercomputer Center (1999), "Methodologies for the Long-Term Preservation of and Access to Software-Dependent Electronic Records," NHPRC Proposal, June 1, 1999. Available at http://www.sdsc.edu/NARA/Publications/nhprc_latest.pdf. A summary version of the proposal is available at http://www.sdsc.edu/NARA/Publications/nhprc_summary.pdf.

In the future, time and cost concerns as well as new technological capabilities are likely to necessitate that even archival description may be created, at least partially, by automated means, likely including harvesting and re-purposing metadata created by others prior to the records coming into archival custody. For this to be acceptable as an assistance or augmentation to archival description, however, a) the metadata harvested should supplement manual description or should capture some aspect that it is difficult or impossible to do manually; and b) archivists should assess what they do manually in traditional description and identify at the point of recordkeeping systems design what could be captured automatically out of the system. Neither of these activities, however, necessarily usurps the archivist's prerogative to supplement and synthesize the metadata gathered automatically in the process of creating a descriptive instrument. Moreover, because the metadata thus gathered are likely to be in digital form, the archivist would have the option of retaining it both in its original form, as evidence of the records and recordkeeping to which it relates, and to transform it into a form that is more useful for secondary use.

Can metadata-based automated tools support any new kinds of capabilities for the description and use of preserved digital materials?

Recordkeeping metadata are created in a variety of ways and by a variety of agents—they may be created manually (as is the case with most archival description) or automatically (as, for example, would be the case with an inverted index of terms culled from a text document). They may also be automatically inferred, derived or harvested from the records and recordkeeping systems themselves, an approach that looks increasingly attractive as systems developers and information professionals of all types become more aware of the burgeoning overhead of metadata creation and management necessary to support the online provision of trustworthy information. They may even be exploited and re-used for purposes for which they were never intended, such as for corporate knowledge mining, developing new institutional market segments or developing learning objects. In the archival community, research and development activities such as the Archivists' Workbench and PERM Projects of the San Diego Supercomputer Center have begun to explore the development of automated tools for metadata creation and management, as well as for the manipulation of records by end users, and the Clever Recordkeeping Metadata Project identified and prototype innovative ways of multi-purposing harvested recordkeeping metadata.

Approaches such as these potentially not only offer archivists a faster and less labour-intensive way to gain a measure of intellectual control over large volumes of electronic records, but also offer secondary users a much richer set of tools through which to access, manipulate and interpret archival records. They can also potentially support validation mechanisms for recordkeeping metadata and monitor the continued integrity of critical linkages that exist between records and their metadata. Perhaps the most important potential use of automated metadata tools, however, might be to support a metadata management regime, something which, if not automated, would be practically impossible for archivists to implement.

In terms of development work, the researchers hope to revise MADRAS so that it is more usable and useful by communities and researchers who are addressing metadata concerns. This would involve extending MADRAS' content and re-thinking its presentation and outputs. The researchers recognize that in the current incarnation of the reports generated, some of the information entered while registering the elements for each schema (including encoding schemas and repeatability) is not used in the evaluation. An improved report might weight schemas based

on such information. For example, if a requirement is satisfied by a required element or sub-element of a given schema, that schema would be designated stronger in that area than a schema that left such requirements to their non-mandatory elements. In addition, the researchers might make use of the presence or lack of an encoding schema specified for an element or the sub-elements of a schema. A schema whose element or sub-element has an encoding schema would be considered more robust than one that does not. One could also see that refining the report to provide the user with an analysis based specifically on how the schema performed within the various recordkeeping entities would be useful. In this way, the user could learn not only the strengths and weaknesses of the schema, but also more clearly where those strengths and weaknesses lie.

Integrating element-description-level information into the analysis and then testing the implications of an element's repeatability or its optional/mandatory status would greatly enhance the analysis of a schema's recordkeeping capability. It would be helpful to increase the amount of analytical information about the encoding schemas required by each schema. The assumption is that there will be times when the analysis can demonstrate that a schema element with specified encoding schema is stronger than one with no encoding schema. This may not be the case for all elements, however. For example, "title" would rarely be made stronger by the use of an encoding schema. Moreover, when registering elements, the Description Cross-domain researchers found that it is rarely the case that a metadata schema *requires* a given encoding schema for a particular data value. This information should be taken into account. Nevertheless, in cases where two recordkeeping schema each have elements covering the five recordkeeping entities (record, agent, mandates, business process and RK process), could the researchers compare these schemas by looking at any encoding schemas which are or are not required for each? Furthermore, does this vary from domain to domain? Would a schema used in the arts domain have different encoding schema requirements for the recordkeeping entities? Encoding schemas facilitate information retrieval, however, and at present MADRAS focuses on issues of metadata creation/preservation. To increase the emphasis on issues such as encodings would suggest an alteration in the focus of the tool. Another approach to increasing the information gleaned from the registration of the metadata elements might be to type the elements into certain categories (content, context and structure, for example) to get a feel for the overall goal of the schema. Then the analysis could take this information into consideration and not judge a description-heavy schema in the same way it does a context-heavy one.

A future iteration of MADRAS should examine whether the ranking of questions should be re-thought and apply that information in the generation of reports. This would require evaluating each question and giving it a weight as well as deciding what element information is absolutely necessary. For example, does a subject classification have more or less weight/importance than, say, the identification of an agent? Another issue for further examination is whether the division of questions by recordkeeping entity actually works well for MADRAS. Automating the analysis tool forced the Task Force researchers to, in effect, make the relationship between the two instruments (and the questions themselves) very rigorous and, as a result, many issues had to be framed as absolutes. In future implementations of MADRAS, the researchers would like to see the reporting become much more sophisticated such that these seemingly cut-and-dried questions could regain much more of their original nuance.

Appendix 17

Metadata Schema Analysis Questions

QID	Question	IP	ISO	CV Type	Controlled Vocab
	General Schema Information				
100	Is the schema intended for recordkeeping purposes?			Single CV Select (without Element List)	Yes#No
101	Whether or not the schema is intended for recordkeeping purposes, indicate which recordkeeping entities can be described. Select all that apply.			Multiple CV Select (without Element List)	Agents#Business Processes#Mandates#Records#Records Management Processes#Relationships#
102	Are the elements of the schema categorized according to a particular scheme?			Single CV Select (without Element List)	Yes#No
102	If the elements in the schema are categorized according to a particular scheme, indicate the categories used to classify the elements.			Open End	
103	In general, which of the following views of records management metadata can the schema address? Select all that apply.			Multiple CV Select (without Element List)	Business Perspective#Records Management Perspective#Use Perspective#
104	Schemas may be classified as single object schemas or multiple object schemas. What is the schema type of the schema?			Single CV Select (without Element List)	Single Object Schema#Multiple Object Schema
	Record Metadata Questions				
200	IDENTIFICATION OF RECORD AGGREGATION LEVELS			Dummy	
201	How does the schema describe a record object?			Multiple CV Select (without Element List)	Element Set#Element(s)#Other#Not Applicable
202	How does the schema describe record aggregations?			Multiple CV Select (without Element List)	Element Set#Element(s)#Other#Not Applicable
203	How does the schema describe recordkeeping systems or corporate archives?			Multiple CV Select (without Element List)	Element Set#Element(s)#Other#Not Applicable

204	How does the schema describe collective archives?			Multiple CV Select (without Element List)	Element Set#Element(s)#Other#Not Applicable
205	GENERAL RECORD METADATA			Dummy	
206	Does the schema contain elements to identify date related information for a record? If so, indicate the applicable element(s).	A1.a	9.2.1 a	Single CV Select with Element List	Yes#No
206	Indicate which of the following date types the schema describes. Select all that apply.	A1.a.iii		Multiple CV Select (without Element List)	Chronological Date#Creation Date#Received Date#Archival Date#Transmission Date#
207	Does the schema contain elements to identify time related information for a record? If so, indicate the applicable element(s).	A1.a	9.2.1 a	Single CV Select with Element List	Yes#No
207	Indicate which of the following time types the schema describes. Select all that apply.	A1.a.iii		Multiple CV Select (without Element List)	Creation Time#Compiled Time#Received Time#Archival Time#Transmission Time#
210	STRUCTURAL METADATA			Dummy	
211	Does the schema contain elements that directly identify record structure? If so, indicate the applicable element(s).		9.2.1 c	Single CV Select with Element List	Yes#No
212	Does the schema contain elements that directly identify record form? If so, indicate the applicable element(s).		9.2.1 d	Single CV Select with Element List	Yes#No
213	Does the schema contain elements that directly identify the chemical and physical properties of the record? If so, indicate the applicable element(s).	A3;A4	9.2.1 e	Single CV Select with Element List	Yes#No
214	Does the schema contain elements that directly identify the technical characteristics and dependencies of a record? If so, indicate the applicable element(s).	A4	9.2.1 f	Single CV Select with Element List	Yes#No
215	Does the schema contain elements that directly identify the technical requirements to render or reproduce records? If so, indicate the applicable element(s).	B1.c	9.2.1 h	Single CV Select with Element List	Yes#No
216	RECORD ACCESS			Dummy	
217	Does the schema contain elements that directly identify record aggregation level? If so, indicate the applicable element(s).	A1.a.iv; B3	9.2.1 m;9.2.3.1 a	Single CV Select with Element List	Yes#No
218	Does the schema contain elements that directly identify record location information (whether physical or logical)? If so, indicate the applicable element(s).	A2; A7; A8; B1.a; B3	9.2.3.1 c;9.6.1 h	Single CV Select with Element List	Yes#No
219	Does the schema contain elements that directly identify record title? If so, indicate the applicable element(s).	A1.a.iv	9.2.3.1 e;9.6.1 h	Single CV Select with Element List	Yes#No
220	Does the schema contain elements that directly identify record subject classification? If so, indicate the applicable element(s).	A1.a.iv	9.2.3.1 e; 9.6.1 h	Single CV Select with Element List	Yes#No
221	Does the schema contain elements that directly identify record descriptive keywords? If so, indicate the applicable element(s).	A1.a.iv	9.2.3.1 e;9.6.1 h	Single CV Select with Element List	Yes#No

