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

REPORT ON THE WORK AND FINDINGS OF THE INTERPARES 2 DESCRIPTION CROSS DOMAIN GROUP

Anne Gilliland, Joanne Evans, Alison Langmead, Monique Leahey-Sugimoto, Lori Lindberg, Victoria McCargar, Joe Tennis, Holly Wang

1. INTRODUCTION

Metadata that is associated with either an information system or an information object for purposes of description, administration, legal requirements, technical functionality, use and usage, and preservation, plays a critical role in ensuring the creation, management, preservation and use and re-use of trustworthymaterials, including records. Recordkeeping[1] metadata, of which one key type is archival description, plays a particularly important role in documenting the reliability and authenticity of records and recordkeeping systemsas 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 is also the means by which it is possible to identify how record components – those constituent aspects of a digital record that may 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.

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, regardless of the community, sector or institution within which they are being created. The InterPARES 2 Description Cross-Domain Group (DCD) examined the conceptualization, definitions, roles, and current functionality of metadata and archival description in terms of requirements generated by InterPARES 1[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 InterPARES2, the Description Cross-Domain also addressed its research goals with reference to wider thinking about and developments in recordkeeping and metadata.

---

[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 life cycle 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 life cycle model, "record keeping" refers to the phase in the life cycle that comes between "record creation" and "record preservation."

[2] The members of the Description Cross-Domain Group were Dr. Martine Cardin, Lavalle University, Quebec; Chia-Ning Chang, University of British Columbia, Vancouver (doctoral student); Terry Eastwood, University of British Columbia (Chair, 2005-2006), Vancouver; Joanne Evans, Monash University, Melbourne (doctoral student); Michael Garabedian, University of California, Los Angeles (master's student); David Gibbs, University of California, Los Angeles (master's student); Anne Gilliland, University of California, Los Angeles (Co-Chair 2001-2005); Hans Hofman, National Archives of the Netherlands; Dr. Alison Langmead, University of California, Los Angeles (master's student); Tracy Luairault, Carleton University, Ottawa (doctoral student); Monique Leahey-Sugimoto, University of California, Los Angeles (former master's student); Lori Lindberg, San Jose State University and University of California, Los Angeles (doctoral student); Dr. Richard Marciano, San Diego Supercomputer Center; Victoria McCargar, consultant, Los Angeles; Dr. Sue McKemmish, Monash University, Melbourne (2001-2005); Randy Preston, University of British Columbia, Vancouver; Nadav Rouche, University of California, Los Angeles (master's student); Stuart Sugarman, University of California, Los Angeles (master's student); Shannon Supple, University of California, Berkeley (former UCLA master's student); Dr. Joe Tennis, University of British Columbia, Vancouver; Dr. James Turner, Universitė de Montrėal; Holly Wang, University of California, Los Angeles (former master's student); Eunha Youn, University of California, Los Angeles (doctoral student); and Yuchai Zhou, Carleton University, Ottawa (master's student).

InterPARES2 addressed not only records, however, but a range of digital information objects (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 government, scientific and artistic 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 Groups. 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 be considered to be or treated as records by the creator. "Potential records" is a new construct that indicates that a 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 Group, 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 DCD sought to name and type categories of metadata. It also addressed incentives for creators to generate appropriate metadata, as well as issues associated with the retention, maintenance and eventual disposition of the metadata that aggregates around digital entities over time.

2. RESEARCH QUESTIONS

Metadata investigations in the digital environment tends to cover a lot of territory, and the scope of the DCD 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[3] in records creation, control, maintenance, appraisal, preservation, and use in traditional record-keeping 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 record-keeping 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?

[3] Used throughout to refer to metadata in the broadest sense, as well as archival description specifically.

- 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 geo-spatial data community) support and inform requirements such as those developed by InterPARES 1? 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 questions came from the projects funded by the United States National Science Foundation and the National Historical Publications and Records Commission that supported the US Teasm's participation in InterPARES 2:

Formulation and testing of technological, metadata, and policy models in order to formulate, analyze and test:

- new and existing methodologies and strategies for ensuring that records created using these 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 should also develop hypotheses of metadata necessary for prototype systems; and rules for the ongoing description of electronic records.

In the course of its work, the DCD 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., life cycle, 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 labeling of the models being developed.

3. RESEARCH DESIGN AND METHODS

Multiple, inter-dependent activities and associated methods were used to generate products and data that could be triangulated in order to answer the research questions (the researchers primarily engaged in each activity are indicated in parentheses).