222	Does the schema contain elements that directly identify a record abstract? If so, indicate the applicable element(s).	A1.a.iv	9.2.3.1 e;9.6.1 h	Single CV Select with Element List	Yes#No
223	Does the schema contain elements that directly identify the language of a record? If so, indicate the applicable element(s).	A1.a.iv	9.2.3.1 e;9.6.1 h	Single CV Select with Element List	Yes#No
224	Does the schema contain elements that directly identify record indexing? If so, indicate the applicable element(s).	A1.a.iv	9.2.3.1 e;9.6.1 h	Single CV Select with Element List	Yes#No
225	RECORD SECURITY			Dummy	
226	Does the schema contain elements that directly identify record and record aggregation access restrictions? If so, indicate the applicable element(s).	B1.b	9.2.4.1 a	Single CV Select with Element List	Yes#No
227	Does the schema contain elements that directly identify business process access restrictions? If so, indicate the applicable element(s).	B1.b	9.2.4.1 a	Single CV Select with Element List	Yes#No
228	Does the schema contain elements that directly identify agent access restrictions? If so, indicate the applicable element(s).	B1.b	9.2.4.1 a	Single CV Select with Element List	Yes#No
229	Does the schema contain elements that directly identify time limitations on record access restrictions? If so, indicate the applicable element(s).	B1.b	9.2.4.1 c	Single CV Select with Element List	Yes#No
230	Does the schema contain elements that directly identify time limitations on business process access restrictions? If so, indicate the applicable element(s).	B1.b	9.2.4.1 c	Single CV Select with Element List	Yes#No
231	Does the schema contain elements that directly identify time limitations on agent access restrictions? If so, indicate the applicable element(s).	B1.b	9.2.4.1 c;9.6.1 e	Single CV Select with Element List	Yes#No
232	RELATIONSHIPS TO THE RECORD			Dummy	
233	Does the schema document the relationship between the record and the business transaction or activity that generated it? If so, indicate the applicable element(s).		9.2.1 l	Single CV Select with Element List	Yes#No
234	TBD Does the schema capture the relationships between the data and the format element(s) comprising the record? If so, indicate the applicable element(s).		9.2.1 g	Single CV Select with Element List	Yes#No

Agent Metadata Questions

300	IDENTIFICATION OF AGENT GROUPS	Dummy	
301	How does the schema identify a person or actor?	Multiple CV Select (without Element List)	Element Set#Element(s)#Other#Not Applicable
302	How does the schema identify organizational unit and/or work group?	Multiple CV Select (without Element List)	Element Set#Element(s)#Other#Not Applicable

303	How does the schema identify organizations or corporate bodies?			Multiple CV Select (without Element List)	Element Set#Element(s)#Other#Not Applicable
304	How does the schema identify social institutions?			Multiple CV Select (without Element List)	Element Set#Element(s)#Other#Not Applicable
305	GENERAL AGENT METADATA			Dummy	
306	Does the schema contain elements that directly identify a creator of records? If so, indicate the applicable element(s).	A1.a.i	9.4.1 a;9.2.1 b	Single CV Select with Element List	Yes#No
307	Does the schema contain elements that directly identify agents involved in records management processes? If so, indicate the applicable element(s).	A1.a.i	9.4.1 b	Single CV Select with Element List	Yes#No
308	Does the schema contain elements that directly identify agents involved in records transactions (addressees, originators, etc.)? If so, indicate the applicable element(s).	A1.a.i	9.5.1 c	Single CV Select with Element List	Yes#No
309	AGENT SECURITY	A2		Dummy	
310	Does the schema contain elements that directly identify agent authorizations for records management processes? If so, indicate the applicable element(s).	B1.b; B2.a	9.4.1 b	Single CV Select with Element List	Yes#No
311	Does the schema contain elements that directly identify access privileges for authorized agents to create, use, modify, destroy, etc. records? If so, indicate the applicable element(s).	B1.b	9.4.1 c;9.6.1 d	Single CV Select with Element List	Yes#No

Business Process Metadata Questions

400	IDENTIFICATION OF BUSINESS PROCESS CATEGORIZATIONS			Dummy	
401	How does the schema identify business transactions?			Multiple CV Select (without Element List)	Element Set#Element(s)#Other#Not Applicable
402	How does the schema identify business activities?			Multiple CV Select (without Element List)	Element Set#Element(s)#Other#Not Applicable
403	How does the schema identify business functions?			Multiple CV Select (without Element List)	Element Set#Element(s)#Other#Not Applicable
404	How does the schema identify ambient functions?			Multiple CV Select (without Element List)	Element Set#Element(s)#Other#Not Applicable
405	GENERAL BUSINESS PROCESS METADATA			Dummy	
406	Does the schema contain elements that directly identify business processes? If so, indicate the applicable element(s).		9.5.2 a	Single CV Select with Element List	Yes#No

407	Does the schema contain elements that directly identify the security and access rules for business processes and transactions? If so, indicate the applicable element(s).	B1.b	9.5.1 d	Single CV Select with Element List	Yes#No
408	Does the schema contain elements that directly identify business process classification schemes? If so, indicate the applicable element(s).		9.2.3.1 f; 9.5.1 f; 9.6.1 g	Single CV Select with Element List	Yes#No
409	Does the schema contain elements that directly identify the date of a transaction? If so, indicate the applicable element(s).	A1.a.iii	9.5.1 h	Single CV Select with Element List	Yes#No
410	Does the schema contain elements that directly identify the time of a transaction? If so, indicate the applicable element(s).	A1.a.iii	9.5.1 h	Single CV Select with Element List	Yes#No
411	RELATIONSHIPS TO BUSINESS PROCESS			Dummy	
412	TBD Does the schema document the relationship between business processes and the records? If so, indicate the applicable element(s).		9.5.1 b	Single CV Select with Element List	Yes#No
413	Does the schema document the relationship between business processes and agents? If so, indicate the applicable element(s).		9.5.1 b	Single CV Select with Element List	Yes#No

Mandate Metadata Questions

500	INTERNAL BUSINESS RULES & SYSTEM CONTROLS			Dummy	
501	Does the schema contain elements to identify metadata schemes or schemas used in organization business systems? If so, indicate the applicable element(s).		9.3.1 a	Single CV Select with Element List	Yes#No
503	Does the schema contain elements to identify rules or controls that regulate record creation? If so, indicate the applicable element(s).		9.3.1 b	Single CV Select with Element List	Yes#No
504	Does the schema contain elements to identify rules or controls that regulate records management? If so, indicate the applicable element(s).	B1.b	9.3.1 b	Single CV Select with Element List	Yes#No
505	Does the schema contain elements to identify rules or controls that regulate record access? If so, indicate the applicable element(s).	B1.b	9.3.1 e	Single CV Select with Element List	Yes#No
506	Does the schema contain elements to identify rules or controls that regulate the rights to records? If so, indicate the applicable element(s).	B1.b	9.3.1 e	Single CV Select with Element List	Yes#No
507	Does the schema contain elements to identify rules or controls for records management operations? If so, indicate the applicable element(s).	B1.b	9.3.1 d	Single CV Select with Element List	Yes#No
508	Does the schema contain elements to identify rules or controls for metadata creation? If so, indicate the applicable element(s).		9.3.1 c	Single CV Select with Element List	Yes#No

509	Does the schema contain elements to identify rules or controls for metadata management? If so, indicate the applicable element(s).	B1.b	9.3.1 c	Single CV Select with Element List	Yes#No
510	EXTERNAL MANDATES AND REGULATIONS			Dummy	
511	Does the schema contain elements to identify mandates or regulatory requirements for record creation? If so, indicate the applicable element(s).		9.3.1 f	Single CV Select with Element List	Yes#No
512	Does the schema contain elements to identify mandates or regulatory requirements for records management? If so, indicate the applicable element(s).		9.3.1 f	Single CV Select with Element List	Yes#No
513	Does the schema contain elements to identify mandates or regulatory requirements for record retention? If so, indicate the applicable element(s).		9.3.1 g	Single CV Select with Element List	Yes#No
514	Does the schema contain elements to identify mandates or regulatory requirements for record security? If so, indicate the applicable element(s).		9.3.1 g	Single CV Select with Element List	Yes#No
515	Does the schema contain elements to identify mandates or regulatory requirements for record destruction? If so, indicate the applicable element(s).		9.3.1 g	Single CV Select with Element List	Yes#No
516	RELATIONSHIPS TO INTERNAL AND EXTERNAL MANDATES			Dummy	
517	Does the schema document the relationship between external mandates and regulations and the records? If so, indicate the applicable element(s).		9.3.1h	Single CV Select with Element List	Yes#No
518	Does the schema document the relationship between external mandates and regulations and internal records management operations? If so, indicate the applicable element(s).		9.3.1h	Single CV Select with Element List	Yes#No

RK Process Metadata Questions

600	GENERAL RECORDKEEPING METADATA			Dummy	
601	Does the schema contain elements to identify retention information applied to a record? If so, indicate the applicable element(s).	B1.a	9.6.1 b	Single CV Select with Element List	Yes#No
602	Does the schema contain elements to identify disposition authority schemes? If so, indicate the applicable element(s).		9.6.1	Single CV Select with Element List	Yes#No
603	Does the schema contain elements to identify security classification schemes? If so, indicate the applicable element(s).		9.6.1	Single CV Select with Element List	Yes#No
604	Does the schema contain elements to identify a record's registration into a record system? If so, indicate the applicable element(s).		9.6.1	Single CV Select with Element List	Yes#No
605	LONG-TERM PRESERVATION			Dummy	

606	Does the schema contain elements to document conversion of records from one format or medium to another? If so, indicate the applicable element(s).		9.6.2 a;9.2.1 i;9.2.2	Single CV Select with Element List	Yes#No
607	Does the schema contain elements to document the agent responsible for record conversion? If so, indicate the applicable element(s).	B2.a	9.6.2 a;9.2.2	Single CV Select with Element List	Yes#No
608	Does the schema contain an element(s) to document the date and time of record conversion? If so, indicate the applicable element(s).	B2.a	9.6.2 a;9.2.2	Single CV Select with Element List	Yes#No
609	Does the schema contain an element(s) to document any changes created by the record conversion? If so, indicate the applicable element(s).	B2.c	9.6.2 a;9.2.2	Single CV Select with Element List	Yes#No
611	(TBD)?? Relationship between the records acquired from the creator and the copies produced by the preserver.	B2.b		Single CV Select with Element List	Yes#No
612	(TBD) Does the schema contain an element(s) to document the reason for record conversion? If so, indicate the applicable element(s).		9.6.2 a;9.2.2	Single CV Select with Element List	Yes#No
616	Does the schema contain elements to document migration of records from one hardware or software system to another? If so, indicate the applicable element(s).	B2.c		Single CV Select with Element List	Yes#No
617	Does the schema contain elements to document the agent responsible for record migration? If so, indicate the applicable element(s).			Single CV Select with Element List	Yes#No
618	Does the schema contain an element(s) to document the date and time of record migration? If so, indicate the applicable element(s).			Single CV Select with Element List	Yes#No
619	Does the schema contain an element(s) to document any changes created by the record migration? If so, indicate the applicable element(s).			Single CV Select with Element List	Yes#No
620	(TBD) Does the schema contain an element(s) to document the reason for record migration? If so, indicate the applicable element(s).			Single CV Select with Element List	Yes#No
621	Does the schema contain elements to document other procedures to counteract media fragility and technological obsolescence ? If so, indicate the applicable element(s).	A4		Single CV Select with Element List	Yes#No
624	RECORD BACKUP			Dummy	
625	Does the schema contain elements to document regular record back-ups performed on records? If so, indicate the applicable element(s).	A3		Single CV Select with Element List	Yes#No
626	Does the schema contain elements to document the agent responsible for record backups? If so, indicate the applicable element(s).	A3		Single CV Select with Element List	Yes#No
627	Does the schema contain an element(s) to document the date and time of record backup? If so, indicate the applicable element(s).	A3		Single CV Select with Element List	Yes#No

628	Does the schema contain an element(s) to document any changes created by the record backup? If so, indicate the applicable element(s).	A3		Single CV Select with Element List	Yes#No
629	(TBD) Does the schema contain an element(s) to document the reason for record backup? If so, indicate the applicable element(s).			Single CV Select with Element List	Yes#No
635	RECORD REPRODUCTION			Dummy	
636	Does the schema contain elements to identify record reproduction activities (other than backups) for a record? If so, indicate the applicable element(s).			Single CV Select with Element List	Yes#No
637	Does the schema contain an element(s) to indicate the agent responsible for record reproduction? If so, indicate the applicable element(s).	B2.a		Single CV Select with Element List	Yes#No
638	Does the schema contain an element(s) to indicate the date and time of record reproduction? If so, indicate the applicable element(s).	B2.a		Single CV Select with Element List	Yes#No
639	Does the schema contain an element(s) to document the relationship between records acquired from the creator and copies produced by the preserver? If so, indicate the applicable element(s).	B2.b		Single CV Select with Element List	Yes#No
640	Does the schema contain an element(s) to identify any changes to the record due to copy procedures? If so, indicate the applicable element(s).	B2.c		Single CV Select with Element List	Yes#No
641	Does the schema contain an element(s) to document changes and provide such information both to the preservers and users? If so, indicate the applicable element(s).	B2.d		Single CV Select with Element List	Yes#No
650	RECORD AUTHENTICATION			Dummy	
651	Does the schema identify authentication declarations for records or groups of records? If so, indicate the applicable element(s).	A6	9.6.1 c	Single CV Select with Element List	Yes#No
652	Does the schema identify rules under which authentication declarations are utilized? If so, indicate the applicable element(s).	A6	9.6.1 c	Single CV Select with Element List	Yes#No
653	Does the schema identify agents providing authentication declarations? If so, indicate the applicable element(s).	A6	9.6.1 c	Single CV Select with Element List	Yes#No
660	RECORD TRACKING			Dummy	
661	Does the schema contain elements to identify record access? If so, indicate the applicable element(s).	B1.b		Single CV Select with Element List	Yes#No
662	Does the schema contain elements to identify an agent who accessed the record? If so, indicate the applicable element(s).	B1.b		Single CV Select with Element List	Yes#No
663	Does the schema contain elements to identify the date a record was accessed? If so, indicate the applicable element(s).	B1.b		Single CV Select with Element List	Yes#No

664	Does the schema contain elements to identify the action taken on record? If so, indicate the applicable element(s).	B1.b		Single CV Select with Element List	Yes#No
665	Does the schema contain elements to identify the physical transfer of records from one location to another? If so, indicate the applicable element(s).	B1.a		Single CV Select with Element List	Yes#No
666	Does the schema contain elements to identify the agent responsible for the transfer of records from one physical location to another? If so, indicate the applicable element(s).	B1.a		Single CV Select with Element List	Yes#No
667	Does the schema contain elements to identify the date and time of record transfer? If so, indicate the applicable element(s).	B1.a		Single CV Select with Element List	Yes#No
668	Does the schema contain elements to identify the transfer location name? If so, indicate the applicable element(s).	B1.a		Single CV Select with Element List	Yes#No
670	Does the schema contain elements to identify the custodial transfer of records from one agent to another? If so, indicate the applicable element(s).	B1.a		Single CV Select with Element List	Yes#No
680	Does the schema capture changes to the classification of a record over time? If so, indicate the applicable element(s).		9.2.2	Single CV Select with Element List	Yes#No
681	AGENT TRACKING			Dummy	
682	Does the schema capture changes to personnel (including records creators, managers, users, etc.) over time?		9.2.3.2 b	Single CV Select with Element List	Yes#No
683	Does the schema capture changes to the agents' roles (including security classifications) over time? If so, indicate the applicable element(s).		9.4.2	Single CV Select with Element List	Yes#No
684	BUSINESS PROCESS TRACKING			Dummy	
685	TBD: Does the schema capture changes to records-producing business processes over time? If so, indicate the applicable element(s).		9.5.2;9.2.3.2 a;9.2.3.2 g	Single CV Select with Element List	Yes#No
686	Does the schema capture changes to business process classification schemes over time? If so, indicate the applicable element(s).		9.5.2	Single CV Select with Element List	Yes#No
687	MANDATE TRACKING			Dummy	
688	Does the schema capture changes to business rules and system controls over time? If so, indicate the applicable element(s).		9.3.2	Single CV Select with Element List	Yes#No
689	Does the schema capture changes to external mandates and regulatory requirements over time? If so, indicate the applicable element(s).		9.3.2	Single CV Select with Element List	Yes#No
690	METADATA RECORD TRACKING			Dummy	
691	Does the metadata schema contain elements to manage metadata about the metadata record ? If so, indicate the applicable element(s).		8.3.9.2	Single CV Select with Element List	Yes#No
691	Indicate what metadata record information the schema describes. Select all that apply.		8.3.9.2;8.3.8	Multiple CV Select (without Element	Date and time of metadata creation or alteration#Date and time of metadata

			List)	alteration#Agent responsible for metadata creation or alteration#Activity undertaken to metadata record#Metadata version number#Security and access restrictions for metadata#
697	RELATIONSHIPS TO THE RECORDKEEPING PROCESS		Dummy	
699	Does the schema document the relationship between the record, the agent and recordkeeping processes the agent performs on the record? If so, indicate the applicable element(s).	9.6.1 i	Single CV Select with Element List	Yes#No

Appendix 18

Case Study Data Relating to Metadata

Focus 1. Artistic Activities

General information regarding metadata

CS01, Arbo Cyber, théâtre (?)

The report states that no descriptive schemas and metadata are employed. However, records are classified by date of the performance (not by their date/time of digitization) to which they are linked. Individual practices are used to relate to the functional and technological needs of the *Ludosynthese*. However, the report reveals that if Arbo Cyber, theatre (?) decides to enter digital information, these properties will be limited to the programs' capabilities.

CS02, Performance Artist Stelarc

The report states that Stelarc has no recognized system of organization from an archival point of view with regard to his digital materials. Instead, the materials are arranged according to Stelarc's performance and publicity needs. No documented processes or procedures are used to identify, retrieve or access his digital materials. Although some records access and modification restrictions are in place, these do not appear to be formally documented. In effect, there are few, if any, formal recordkeeping practices and no metadata are consciously or intentionally recorded.

CS03, HorizonZero/ZeroHorizon Online Magazine & Media Database

The report states that the organization of the files pertaining to each issue of *HorizonZero* is ad hoc and is generally organized by the issue for which they were created. These files are accessible through a shared space that can be navigated using tracking software that organizes the posting into threads.

These tracker entries are saved using an archival function implemented in the tracker software (Mantis 0.18.0A4).

CS09(01), Altair4 di Roma

The report states that there is neither a recordkeeping system nor metadata schemas; however, Altair4 uses the "Where is it" program to reorganize and retrieve digital entities. To use them, it is necessary to know the filename, path and approximate date of production.

CS09(02), National Film Board

The report states that all work done using the computer as an intermediary [...] is kept on the server system and is related to a given project by the project number (assigned before a production is given the go-ahead; once a production is approved, it is given a unique production number).

CS09(03), Commercial Film Studio

The report states that only those digital entities that are archived have metadata. The standards used are Dublin Core, the Thesaurus for Graphic Materials I & II and AACR2.

CS09(04), WGBH Boston

It is important to note that the current production entity investigated during the case study consisted of a mixed analogue/digital production system. At the time of the case study, the creator was in the process of converting to a digital asset management (DAM) system, while at the same time maintaining its collection of analogue film, tapes and audio content that

dates back to the 1950s. Catalogue records for these materials are kept in a FileMaker Pro 7.0 database designed and developed in-house. The DAM system is an Artesia TEAMS product that has been customized by WGBH.

CS10, Danube Exodus

File naming is largely ad hoc and some individuals develop their own system. Therefore, there is no formal recordkeeping system; furthermore, there is no system to track the changes, actions or transactions to the digital files.

CS13, *Obsessed Again...*

The report implicitly states that no metadata schemas or standards are employed. There is no formal recordkeeping system. All digital entities are stored on computer disks, which remain in the possession of the composer. These entities are only identified through the assignment of a semi-descriptive filename.

CS15, *Waking Dream*

The reports states that metadata is not consciously captured. The digital entities are kept in simple directories and are not entered in any sophisticated recordkeeping system. Professor Fels wrote the code used in *Waking Dream* and maintains it on his computer. Thus, retrieval and access of these digital entities is dependent on whether or not the computer in question contains the necessary application.

Metadata information in the 23 questions:

4d. How are the digital entities identified (e.g., is there a [persistent] unique identifier)?

CS01, Arbo Cyber, théâtre (?)

Arbo Cyber, theatre (?) does not make use of a persistent or unique identifier for electronic records, but they do use a naming convention. This was referred to during the interviews as the "nomenclature": it makes use of a strict set of punctuation and spelling rules and relies on signifying and representative values[24]. This abbreviation code is very important in the *Ludosynthese*, as it indicates location within the site.

CS02, Performance Artist Stelarc

The digital entities are identified under project titles, event series and biographical content on the Web site.

CS03, *HorizonZero/*ZeroHorizon Online Magazine & Media Database

The digital entities are identified by naming conventions that are ad hoc, though some staff members have evolved consistent naming conventions for their own work.

CS09(01), Altair4 di Roma

These conventions comprise the folder with project name/file object name/number of version and the last version file object name/final version.

CS09(02), National Film Board

No information provided.

CS09(03), Commercial Film Studio

Strict naming conventions are used to identify the digital entities, and all those having a role in manipulating the file are required to adhere to these conventions. Among other elements, the name of the file contains information on the sequence, the scene, the name of the object as well as numerical information to identify the version. The sequence of information in the file name is:/studio/title/sequence/scene/object/version.