- Collecting and compiling data on the types and sources of metadata currently being used in real-life settings as identified through case studies being conducted in other InterPARES 2 groups. Method used: *case studies* (Focus Group case study researchers, UBC project staff).

- Conducting a special case study 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 is to be integrated into recommendations made by the Description and other InterPARES2 Groups with regard to evaluating, extending or revising existing descriptive and metadata schemas as well as promoting the Metadata Specification Model. Method used: *literary warrant analysis* (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, as well as identifies existing cross-walks between them. Methods used: *iterative systems design* (Gilliland, McKemmish, Hofman, Marciano, Lindberg, Evans, Rouche, Wang, Leahey-Sugimoto, Langmead, Zhou).

- 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)Methods used: *requirements operationalisation, warrant analysis, schema analysis, metadata mapping* (Gilliland, McKemmish, Hofman, Marciano, Lindberg, Evans, Rouche, Wang, Leahey-Sugimoto, Langmeade, Youn).

- Developing metadata specifications relating to the activity models developed by the Modeling Group identifying the type, source and application of metadata identified in the models and when, how and by whom it should be created.[4]. These specifications can also form the basis for developing a set of specifications for automated tools (not to be confused with descriptive instruments) that can be used to assist with the creation, capture, management and preservation of essential metadata for active and preserved records. Method used: *modeling and empirical instantiations* (Tennis, Eastwood and Preston).

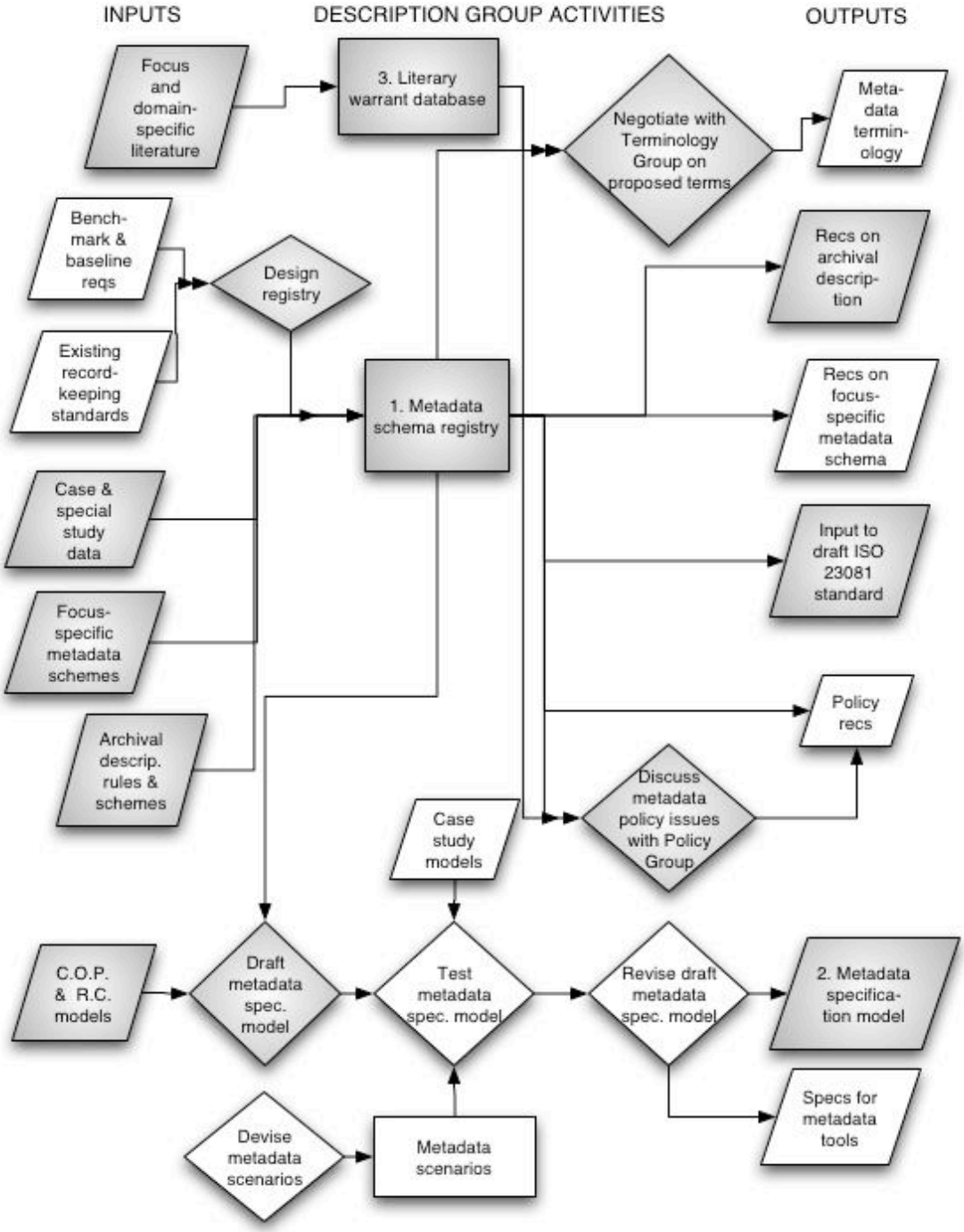
- Interfacing with other relevant research and development activities such as the development of the ISO 23081 Records Management Metadata Standard, the Monash University-based Clever Recordkeeping Metadata Project[5] and the work of the San Diego Supercomputer Center on the development of metadata tools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