Interpretation of this information is as follows: "Studio" refers to the name of the studio that owns the artwork, since occasionally artwork is outsourced to another studio or a subsidiary. "Title" refers to the working title of the film being produced. "Sequence" and "Scene" refer respectively to these parts of the film (in the parlance of the studio studied, "scene" is the equivalent of "shot"). "Object" refers to the particular piece of artwork in hand. Finally, a version number is added to identify the precise iteration of the file. Sometimes in PODS (a proprietary system) or at the story stage, there is also an abbreviation for information such as the sequence date and the name of the artist. There has been some attempt to develop a consistent taxonomy. Specific terms to describe each object in development are selected in the brainstorming stage by the production team. Thus there is agreement by committee on the naming conventions to be used for each production. These, however, do not extend from one production to another.

CS09(4), WGBH Boston

Current: Yes, and the unique identifier links the catalogue red in the log with the original footage. The original footage and logs follows naming conventions that link them together and to the final production. Please see question 4(f).

DAM: Same as above.

CS10, *Danube Exodus*

No alternative attempt to apply persistent unique identifiers was noted. Most files were organized in folders whose directory structure seemed to follow the intellectual conceptualization of the project.

CS13, *Obsessed Again...*

The report states that the format of each digital file is dictated by the specifications of the individual software programs with which they were created. The NoteWriter, Max/MSP and Editor/Librarian files are proprietary, binary formats and, as such, their specifications are unreleased. The MIDI files used by the Max/MSP patches are standard text files following the MIDI specification.

CS15, *Waking Dream*

The report states that the digital entities are uniquely identified with file names and, when changes have been made, with version numbers.

18b. From what applications do the recordkeeping system(s) inherit or capture all digital entities and the related metadata (e.g., e-mail, tracking systems, workflow system, office system, databases, etc.)?

CS01, Arbo Cyber, théâtre (?)

This question does not really apply to Arbo Cyber, theatre (?), but it can be said that the documents are influenced by the programs used by the artists, such as Photoshop, Illustrator or Flash. However, the properties gained through these programs have no real significance and therefore cannot be seen to have any real value for the recordkeeping system.

CS02, Performance Artist Stelarc

The applications that Stelarc captures their digital entities and related metadata are from the following, the mail system, Web-driven database operated by Web host, Internet networks, public databases functioning as sources for data mining and conversion into performance images.

CS03, HorizonZero/ZeroHorizon Online Magazine & Media Database

The report states that the recordkeeping system is not an RMA; the documents are "captured" by transferring them from individual hard drives to the shared server space. Metadata are attached to those documents (once again, not automatically) that are subsequently transferred to the ZeroHorizon database.

CS09(01), Altair4 di Roma

Not applicable.

CS09(02), National Film Board

The records in the recordkeeping system come primarily from office systems such as Microsoft Office, as well as from various graphics systems (for photography and posters, for example).

CS09(03), Commercial Film Studio

Another database, built on FileMaker Pro and called ArchiveWorks, is used for tracking physical pieces of artwork that are not digital.

CS09(04), WGBH Boston

Current: Productions stand alone FileMaker databases feed into the Archives database.

DAM: Same as above and through direct user input.

CS10, *Danube Exodus*

None of the subjects have a formal or automated recordkeeping system, though all have some process by which records are kept. There is therefore no system in place to track changes, actions or transactions to digital files, beyond renaming by individuals and such strategies, and, as far as can be ascertained, none of the subjects employ any kind of digital or media asset management system that could perform similar functions. (It has not been possible to confirm this with C_3.) All the subjects stated that they attempted to keep all relevant files, despite only really being concerned about the fate of work files and any secondary files that would allow them to remain functional. What constituted relevant or important files was largely left to the discretion of whatever individual was regarded as responsible for the project; for instance, the Project Manager at the Labyrinth Project.

CS13, *Obsessed Again...*

None.

CS15, *Waking Dream*

Not applicable.

18d. Does the recordkeeping system provide ready access to all relevant digital entities and related metadata?

CS01, Arbo Cyber, théâtre (?)

The report states that access is not direct, because the preservation strategy involves transferring records and placing them on external storage devices. Furthermore, Arbo controls their own entities without any need for particular measures of control.

CS02, Performance Artist Stelarc

Yes. Links are also present to make collaborators' Web sites and other relevant internet locations accessible. If general links become obsolete, the webmaster will keep them on the Web site as dead links. If important links become obsolete, new links will be set up to make that information accessible.

CS03, HorizonZero/ZeroHorizon Online Magazine & Media Database

Yes.

CS09(01), Altair4 di Roma

Not applicable.

CS09(03), Commercial Film Studio

Yes, access is maintained for all relevant digital entities and their metadata. Everything in the system that can be opened can be downloaded.

CS09(04), WGBH Boston

Current: No, the analogue/digital hybrid nature makes access cumbersome, though possible.
DAM: The fully digital nature of the recordkeeping system allows for greatly improved access, as well as the implementation of automatic standard language applications and thesaurus capability.

CS10, *Danube Exodus*

The report does not explicitly state how it provides access to the digital entities.

CS13, *Obsessed Again...*

Again, no system exists, but Dr. Hamel currently has ready access to all relevant digital entities.

CS15, *Waking Dream*

Not applicable.

18e. Does the recordkeeping system document all actions/transactions that take place in the system re: the digital entities? If so, what are the metadata captured?

CS01, Arbo Cyber, théâtre (?)

The lack of a true recordkeeping system makes it difficult to apply this question. The entities are saved on external storage devices; thus, it is impossible to modify them or for the system to document these modifications.

CS02, Performance Artist Stelarc

No, the Web master does not keep a record of specific updates to the Web site. The report states that the metadata are unknown.

CS03, HorizonZero/ZeroHorizon Online Magazine & Media Database

The report states there are no recordkeeping system.

CS09(01), Altair4 di Roma

Not applicable.

CS09(02), National Film Board

If different versions of digital entities are created by the animator, these must have a separate identification in order that they be retrievable. However, it is not known what metadata are captured as the NFB's Synchrone system (an intranet comprised of an integration of multiple databases created through in-house software developments) used is unique to the National Film Board and the subject was not queried during the interview process.

CS09(03), Commercial Film Studio

No, for the moment only the check-in and check-out transactions are documented. Some transactions modify a record's metadata, but these are not documented at present.

CS09(04), WGBH Boston

Current: Partially. Use of tapes is tracked in a FileMaker database, but re-use of shots is not tracked. DAM: Yes, each use will be noted along with versioning.

CS10, *Danube Exodus*

No.

CS13, Obsessed Again...

No such documentation exists.

CS15, *Waking Dream*

No metadata is consciously captured.

22. What descriptive or other metadata schema or standard are currently being used in the creation, maintenance, use and preservation of the recordkeeping system or environment being studied?

CS01, Arbo Cyber, théâtre (?)

The report states that FLA files in Flash allow for notes in a "grey-zone" that are inaccessible to users. They are used as memory aids, and no specific data is required. Furthermore, the notes only deal with content. These "grey-zones" also fail to capture information concerning the records themselves. The informant also did not see the use in identifying metadata. The informant had no knowledge of the information that can be captured in digital images. The only data attached to these images was that created automatically by the computer at the moment of creating and saving files.

CS02, Performance Artist Stelarc

This is unknown.

CS03, HorizonZero/ZeroHorizon Online Magazine & Media Database

No descriptive or metadata schema are consistently used for the records of *HorizonZero* pertaining to the production of each issue. There are naming conventions that describe the content of some records, but most records can be identified only by their context in the filing system.

CS09(01), Altair4 di Roma

There are no standards for activity of a creative nature. Since Altair4 uses no recordkeeping system, no reference is made to standards of description and/or indexing.

CS09(02), National Film Board

The NFB is introducing the use of MPEG-7 and MPEG-21 as standards for encoding content and rights about films. These are being introduced to simplify commercialization.

CS09(03), Commercial Film Studio

There are no standards for creation of the assets in the workflow pipeline. However, the archivist has introduced standards for description and indexing which cover those assets that make it to the archive. These include the Categories for the Description of Works of Art (CDWA), the Dublin Core (DC), the Thesaurus for Graphic Materials I: Subject Terms (TGMI), the Thesaurus for Graphic Materials II: Genre and Physical Characteristics Terms (TGM II). The Anglo-American Cataloguing Rules are used to describe scripts, manuscripts, partial notes and such. Some tracking information about other documentation is recorded using the Turabian Style Guide and The Chicago Manual of Style.

CS09(04), WGBH Boston

Current: In-house descriptive standards combined with modified Library of Congress Subject Headings. DAM: The above plus Dublin Core and PBCore (i.e., Public Broadcasting Core) compliant.

CS10, *Danube Exodus*

The interim report states that neither standards nor schemas are being used consistently in the environments studied. Forgács does capture metadata in the course of his work, but it is a system largely based on individual need, as informed by standard professional filmmaking practice. However, to date it is uncertain to the extent to which any metadata schema is currently used within the institution.

CS13, *Obsessed Again...*

There are no descriptive or other metadata schemas or standards currently being used.

CS15, *Waking Dream*

No descriptive or metadata standards are currently being used. There is no recordkeeping system being used.

23. What is the source of these descriptive or other metadata schema or standards (institutional conventions, professional body, international standard, individual practice, etc.?)

CS01, Arbo Cyber, théâtre (?)

Arbo does not use any descriptive or metadata standards. The report states that the "grey-zones" list information; thus, are not standardized.

CS02, Performance Artist Stelarc

The final report states that it is likely individual practice by Stelarc and his Web master that are the sources for any descriptive standards.

CS03, HorizonZero/ZeroHorizon Online Magazine & Media Database

The CanCore standard is derived from the Dublin Core metadata set, and is based on and fully compatible with the IEEE Learning Object Metadata standard and the IMS Learning Resource Meta-data specification. Other metadata sets are the result of individual practice.

CS09(01), Altair4 di Roma

The final report states that only material that is archived are then governed by international standards.

CS09(02), National Film Board

The NFB participates in international standards making bodies and is in some instances responsible for either assisting in developing these or in adapting them to the Canadian scene. These standards are, however, technical rather than descriptive.

CS09(03), Commercial Film Studio

Institutional convention governs practice during the workflow stage for any particular production. A snapshot of the entire directory structure for each production is kept, but users trying to access materials from even recent productions have been unsuccessful because of hardware and software changes that occurred in the meantime. Material that is archived is done so using the tools listed in the answer to Question 22, so professional bodies and international standards govern these activities.

CS09(04), WGBH Boston

Current: In-house data entry personnel with professional archives and library training, Library of Congress published and on-line sources. DAM: The above plus Dublin Core and PBCore (i.e., Public Broadcasting Core) reference resources.