Figure 1 indicates the relationships between the constituent components and some of the associated activities of the Description Cross-Domain Group. Numbers 1-3 on the flowchart indicate the primary locuses of activity and eventual products:

[4] The metadata specification model for the Business Process Model also developed by the InterPARES2 Modeling Group has still to be developed.

[5] See Create Once, Use Many Times - The Clever Use of Metadata in eGovernment and eBusiness, available http://www.sims.monash.edu.au/research/rcrg/research/crm/ .

Figure 1. Flowchart of Description Group Actvities



4. ACTIVITIES UNDERTAKEN AND RESULTS PRODUCED

4.a. The Metadata and Archival Description Registry and Analysis System (MADRAS)[6]

MADRAS was initially envisioned as a metadata registry by which the DCD could identify relevant metadata sets and schemas that the group wished to evaluate in order to generate recommendations in response to its research questions. However, it quickly became clear that if the DCD operated on the assumption that metadata was essential to the creation of reliable and preservation of authentic records in electronic systems, then it also needed to address issues associated with how trustworthy metadata is created and maintained. It was also clear that the DCD needed to operationalise 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 US research funds until June 2007, it is now envisaged that the beta production version completed in InterPARES 2 and used by us to answer our 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.

The purpose of MADRAS is fourfold:

a. To support the unambiguous registration of relevant metadata schemas, sets and application profiles;
b. To support the analysis of registered items against requirements derived from the InterPARES1 Benchmark and Baseline Requirements as well as 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.
c. 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.
d. To generate analytical data to be provided to the working group (ISO TC46/SC11-WG1)

---

[6] For further details on the development of MADRAS, see Gilliland, Anne J., Nadav Rouche, Joanne Evans, and Lori Lindberg, "Towards a Twenty-First Century Metadata Infrastructure Supporting the Creation, Preservation and Use of Trustworthy Records: Developing the InterPARES2 Metadata Schema Registry," *Archival Science* (2005): 43-78; Evans, Joanne and Nadav Rouche, "Utilizing Systems Development Methods in Archival Systems Research: Building a Metadata Schema Registry," *Archival Science* 4 nos 3-4 (December 2004):315-334; Evans, Joanne and Lori Lindberg, 'Describing and Analysing 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 Scientific and Technological Literature Publishing House, Shanghai, China, 2004, pp. 75-80; Gilliland-Swetland, Anne J. 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 Lindberg, Lori, 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," *Proceedings of the Third IS&T Archiving Conference* (Springfield, VA: Society for Imaging Science and Technology, 2006).

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 studies undertaken by the Focus Groups.
e. Other relevant Focus-specific metadata schemas and sets identified by Focus Groups or by the Description Group (e.g., GIS metadata standards).
f. Archival description rules, sets, and related practices (e.g., ISAD(G)/ISAAR, EAD/EAC/DACS, RAD, and the Australian Series System).

4.a.1. MADRAS Registry Component

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 recordkeeping 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 that metadata and 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 these both of these concerns[9].

---

[7] See Kate Cumming, Ph.D. Thesis, Monash University 2005.
[8] Chris Hurley, "Abandoned Children to Zoos," *Archivaria* 40, Fall 1995. Available: http://www.sims.monash.edu.au/research/rcrg/publications/ambientf.htm
[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

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 DCD 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 assess a schema they were using or planned to use against InterPARES requirements. 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 DCD 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, with some modifications. Later, we decided also to address the requirements contained in the ISO 23081 Standard as the current international standard, so that users could both assess their schemas and compare differences between requirements for InterPARES and ISO. We could also use the tool to run our own tests based upon selective schema input.

- 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. Why a DTD and not a schema? We decided to move ahead with the analysis component of MADRAS separately using a spreadsheet-based worksheet to allow for parallel work activities while the system was still being built. An automated form of the analysis worksheet used in the analytical framework was subsequently integrated into the DTD. The final step in MADRAS development was designing the prototype public interface and evaluative report forms.

- In order to develop the DTD, we examined how metadata should themselves be sourced in order to ensure their reliability and authenticity, e.g., through recordkeeping requirements for metadata registries described in the ISO standard for metadata registries[10].

- Once we had developed the DTD, we identified multiple key metadata and descriptive schemas and sets (both archival and from those in use in sectors within the three InterPARES2 focus areas) and registered them in the database in order to test and refine it.

- Guidelines for registering, describing and analyzing schemas were developed and refined as our 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.

- Documentation of system functionality and requirements was developed to support the ability to maintain the system and facilitate eventually transferring it from UCLA where it was developed to a maintenance agency.

---

archivists give to the records and recordkeeping metadata of others in order to assess and validate their management and reliability, they must also give to their own.

[10] ISO/IEC 1179 Information Technology – Metadata Registry (MDR).

Example:[11]
In MADRAS a record is created for a version of a metadata standard. So for example there are separate MADRAS entries for Version 1.2 of the Metadata Encoding and Transmission Standard (METS) and for Version 1.3.[12]. Statements in the IDENTIFICATION, DESCRIPTION and RIGHTS sections describe the version of the metadata standard, whilst statements in the DOCUMENTATION, PROVENANCE and RELATIONSHIPS sections capture relationships to other versions of standards and to other kinds of entities. The ADMINISTRATION section captures statements relating to the registry description itself.

The registration process involves giving the metadata record for the standard a unique identity within the schema registry (ID) and completing as much of the identity details as possible. The identity details include:

UGI	The unique global identifier – made up of the domain and the identifier
OFFICIALNAME	The official name of the metadata standard.
ALTERNATENAME	Any alternate names under which the standard is known. This is repeatable with an attribute to specify the type and language of the alternate name.
VERSION	The version of the standard, as used in the documentation.
STATUS	The status of the version – e.g. whether draft, beta, etc. as used in the documentation
PUBLICATION	A publication statement incorporating the publisher, place, date, and if relevant the date the version was superseded.
DATERANGE	The date range for the validity of the schema

Encodings of versions of standards are identified in the documentation section. The metadata captured about the encoding is therefore based on citing the encoding as a reference. It includes:-

REFTYPE	The type of reference e.g. DTD, XML Schema
REFTITLE	Its title e.g. METS Schema 1.3
REFAUTHOR	The author
REFIMPRINT	An imprint statement incorporating place, publisher
DATE	The date of publication
LANGUAGE	The language
EXTPTR	Its url

The registration process of MADRAS has been conceived as one that involves a person entering values into templates. Population of the prototype database showed the wide variation in the way schemas are published and information about them is presented. In such circumstances manual processes involving human cognition, collation and data entry become the only viable registration method. With the lack of standardized ways of publishing the required meta-information, no automated, or even semi-automated, processing could be considered. The best 'technology' to translate the existing meta-information into the MADRAS schema is a person, as they can negotiate the situation specific mappings and deal with the gaps and ambiguities.

[11] Taken from Joanne Evans, draft doctoral dissertation, appendix, *Guidelines for Analysis of Metadata Schemas v4-1*.
[12] The MADRAS entry for METS 1.2 is available at http://www.gseis.ucla.edu/us-interpares/madras/summary_view_public.php?id=8 . The MADRAS entry for METS 1.3 is available at http://www.gseis.ucla.edu/us-interpares/madras/summary_view_public.php?id=7. (Accessed 15 June 2006)

However this also introduces scalability and sustainability issues with 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. This raises the question of what meta-information should be made available as part of the publication of metadata standards, for the consumption of what types of agents?

The current beta environment for MADRAS is implemented with PHP, a server-side scripting language that provides web developers tools for building dynamic websites. 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. The Education Technology Unit (ETU) from the Graduate School of Education and Information Studies at UCLA is hosting MADRAS and provides server-side support. 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. We 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.)