CS10, *Danube Exodus*

The interim report states that this is not applicable.

CS13, Obsessed Again...

No such schema or standards are employed.

CS15, *Waking Dream*

Not applicable.

Focus 2. Scientific Activities

General information regarding metadata

CS06, Cybercartographic Atlas of Antarctica

The final report states that metadata in the field of geomatics are critical to business processes. The Cybercartographic Atlas of Antarctica acquires data from a number of organization and these data sets are accompanied by metadata (see Appendix K of the final report for details). The Atlas itself adheres to the ISO19115 geographic metadata standard for each module that has been entered into the MADRAS Registry developed at UCLA. Digital multimedia information objects (e.g., video clips, photos, audio, webcams, etc.) are also fully referenced and include metadata embedded into the object and/or accompanying the object and/or referenced as a caption and acknowledged in the bibliography of each content module. CS06 includes metadata-specific documents as follows:

- Excerpt - *Elements of geospatial data quality, March 8, 2002*
- *Multimedia Metadata Discussion Document, December 2003*
- Appendix P, List of Standards Adhered to on the Project

CS08, Mars Global Surveyor Data Records (NASA)

The PDS (Planetary Data System) uses self-describing data files as a preservation strategy. The labels of self-describing files describe the file format of attached data as well as the context in which the data were created. The PDS is referred to as an "active archive," whereas the National Space Science Data Center's (NSSDC's) repository is referred to as a "deep archive." The PDS is the entrance for Planetary Science data into the NSSDC archives for long-term preservation.

CS14, Archaeological Records in a GIS

The final report states that process for creating and maintaining the digital entities is ad hoc, even though GIS dynamically links geospatial data and descriptive attribute data from a wide variety of sources, and thus is a spatially referenced data set with specific metadata.

CS19, Preservation and Authentication of E-Engineering and Manufacturing Records

The main purpose of the engineering experiment examined by CS19 was to develop an open-source preservation format for digital computer-aided design (CAD) records of solid models used in high-tolerance manufacturing of complex assemblies. The experiment used Web Ontology Language (OWL), a W3C specification that extends XML to allow representation of semantics within metadata schemas, to persist the geometry, topology and functional characteristics of CAD model objects. The semantic format enabled automated querying of the digital entity's meaning, expressed in its metadata in order to assess its authenticity. The CAD model objects were developed using proprietary reasoning programs and instantiated in accordance with ISO 10303, Standard for the Exchange of Product Model Data (STEP), AP 203 and part 21 EXPRESS. STEP is ISO's metadata standard for the representation and electronic exchange of industrial product data between computer-based product life-cycle

systems. AP 203 specifies the complete boundary representation of a solid model and EXPRESS defines its elements and attributes using an object-oriented approach (see 4a, below). Metadata elements were stored in the metadata catalogue management system (MCAT) of the ISO 14721, Open Archival Information System-compliant pilot preservation system managed by CS19 partners the Electronic Records Archives (ERA) Program of the U.S. National Archives and Records Administration (NARA), the University of Maryland and the San Diego Supercomputer Center (SDSC). This preservation system also incorporated SDSC's Storage Resource Broker technology, a middleware application that uses grid and metadata technologies to transparently manage data. The intent of the experiment was to preserve not only the geometric specifications of the model but also its semantically encoded metadata, joined to make a "new logical preservation format" for archival purposes. By logical preservation format, the experiment partners in CS19 meant a format encompassing not only the fixed form and content of information representing the model, but also instructions encoded within its metadata in a way that reasoning engines of the future can conduct "proofs" against the object to authenticate it as fit to support the procedural action for which it was designed to be used.

CS26, MOST Satellite Mission

The final report states that the MOST researchers chose file formats based upon best practice; thus, resulting in metadata based upon the file format chosen.

Metadata information in the 23 questions:

4a. What are the key formal elements, attributes, and behaviour (if any) of the digital entities?

CS06, Cybercartographic Atlas of Antarctica

The information expressed is primarily cartographic, according to the functionality of each of the file types below.

Text

- HTML
- XML with XSL style sheets
- Feedback / comment forum or blog
- Databases
- PostgreSQL—open source
- PostGIS (e.g., polygons, etc.)—open source
- Excel spreadsheet (scientific numeric data—e.g., local databases)
- ESRI EOO (e.g., Antarctic Digital database)
- Flat binary (e.g., National Snow and Ice Data Center (NSIDC) at NASA)
- Graphics (e.g., remote sensing data, terrain models, Digital Elevation Models
- (DEM), Radar data, pictures, etc.)
- 2-dimensional—BMP
- 2-dimensional—GIF
- 2-dimensional—JPEG
- 2-dimensional—TIF
- 2-dimensional—PNG
- 2-dimensional—GEOTiff

- 3-dimensional—VRML and the viewer(s) required to access it (e.g., Cortona or other that works with Firefox and Mozilla browsers)

Sound

- OGGVorbis—open source
- WAV
- AIF
- AU
- Moving images
- Quicktime
- MPEG4
- Animation
- SVG (Scalable Vector Graphics)—open standard
- Flash
- Virtual reality fly-through
- VRML or video
- Games
- Online quizz
- Programming languages and technical specifications
- Javascript
- Java
- SVG
- DHTML
- XML (schema files)
- GML (Geographic Markup Language)
- VRML (Virtual Reality Modeling Language)
- Haptics (e.g., a vibrating mouse, shaking chair, force feedback devices)
- Feasible if creator wishes to do so
- Operating System, Middleware
- Linux Redhat Enterprise V4
- Apache Server—open source
- TomCat - Java—open source
- PROJ—open source
- GEOS—open source
- Geoserver—open source
- Deegree—open source
- Java SDK—open source
- XML Libraries—open source
- WFS
- WMS/WCS

For additional details about the digital entities in use, please consult the following report appendices:

- Appendices H & J: Hardware and software lists
- Appendix J: Mime Encoding of Project Software
- Appendix K: List of Data Sources for the CAA

- Appendix M: Atlas Framework, Model and File Types - Freiburg Paper and Presentation

Given the complexity of the CAA, it is not possible to list all the digital components or their individual specifications. Please refer to Figure 2 in the report for a diagram of the overall technical architecture of the CAA.

CS08, Mars Global Surveyor Data Records (NASA)

The PDS data nomenclature standards define the rules for constructing Data Element and Data Object names. A keyword (standard data element name) is an element of the Planetary Science Data Dictionary (PSDD) that defines a named property of an object. The keyword plus its value is an attribute. A label (product label) is a resource description stored in a file. If the label is in the same file as the resource, it is called an attached label. If it is in a separate file, it is called a detached label. Labels also describe the structure or format of the data. Object Description Language (ODL) is used to create labels (data descriptions) for data files and other objects such as software and documents. The PDS labels contain the key attributes of the digital objects. The behaviours of a digital object consist of the various operations that can be performed on the object. For instance, an image object is a regular array of sample values. Image objects are normally processed with special display tools to produce a visual representation of the sample values. This operation on the digital object to produce a visual representation is a behaviour of an image object. Other behaviours of these digital objects consist of the processing and analytic tools that can be used to create other objects, e.g., a tool to produce an image histogram from an image.

CS14, Archaeological Records in a GIS

The final report states that the core data set is represented in both text and numeric characters, while the outputs are textual and graphic in nature (map(s) alongside tabulated data). Furthermore, the process for creating and maintaining these entities is ad hoc.

CS19, Preservation and Authentication of E-Engineering and Manufacturing Records

There were five (5) digital entities in the CS19 engineering experiment. The first two entities listed below are produced during actual computer-aided design (CAD) and computer-aided manufacturing (CAM) processes of the original experiment partner. In the actual business process, these entities are stored with a TIFF rendition of designs as an archival aggregate in a product data management system. They were extended in CS19's engineering experiment by three additional digital entities. Each iteration of format in the experiment was chosen to either strengthen semantic expressiveness or to capture knowledge representation in a persistent, open source format:

1. The original entities (1) are created by product designers using proprietary Pro-Engineer CAD systems and are provided to colleagues charged with computer-aided manufacturing of high-tolerance, high-assurance objects used in complex assemblies. There is no formal definition of this format in the public domain as the file has a proprietary format.
2. Corporate business rules of the original experiment partner ensure that the proprietary CAD design record (1) is translated into (2) Standard for the Exchange of Product Model Data (STEP) AP203 format (ISO 10303). The formal element, attribute and behaviour definition of the objects in the STEP file are contained in ISO 10303 AP 203. The standard describes the formal representation of the Euler complete boundary representation definition of a solid model. The definition of the elements and attributes are described in an object-oriented representation language called EXPRESS that is ISO 10303 Part 21. EXPRESS schemas are computer-processable and can be verified

automatically for syntactical correctness and for the existence of appropriate links to other schemas. Instances of the defined entities form the actual exchanged data. Entity definitions include rules that can be checked at translation time to verify certain aspects of semantic validity of the transferred instances.

3. From there, the experiment took the logical form of this STEP record (2) and enhanced it into another logical form (3) that supported the delineation of additional geometric relationships and reasoning about part shape and action or process semantics using C++ based knowledge representation tools. The derived features and action semantics able to be represented by this format allow for their automated interrogation by reasoning programs, establishing semantic metadata to enable automated archival authentication of the digital solid model.
4. These entities (3) were then taken through a proprietary reasoning engine (Logistica) to complete rendition of a format (4) with all required attributes and metadata, including the formulation of logical predicates. Although the Logistica format is proprietary, it can be said that it contains a knowledge component and a procedural reasoning component.
5. The Logistica entity (4) was converted to Web Ontology Language (OWL) format (5), an open source, public domain XML specification of the World Wide Web Consortium (W3C) for persistent archiving purposes. The OWL form is in ASCII. The logical components of this form are defined mathematically by concepts of descriptive logic, and the syntax of this form is defined by the W3C in the specifications. OWL is a semantic XML format to represent machine interpretable content when the content needs to be processed by applications rather than just structured for presentation to humans. This requirement applies not only to the World Wide Web but to the digital holdings of any given domain within it, including records repositories. OWL can be used to explicitly represent the meaning of terms in vocabularies and the relationships between those terms; in other words, to operationalize an ontology. OWL has more facilities for expressing meaning and semantics than XML, RDF and RDF-S, and thus OWL goes beyond these languages in its ability to represent machine interpretable content.

CS26, MOST Satellite Mission

The key elements are mainly textual, but there are graphic elements as well.

4d. How are the digital entities identified (e.g., is there a [persistent] unique identifier)?

CS06, Cybercartographic Atlas of Antarctica

There are no unique and/or persistent identifiers, and there is no formal ID lookup system.