4.a.2. 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. While 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. The Benchmark Requirements are based on the notion of a trusted record-keeping system. They include requirements that support the presumption of the authenticity of electronic 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 electronic 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 in the InterpARES 1 Authenticity Task Force Report,

> 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'.[i] 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. [p. 3]
>
> 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 operationalised 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 InterPARES1 be manifested or documented through metadata? One way of addressing this problem 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 in order to identify their constituent metadata manifestations.[15]

Accordingly, the development of the analytical framework to be used in MADRAS sought to operationalise 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, DCD researchers analysed 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

---

[13] Report of the Authenticity Task Force of the InterPARES Project. Available: http://www.interpares.org.

[14] InterPARES 1 Project. *The Long-term Preservation of Authentic Electronic Records: Findings of the InterPARES Project*. Available: http://www.interpares.org/book/index.cfm

[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 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.

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 generally available, anyone will be able to 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.

In order to draw on as many perspectives as possible and to try to identify where there might be consensus or divergences about relevant recordkeeping requirements (especially where there might appear to be differing view points emerging from the life cycles 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) specifying requirements for Electronic Records Management Systems (ERMS).

The requirements were expressed in the form of evaluative questions, and the questions were 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 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 (see APPENDIX A for Analysis Questions). 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 in order to determine their feasibility, their granularity, their usefulness and 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 view of what metadata is 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. We employed an iterative development process, focused on refining and arranging questions. We 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.

As mentioned above, the instantiations also provided an interesting commentary on the status of metadata schema publication and documentation practices. It raised issues as to 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.

We decided that the first iteration of the system would be for InterPARES' researchers and then the system should be revised for future use by the general public. The analysis worksheet underwent a number of versions and changes through the initial testing and validating that resulted in a final accounting of 4 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, we used FileMaker Pro to work up a model for the display of information in MADRAS that was eventually recreated in the actual MADRAS system.

Challenges in the development of the analytical framework ranged in complexity. Often, we returned to first principles. For example, during the process of creating MADRAS, we were asked to come to an agreement (or not) on the meaning of the word "record". What do we consider a record? A relationship? Along the same lines, we found we needed to consider what our base unit of analysis should be (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.

We did experiment with developing various versions of a decision tree, but, lacking consensus, decided not to use any of them in the current production 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 developing a mapping between related InterPARES and ISO requirements; the development of controlled vocabularies for classification of 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/around records and related metadata, 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 some of the metadata relate to the recordkeeping processes that manage the record.

The 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 similar questions that sought similar answers were asked over more than one section of the spreadsheet as a result. The decision was made to map each of the reference instruments against one another in order 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 is created to satisfy one of seven particular purposes:

- Unique identification
- Authentication of records
- 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
- Administering terms and conditions of access and disposal
- Tracking and documenting use history, including recordkeeping and archiving processes
- Enabling discovery, retrieval and delivery for authorized users
- Restricting unauthorized use

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 we could ensure consistency across different analysts. We 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. We had to assume certain pre-existing knowledge on the part of the user to the system. It was decided that users would most likely be experienced recordkeepers or those familiar with archival terminology.

During the automation process, we assessed the strengths and weaknesses of the original analysis spreadsheets in order to clarify and bolster the effectiveness of MADRAS. The issues found with the original spreadsheets included that:

- The original worksheet facilitated documenting rather than analyzing a metadata schema. (What we tried to do: focus on analyzing rather than on documenting the schema)

- The original worksheet was repetitious. Information documented in one section was repeated in another. (What we tried to do: eliminate redundancy)

- The original worksheet and evaluation instruments had confusing language. (What we tried to do: simplify and add documentation (for example, we created a definition file which 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 database/online worksheet. (What we did: created 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. (What we tried to do: create a system where we could eliminate ambiguity as much as possible)

- The original analysis process did not allow for the discovery of other types of metadata which might be present in the schema but not in any of the analysis instruments. (It was not possible to address this as the tool was so strongly focused on these particular IP and ISO instruments), and

- While the original analysis process asked for repeatability and obligation value for each element, the Excel worksheet did not ask for this information. (What we tried to do: Separated out the repeatability (or lack thereof) of fields as well as whether a field a is mandatory into the element registration process. We also hope to use this information in future implementations.)

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

We attempted to confront the issue of how one separates what is explicitly stated in schema documentation and what is implicit, since we 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."

4.a.3. 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 traveled 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 we revised the framework somewhat to be more entity-focused.