- The digital objects are identified by a unique combination of a file name and a location in the system
- Some objects are identified in databases, with location information included with other metadata (see Question 22 below).
- There are also some metadata embedded within some digital objects. The modules are associated with metadata. Within a module, metadata are available to reference any entity via a citation.
- Some of the maps will have embedded Geographic Markup Language (GML) to link to and describe related geo-referenced objects, such as images or sounds.
- A multimedia metadata schema is being developed. Some of the elements will be embedded within the information objects themselves and some will be linked to the

object. This will become a part of the Authors' Toolkit, which includes a template of the XML schema that is completed by the content creators.

CS08, Mars Global Surveyor Data Records (NASA)

A product ID data element represents a permanent, unique identifier assigned to a data product by its producer. In the PDS, the value assigned to a product ID must be unique within its dataset. The PDS Standards Reference also specifies the rules for dataset and volume names and IDs. Each PDS dataset must have a unique data set name and unique data set ID, both formed from up to seven components. Within datasets, there are unique volume IDs. Within volumes, the file names are unique.

CS14, Archaeological Records in a GIS

Digital entities are identified through file naming conventions. Aggregations of files within certain folders can also create an associative identity of their own.

CS19, Preservation and Authentication of E-Engineering and Manufacturing Records

In the business activities of the originating experiment partner, digital entities (1) and (2), along with a TIFF version of a solid model design, are stored according to documented company policies in a proprietary product data management system (PDM). The PDM in use is a commercial records management application. This aggregate, termed a "bill of materials," is filed in the PDM according to a numbered schema corresponding with design/manufacturing procedures and there under by project number. Within digital entities (3), (4) and (5), the underlying format allows the assignment of unique identifiers at the file level depending upon business needs. This is especially true of files formatted according to the ISO 10303 STEP AP203 and part 21 EXPRESS metadata schemas, which among their functions support specification of the bond between components in complex mechanical assemblies. It also should be noted that within individual CAD files and the semantic extension formats the representation of each individual attribute or element also has persistent unique identifiers. However, the protocol of the engineering experiment did not require the unique identification of each digital entity, since there was only one instance of each of the five entities. Furthermore, CS19 is founded on the proposition (already operational in the Semantic Web) that simple enumeration of discrete identity and integrity metadata is inadequate to the demands for discovery and authentication facing the future of archives. The conception of intrinsic documentary form needs to go much further into recognizing the characteristic patterns (classes, relationships, constraints) that cohere among and between otherwise static identity attributes.

CS26, MOST Satellite Mission

Digital entities are uniquely identified by file names [managed by 1) primary target (star) and 2) date]. In addition to this, the metadata provide another set of unique identifiers. The report does not explain what these identifiers are.

18b. From what applications do the recordkeeping system(s) inherit or capture all digital entities and the related metadata (e.g., e-mail, tracking systems, workflow system, office system, databases, etc.)?

CS06, Cybercartographic Atlas of Antarctica

The final report states that CAA relies on the XML-tagged content modules for the creation of metadata. CAA content modules are developed by content creators in such a way that the linkages between the information objects, their functionality and associated metadata are

described in an XML document (created within the specified XML project schema), where the markup language indicates what to display. Subversion maintains all code, and all versions of that code are tracked. Subversion is from Tigris.org—an open-source content versioning system (CVS) for use with the most popular operating systems. The Subversion database is backed up regularly.

CS08, Mars Global Surveyor Data Records (NASA)

Instrument measurements are sent as data packets from spacecraft through the Deep Space Network to computers at the Mission Ground Station at JPL (Jet Propulsion Laboratory). Computer workstations of the various project institutions are connected via NASCOM and Ethernet links to a project database (PDB) at JPL. The workstations are used to create standard data products, documentation and index tables. These are packaged into archive volumes and sent to the Science Data Validation Team (SDVT) for validation. The SDVT transfers the archive volumes to the PDS where there is additional validation.

CS14, Archaeological Records in a GIS

The final report states that there is no recordkeeping system external from the applications; therefore, no formal capture activity. There are numerous capture activities within the GIS. Other than other elements of the Microsoft Office Suite, there are no collective capture tools for the information within the GIS. Groups of data are captured temporarily within the GIS application, ArcView while analysis is being conducted, but then is exported to its appropriate areas outside of the GIS application, either from Microsoft Excel or Access files.

CS19, Preservation and Authentication of E-Engineering and Manufacturing Records

In the business activities of the originating experiment partner the digital entities created in the CAD system are captured in the corporate PDM, which is a commercial records management application system (cf. 4d, above). The expression of the experiment digital entities into the final logical preservation format was a process of derivation and extension from both proprietary and open source systems, as detailed in 4a, above. Within the protocol of the CS19 engineering experiment, the digital entities and related metadata were captured by SDSC's Storage Resource Broker and NARA-ERA's Metadata Catalog Management System.

CS26, MOST Satellite Mission

The final report states that there is no formal capture system in place, beyond the tools within Microsoft Windows.

18d. Does the recordkeeping system provide ready access to all relevant digital entities and related metadata?

CS06, Cybercartographic Atlas of Antarctica

The final report states that a "multimedia metadata schema is being developed where some of the elements will be embedded within the information objects themselves and some will be linked to the object and these will become part of the Authors' Toolkit, which includes a template of the XML schema which is completed by the content creators." The ISO19115 metadata standard will be adhered to at the module level.

CS08, Mars Global Surveyor Data Records (NASA)

Presumably, but this is not clarified in the final report.

CS14, Archaeological Records in a GIS

No. As mentioned earlier, the recordkeeping environment is a dispersed and does not provide organized access. The creator is the intermediary between the files when access is needed,

especially because the majority of the files are in the file directory or on the hard drive of the creator.

CS19, Preservation and Authentication of E-Engineering and Manufacturing Records

In the business context of the originating experiment partner the PDM system allows ready access to all digital entities and related metadata. Access is accomplished through standard queries invoked by menu picks by such attributes as procedure number, job, creator, design-change number, design release version number, etc. For the CS19 engineering experiment the SRB and MCAT systems provide a variety of means to access digital entities and any combination of metadata. In addition, the experiment protocol called for the logical querying of the semantic metadata of formats (3), (4) and (5), to authenticate the digital entity's identity, integrity and suitability for the manufacturing process for which it was designed.

CS26, MOST Satellite Mission

The final report reveals that it is possible to access all digital entities via Windows Explorer but does not actually mention how access is provided to the metadata prescribed by the MOST researchers.

18e. Does the recordkeeping system document all actions/transactions that take place in the system re: the digital entities? If so, what are the metadata captured?

CS06, Cybercartographic Atlas of Antarctica

Although this question is not directly applicable to CS06, answers to the following of the 23 research questions provided in the report do touch on this issue.

Question 8: Any digital object that forms part of the CAA must be described by the creator, using metadata standards adopted or developed by the project. See Question 20 in the report and Appendix P, which includes the project's metadata standards. Retrievability of, and access to, the digital objects are based on a number of adopted OGC interoperability specifications (see Appendices P and N in the report).

Question 10: Data are acquired from authoritative sources and are peer-reviewed (e.g., British Antarctic Survey, Scientific Committee on Antarctic Research, scientific and academic journals and books). Each is assessed against the Elements of Spatial Data Quality, which include:

- Lineage
- Positional accuracy
- Attribute/thematic accuracy
- Completeness
- Logical consistency
- Semantic accuracy
- Temporal information

See Appendices U and K in the report for the list of data sources.

Authenticity in geography is captured in standard metadata as data lineage. Quality measures are dependent on the type of data and their function (e.g., the acceptable margin of error for the precise location and size of a particular ice flow to inform tourist ships is smaller than fish counts to inform fisheries and ecological modeling). In addition, each scientific domain

is governed by its particular data quality standards, measures and assurances and these are included in the metadata. Appendix P in the report includes a list of such standards.

Question 13: Changes to the code are captured in Subversion, a source repository system used by the project. Subversion maintains all code, and all versions of that code are tracked. Subversion is from Tigris.org—an open source content versioning system (CVS) for use with the most popular operating system. The Subversion database is backed up regularly. Other digital objects that form part of the CAA are not captured by Subversion.

The Authors' Toolkit will eventually allow changes to associated metadata to be tracked as well. Also see:

- Excerpt—*Elements of geospatial data quality, March 8, 2002*
- *Multimedia Metadata Discussion Document*, December 2003

CS08, Mars Global Surveyor Data Records (NASA)

The PDS logs accesses to restricted areas of the system. User ID, date, time and operation are logged.

CS14, Archaeological Records in a GIS

The report explicitly states that there is no audit trail. The GIS Specialist is in the process of creating metadata relating to the source of the data, including the original author, date or recording, etc.

CS19, Preservation and Authentication of E-Engineering and Manufacturing Records

The PDM system used by the originating research partner in actual business processes captures actions, events and changes to the digital entities (1), (2) and the bill of materials aggregate. Metadata is typically name of creator, release version numbers, date of release, etc. The SRB and MCAT systems captured all changes to the representation of the CAD solid model as it migrated through the semantic format extensions (3), (4) and (5), including the formulation of metadata that support querying by automated reasoning programs.

CS26, MOST Satellite Mission

The final report states that there is no audit trail.

22. What descriptive or other metadata schema or standard are currently being used in the creation, maintenance, use and preservation of the recordkeeping system or environment being studied?

CS06, Cybercartographic Atlas of Antarctica

The final report states that the CAA has solid metadata practices in place; these metadata practices include the following: FGDC and/or British Antarctic Survey DIF (Directory Interchange Format), OGC interoperability specifications and the International Standards Organization 19115 Geomatics Standards. The report also indicates that the ISO 19115 metadata standard for digital mapping data has been explored (see *Multimedia Metadata Discussion Document, December 2003*).

CS08, Mars Global Surveyor Data Records (NASA)

The final report states that the *Planetary Science Data Dictionary* (PSDD) is used in the creation, maintenance, use and preservation of the PDS. The PSDD contains definitions of the standard data element names and objects.

CS14, Archaeological Records in a GIS

The final report states that they are interested in using ArcCatalogue, a metadata tool that is in the new version of ArcView. Their main goal relating to metadata capture surrounds source information relating to CC Database data. The metadata would indicate from what source (publication, repository, Web site, database, etc.) the data was retrieved. In addition, time tagging of georeferenced information is part of the documentation of the processes of creating online digital maps, models and georeferenced visualizations.

CS19, Preservation and Authentication of E-Engineering and Manufacturing Records

The final report states that the first digital entity (1), produced during actual computer-aided design (CAD) and computer-aided manufacturing (CAM) processes of the original experiment partner, originates in a proprietary software tool, thus the precise metadata schema is unavailable. However, the tool produces models in conformance with the ANSI Y-14.5 tolerance standard and provide export files (2) compliant with ISO 10303 Standard for the Exchange of Product Model Data (STEP), AP 203 and part 21 EXPRESS. Corporate metadata standards and procedures govern the filing of these two digital entities with a third TIFF export of the model view into a commercial Product Data Management System. The formats of CS19's digital entities (3) and (4) included the formulation of additional semantic metadata by in-house computer scientists expert in knowledge representation systems that supported the delineation of additional geometric relationships of the CAD solid model and reasoning about part shape and action or process semantics. Although some of the metadata supporting action semantics was lost in the translation to digital entity (5), OWL XML, it was able to persist and authenticate precise specifications about part shapes and relationships, including the classes, predicates and constraint rules that govern the identity and behavior of the CAD solid models.

CS26, MOST Satellite Mission

The metadata schema that is used was created by the MOST researchers and is specific for the data/files that are created in the MOST project. The metadata refer to information such as orbital parameters, observational parameters, telemetry information and target image information. The report notes that some of the metadata/descriptive fields in the FITS files are mandatory, due to the file format. In general, no metadata standards are used; the MOST researchers have created their own scheme of important descriptive fields.

23. What is the source of these descriptive or other metadata schema or standards (institutional conventions, professional body, international standard, individual practice, etc.?)

CS06, Cybercartographic Atlas of Antarctica

The source of metadata comes directly from professional bodies, institutional conventions, as well as international standards. The Atlas adheres to ISO 19115 at the modular level and additional research is ongoing regarding metadata at the granular level.

See response to question 22, above.

- International Standards Organization (ISO)
- Open Geospatial Consortium (OGC)
- Scientific Committee on Antarctic Research (SCAR)
- Geomatics and Cartographic Research Centre (GCRC)
- DIF Format (see http://gcmd.gsfc.nasa.gov/User/difguide/difman.html for details)

The CAA project itself: Y. Zhou, MA thesis on this topic entitled "Profiling and Visualizing Metadata for Multimedia Information in a Geospatial Portal."

CS08, Mars Global Surveyor Data Records (NASA)

The *Planetary Science Data Dictionary* is a NASA institutional standard for Planetary Science Metadata. The PDS procedures for assigning standardized names to data elements follow closely the NBS Guide on Data Entity Naming Conventions.

CS14, Archaeological Records in a GIS

Within ArcCatalogue, the user could create, manage and edit metadata based on the Federal Geographic Data Committee (FGDC) Content Standards for Digital Geospatial Metadata or the ISO 19115 Metadata Standard. These metadata would be stored in XML.

CS19, Preservation and Authentication of E-Engineering and Manufacturing Records

ANSI, ISO, W3C, corporate business rules

CS26, MOST Satellite Mission

The metadata that are used for the various files are based on experience and best practice in the astronomical community and on the foreseeable use of the records in the future. There is an internal MOST document that describes the descriptive fields of the FITS files.

Focus 3. Governmental Activities

General information regarding metadata

CS05, Archives of Ontario Web Exhibits

The final report states that the Ontario government has developed a standard look and feel to which all government Web content must adhere. These are standards created or adopted within the Ontario Public Service. One such standard is the metadata, which refers to title, keyword and description and classification metatags.

CS12, Antarctic Treaty Searchable Database

The final report states this "is an automated technology that objectively integrates digital-record entities without markup, metadata or databases." However, the report further states that "unlike subjective content descriptions in metadata or controlled vocabularies, DIGIN® comprehensively searches both the contents of the granules and their categorical tags to objectively identify those granules that match the search queries. DIGIN® is interoperable with metadata, mark-up and databases."

CS17, New York State DMV On-line Services System

The final report states that the DMV provides a highly interactive online system featuring a complex set of interwoven electronic activities, which collects information about the user via cookies, Web protocols and transactional metadata. A third party digital signature company, VeriSign, is used to make transactions legally binding.

CS18, Alsace-Moselle's Land Registry

The final report states that there are no descriptive or other metadata schemas or standards; however, within the relational database the data are linked together through queries. It has been explained that there is a metadata schema that will be completed for the second phase of the project.

CS20, Revenue On-Line Service (ROS)

The final report states that data mining of ROS-created data is used to audit tax details, improve efficiencies, increase customer service and enable fraud detection.

CS21, Electronic Filing System, Supreme Court of Singapore

The law firm is expected to enter information on their cases through a prescribed documentary template in EFS. Some of the metadata elements are fixed as there is a pull-down menu for law firms to select. Some of the metadata elements the file has to enter include the firm's file reference number and party details, which include the party type, name and address of parties and name of solicitor. EFS captures both the metadata of the record and the actual record itself; the court must check both the metadata and the record.

CS24, City of Vancouver GIS (VanMap)

The interim report states that metadata is not a means of tracking how the information is used, but it does reveal what information is being used and when; this is conducted through the generation of statistical reports.

CS25, Legacoop of Bologna Web Site

The final report explains that all paper mail (what is sent to the organization and what is sent directly to the single functionaries) is registered. And, that the electronic mail sent to the organization official email-system is registered when it is determined that the message is of a certain importance. The registry system uses an automated application to register the records. This application provides a profile of the registered incoming and outgoing documents, including the following: classification code, recipients, object, date and type of document.

Metadata information in the 23 questions:

4a. What are the key formal elements, attributes, and behaviour (if any) of the digital entities?

CS05, Archives of Ontario Web Exhibits

Elements and attributes that are considered integral to the validity and completeness to the Web exhibits Elements were determined based on how the exhibits are normally accessed.

Key intrinsic elements include:

- navigation links from the institutional home page to a listing (with or without a précis of the exhibit);
- exhibit content, normally comprised of Web pages containing text, images, and occasionally with sound or video files;
- government visual identity signs, especially the provincial and city logos and the institutional;
    - Provided by a central body for all Ontario Web sites are:
        - Standard disclaimers
        - Instructions for accessing and installing plug-ins
        - Copyright statements
        - Privacy statements
        - Graphics (.gif format) provided for every ministry name
        - Graphics for mandatory toolbars are provided
        - Ontario logo, mandatory for every government Web page, and footer graphics are provided

      The last three are compliant with the W3C's WAI (Web site Accessibility Initiative) requirements, and all text is provided in English and French.

Key extrinsic elements include:

- A corporate standard Web page template;

- The cascading style sheet created for the Web site as a whole;
- The institutional Web site (contains other exhibits, links to databases, external links, etc.)
- The corporate Web environment (contains links to all government Web sites, news releases, etc.)
- HyperText Markup Language, specification version 4.01;
- Navigation bars required at the top/bottom/side of each Web page
- A "feedback form" that utilizes Common Gateway Interface (CGI) script to interface with an email application

Behaviour of the rendering platform takes place on two levels:

1. the feedback form is a CGI program executed in real-time; and
2. the way the user's browser interacts with the HTML coding of the exhibits.

CS12, Antarctic Treaty Searchable Database

Indeterminate from answer provided.

CS17, New York State DMV On-line Services System

The records are live records and have the ability to change over time. They can be placed into a status where they are no longer alterable, as when a driver dies or a vehicle is junked. The official records contain an official crest or logo.

CS18, Alsace-Moselle's Land Registry

The ordonnances take the form of XML files, containing tagged information relative to landowners, land parcels and rights/obligations relative to the property. Associated with the ordonnance is a digital signature of the judge authoring the ordonnance. The structure of the ordonnance is defined through a DTD. The scanned images of the register take the form of TIFF files.

CS20, Revenue On-Line Service (ROS)

There are some elements and attributes common to the three "classes" of records [digital certificates/signatures, tax forms and debit instructional forms] for presentation, Revenue logo, font and style, certification practice statement, privacy policy, terms and conditions, copyright statements, contact details and standard Web page templates.

CS21, Electronic Filing System, Supreme Court of Singapore

The EFS is composed of standardized HTML style sheets, XML files, Visual Basic, PDF and graphic files for the EFS logo.

CS24, City of Vancouver GIS (VanMap)

The data sheets describe VanMap's data layers, features and functionalities; each layer typically contains, layer name, group name, scale, data currency status, responsible department and definition.

CS25, Legacoop of Bologna Web Site

All the entities have at least a title and a body text and a date. Each element is numbered sequentially according to the chronological order.

4d. How are the digital entities identified (e.g., is there a [persistent] unique identifier)?

CS05, Archives of Ontario Web Exhibits

Each Web exhibit is identified by its title and a URL, which has been assigned within the institution's Web domain. When viewing the source coding for each Web page within each exhibit, each page is also titled.

CS12, Antarctic Treaty Searchable Database

Each of the information granules or digital-record entities in the current *Database* contains unique provenance information in a categorical header tag as well as in the title. Unlike metadata, which are stored in repositories separately from the digital entities, the unique identifiers are part of each granule in the *Database*. Thus, with the categorical header tags, there is never a risk for decoupling the unique identifiers and the digital entities.

CS17, New York State DMV On-line Services System

There is a unique identifier connected to each transaction. The transaction and its identifier are stored with the core record, as a result of the transaction. Different sets of identifiers exist for each of the three file types: license, registration, and title.

CS18, Alsace-Moselle's Land Registry

Every inscription in the database is numbered with a persistent, unique identifier and dated. Ordonnances are also numbered and dated. Each scanned image of the registers is numbered according to the system already in place for numbering individual pages of the registers.

CS20, Revenue On-Line Service (ROS)

There is no need for specialized codes and keys beyond those normally used by the Revenue.

CS21, Electronic Filing System, Supreme Court of Singapore

The case number is a unique identifier, which is automatic generated number assigned by the courts.

CS24, City of Vancouver GIS (VanMap)

The HTML and CML pages and embedded GIF images are identified by unique URLs. The data fields, layers and groups are also identified by field names, layer names and group names.

CS25, Legacoop of Bologna Web Site

A primary key in the form of a progressive number (managed as a key field in the database) is the main identification attribute.

4e. In the organization of the digital entities, what kind of aggregation levels exist, if any?

CS05, Archives of Ontario Web Exhibits

The Web exhibits and the Web pages reflect aggregations of text, images and other components of the exhibit which are conceptually linked. The institutional Web sites and the Web exhibits are grouped together for the navigational convenience of the user.

CS12, Antarctic Treaty Searchable Database

The aggregation levels among digital entities are based on the inherent parent-child relationships within the policy documents. In general, the aggregation levels or hierarchy levels reflect the granularity of a digital collection. This collection granularity is represented specifically by:

- Antarctic Treaty Searchable Database > Year > Major Document or Antarctic Treaty Consultative Meeting > "measures"

Dynamic aggregation of digital-record entities with DIGIN® facilitates the discovery of relationships within and between the digital-record series.

CS17, New York State DMV On-line Services System

The DMV does not use directories or subdirectories, but keeps everything in tables and databases. The individual transactions in the audit trail are organized by date and time, category and current status.

CS18, Alsace-Moselle's Land Registry

The database aggregates the data according to the main categories: parcels, persons, rights and obligations. The presentation of data is organized in the same way as the paper register; that is, a single *feuillet* contains information relative to all the properties of a person within a given administrative territory (usually, a commune, or part thereof).

CS20, Revenue On-Line Service (ROS)

All tax records and debit instructions are saved chronologically and are viewable within the Revenue Customer Information Service. They can be sorted and viewed depending upon the field type selected. Regarding digital certificates and signatures: Metadata surrounding the older Digital Certificates, in addition to the security wrapper, are maintained with ROS. Revenue has a separate Archiving Policy for Certificates, but this is considered beyond the remit of ROS.

CS21, Electronic Filing System, Supreme Court of Singapore

The main case files are divided onto various sub-folders based on the type and nature of records filed, such as affidavit, draft order, minute sheet and summon in chambers.

CS24, City of Vancouver GIS (VanMap)

HTML and related pages are grouped into folders for storage, identification and retrieval purposes. The data are organized into layers, with each layer including a single data source or multiple data sources.

CS25, Legacoop of Bologna Web Site

The entities are aggregated according to the main logical categories of the Web site (documents of the association, news from the cooperation world, CVs and announcements, other services related to the Bologna business area).

18b. From what applications do the recordkeeping system(s) inherit or capture all digital entities and the related metadata (e.g., e-mail, tracking systems, workflow system, office system, databases, etc.)?

CS05, Archives of Ontario Web Exhibits

The exhibits are created using Dreamweaver and PageMaker software applications.

Metadata captured would normally be what are automatically captured by the default settings of those applications. None of the interviewees commented that they used the document properties function to add any specific metadata. Metadata captured would normally be what are automatically captured by the default settings of the applications used to create supporting documentation, such as Microsoft Word.

CS12, Antarctic Treaty Searchable Database

According to the final report, this question is irrelevant since, after the initial implementation of the *Database* in 1999, the only captured files are the entire Antarctic Treaty Consultative Meeting (ATCM) Final Reports without metadata that have been published on the ATCM Web sites of the host nations. The new *"measures"* that have been adopted by the Antarctic Treaty Consultative Parties are then extracted and added to the *Database* with header tags that define their unique location in the overall collection.

CS17, New York State DMV On-line Services System

The system that the DMV uses captures IP addresses, system dates and session cookies. The cookies are used only to maintain the session state; they are not stored on the hard drive of the patron.

CS18, Alsace-Moselle's Land Registry

Both ordinances and inscriptions are captured through custom applications. The scanned images of the register were captured once at the onset of the computerization process.

CS20, Revenue On-Line Service (ROS)

Databases and other systems – ITP is held on an Ingress II mainframe back-end system.
ITP – Integrated Taxation Processing [System]

CS21, Electronic Filing System, Supreme Court of Singapore

The EFS captures digital entities from an oracle database, Filenet (document management system), jukbox (for CDs) and visual basic software. See answer to question 5a in final report for a detailed list of the application systems.

CS24, City of Vancouver GIS (VanMap)

DOMINO, PRISM, License and other systems produce the data.

CS25, Legacoop of Bologna Web Site

The report states that the recordkeeping system does not have any relationships with the Web site system.

18d. Does the recordkeeping system provide ready access to all relevant digital entities and related metadata?

CS05, Archives of Ontario Web Exhibits

The report states that related metadata is not readily accessible, even if it has been captured. This is due to the absence of a recordkeeping system and lack of consistent recordkeeping processes around the provision of access to Web exhibits within the two institutions.

CS12, Antarctic Treaty Searchable Database

The final report says yes, through providing comprehensive integrated access to the digital-record entities. And, that the Antarctic Treaty Searchable Databases does not require metadata.

CS17, New York State DMV On-line Services System

If the mainframe system is equivalent to the recordkeeping system, then the answer to this question is yes. Although customers and third party users have access to only a small portion of the digital entities, the system provides DMV personnel with access to all aspects of the digital entities.

CS18, Alsace-Moselle's Land Registry

Yes, as relevant for each category of user.

CS20, Revenue On-Line Service (ROS)

Yes, to both Revenue employees and ROS users. Not all users will view metadata.

CS21, Electronic Filing System, Supreme Court of Singapore

Yes, as relevant for each category of user. Authorized court users can view both the record profile and the PDF record. Specified group of users can, based on their job competency, view certain categories of audit logs.

CS24, City of Vancouver GIS (VanMap)

Metadata in the form of data sheets is also readily available. This study has not yet investigated the types of metadata that may be generated automatically by the various technological processes used to create VanMap.

CS25, Legacoop of Bologna Web Site

Not applicable.

18e. Does the recordkeeping system document all actions/transactions that take place in the system re: the digital entities? If so, what are the metadata captured?

CS05, Archives of Ontario Web Exhibits

The Web logging software documents aspects of all interactions with the institution's Web site. The report presents 21 reports generated by *Analog* based on the data it gathers. Please refer to page 46 of the report for this list.

CS12, Antarctic Treaty Searchable Database

The report states metadata are not captured. However, the report also states that all queries of the Web site version of the *Database* are automatically logged.

CS17, New York State DMV On-line Services System

The recordkeeping system at the DMV tracks all changes to records in the mainframe system through audit trails and user logs.

CS18, Alsace-Moselle's Land Registry

The report reveals that the system includes extensive login capabilities for recording all actions and transactions taking place in the system. Logs may be used for two distinct purposes.

CS20, Revenue On-Line Service (ROS)

The report states that all changes are noted and logged with time/date stamp and name of Revenue employee making change.

CS21, Electronic Filing System, Supreme Court of Singapore

Yes, all actions and transactions are documented in various audit logs, including:

- Transaction Log: Records user ID of user who activates the change, function name, date./time of the change, data items before and after the change
- Financial Audit Log: Records user ID, function name, date/time of the action, case number/document, control number/unique reference number, amount of fees before the change, amount of fees after the change, remarks, approval for exemption/waiver of court filing fees, approval for request of waiver of hearing fees, and approval for refund of hearing fees.
- Violation Log: Records user ID of user who attempts to access functions he or she is not granted access to, unsuccessful log in attempts, function name, date/time of the action, and brief description of the nature of the violation.

CS24, City of Vancouver GIS (VanMap)

The interim report states that the use of the data can be tracked by unique client IDs randomly generated when users download the MapGuide ActiveX Viewer to their workstations. For example, whenever a user accesses VanMap and issues a request for data the transaction results in a log file record containing his or her ID, the date and time of access and strings of numbers representing specific data layers used.

CS25, Legacoop of Bologna Web Site

Not applicable.

22. What descriptive or other metadata schema or standard are currently being used in the creation, maintenance, use and preservation of the recordkeeping system or environment being studied?

CS05, Archives of Ontario Web Exhibits

The Web site coordinator was unfamiliar with metatags, and initially ignored metadata standards. Metatags are only applied to "key pages"; therefore, is nothing that distinguishes an exhibit page from any other page on the Web site. The report states that only comprehensive source of metadata governing the entirety of an exhibit appears to be the "definition document" created for *The War of 1812* exhibit. This document includes the title, reference code, image number (where applicable), location information and a summary of the document/image.

CS12, Antarctic Treaty Searchable Database

Descriptive metadata, as conventionally applied with templates and attributes that reside in repositories, are not used to implement the Antarctic Treaty Searchable Database. However, there is the use of header tags that describe the parent-child provenance. Also, conventional metadata regarding the portal for the Antarctic Treaty Searchable Database are being added to the National Science Digital Library and Digital Library for Earth System Education. The metadata format for these submissions is a modified Dublin-Core metadata with additional fields for the education audiences that are being addressed by these digital libraries.

CS17, New York State DMV On-line Services System

Although data layouts and schema are used, the DMV respondents indicated that they did not feel comfortable revealing specifics about such information to the InterPARES research team.

CS18, Alsace-Moselle's Land Registry

No descriptive nor metadata schema is currently being used.

CS20, Revenue On-Line Service (ROS)

Twenty-two schemas for the tax forms have been made available in XML DTDs for inclusion in the third party compatible software; view www.ros.ie/downloads.html and Appendix IV. All schemas include a DTD and element definitions and explanations. Although an Irish Public Service Metadata standard exists, it is not used with ROS.

CS21, Electronic Filing System, Supreme Court of Singapore

The schemas for the documentary templates of the EFS are based on the workflow and juridical requirements of the court.

CS24, City of Vancouver GIS (VanMap)

The metadata applied by the VanMap developers include layer name; group name; scale at which the data become available; data currency status; responsible department, branch or division; and definition. Not all of these metadata are applied to all of the data layers. Metadata generated automatically upon creation of the data have not yet been investigated.

CS25, Legacoop of Bologna Web Site

Basic metadata related to the registry system are required in the recordkeeping system, but not exported to the Web site management, which handles only a numbering system for each digital entity and a date.

23. What is the source of these descriptive or other metadata schema or standards (institutional conventions, professional body, international standard, individual practice, etc.?)

CS05, Archives of Ontario Web Exhibits

There is no identified source for the Government of Ontario Category Metadata rules. The City's Web coordinator stated that the metadata tags he uses do not conform to any standards.

CS12, Antarctic Treaty Searchable Database

The final report states that conventional metadata are unnecessary with the DIGIN® technologies, which can interoperate with or without metadata to integrate "structured" as well as "unstructured" information. The sources of the descriptive schema are the persistent digital-record entities themselves.

CS17, New York State DMV On-line Services System

The source of these standards was not mentioned or discussed during the case study interview.

CS18, Alsace-Moselle's Land Registry

No descriptive nor metadata schema is currently being used.

CS20, Revenue On-Line Service (ROS)

Institutional practice. Form design and structure is based on existing paper-based forms. Field selection and management is based on requirements and format of ITP applications and data flow to this and other back-end systems. The XML schemas may include other descriptive standards such as ISO Year Standard.

CS21, Electronic Filing System, Supreme Court of Singapore

Institutional practice. The metadata used in the documentary template are based on common data elements associated with the court records that have to be converted into PDF.

CS24, City of Vancouver GIS (VanMap)

Metadata are based on what the VanMap Team thinks will be useful information for the end user.

CS25, Legacoop of Bologna Web Site

The metadata included are strictly related to the professional standard followed for building the Web site (SQL for the database and HTML for the Web site pages).