MADRAS has been developed and constructed by researchers of varying knowledge levels regarding records and recordkeeping and from disparate recordkeeping philosophies. 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 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, the assessment tool developed for MADRAS is planned to be the Part III assessment tool for ISO 23081. This assessment tool must accommodate both of

the major models of records management currently in use in the archives and records management communities, the life cycle model as reflected in the InterPARES research and the continuum model developed in Australia[16].

4.a.4. 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 area. These include METS, the Metadata Encoding and Transmission Standard; the Australian Recordkeeping Metadata Schema; the New South Wales Recordkeeping Metadata Standard; the Recordkeeping Metadata Standard for Commonwealth Agencies; the South Australian Recordkeeping Metadata Standard; the ERS (Victorian Electronic Records Strategy) Metadata Schema; the Record Keeping Metadata Requirements for the Government of Canada; the Arizona Electronic Recordkeeping Systems (ERS) Guidelines - IV Functional Requirements for Recordkeeping Systems; the Minnesota Recordkeeping Metadata Standard; the PERM Preservation Attributes; GILS, ISO 82045-2 Document Management Metadata; the CEDARS metadata specification for preservation; MARC; XrML; Open Digital Rights Language (ODRL); Digital Rights Expression Languages (DREL), Online Information Exchange (ONIX); Preservation Metadata - Networked European Deposit Library (NEDLIB) Metadata for Long Term Preservation; NLA Pandora Metadata Element set; NISO Z39.87-2002 AIM 20-2002 Data Dictionary – Technical Metadata for Still Images, Metadata for Images in XML (MIX); a range of geospatial metadata standards; and the PREMIS metadata set[17].

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.
- Schemas that had a very large number of elements were given a lower priority.

We 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 did have a large number of schemas *and* that had elements categorized according to a schema, we decided that we would enter just the name of the element container and specify the element that satisfied the condition in a note field. In future, we would like to add an “import” function to collect this data automatically from electronic versions of the schemas instead of having to do it all by hand.

Schemas were then selected based on their type. Schemas tagged 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 we had such a high number of schemas for government recordkeeping, we also tried to prioritize by sector. Schemas relating to the InterPARES focus areas such as the

[16] See Lindberg, Lori, 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,” *Proceedings of the Third IS&T Archiving Conference* (Springfield, VA: Society for Imaging Science and Technology, 2006).

[17] Several InterPARES researchers, in particular Victoria McCargar, were involved with the development of the PREMIS metadata set, which occurred concurrent with the work of InterPARES 2.

arts or geospatial applications were given a higher priority than others.

One thing we noticed during the prioritization process was that *all* of the schemas were from English speaking countries, apart from one we identified that was developed in China. We think it would be interesting to try to find more schemas developed by non-Anglo communities and try to analyze those. We also noted that among our schemas, we did not have a wide variety by domain/sector registered in the system. Schemas for the legal or medical fields were not represented, for example. We think it would be useful to get a sampling of these schemas for comparison, especially to see what other information could be revealed.

Other considerations included weighing what we would gain from analyzing schemas that were not developed specifically for recordkeeping purposes. How do they differ? Are the schemas that were not designed for recordkeeping purposes all necessarily faulty when examined in light of recordkeeping requirements? Might they include elements we had not previously considered that might be useful for recordkeeping purposes?

- We tested, cross-validated and revised the analytical framework (also referred to as the Schema and Analysis and Evaluation Instrument) by using three different analysts to encode selected archival schemas, some examples of key metadata schemas from related information fields (for example, Dublin Core), and from scientific and artistic domains independently.

4.a.5. Findings about MADRAS Tools and Instruments

Upon proceeding with analysis of selected schemas, we were somewhat surprised by the spotty nature of schema documentation. Since a schema is analysed based upon its documentation, it vital that this information be clear and concise, but often we found it to be insufficient/deficient. Insufficient schema documentation led us to realise that we needed to refine our questions to make sure that they were focused on eliciting responses about what schema is intended to do versus what the 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, we have found that we spend 85% of our time pouring over the definitions of sub-elements. While we do 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 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 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: addressing MADRAS Question 206, "Chronological Date" v. "Creation Date" -- InterPARES lists four date types in the benchmark standard, "Chronological," "Received," "Archival" and "Transmission." ISO documentation seems 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 recordkeeper can directly attest (i.e., we know the date written on a document (Chronological Date) but we cannot actually be sure that this was its *creation* date), while ISO appears to believe that the recordkeeper 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. We assume is 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. 9.3.1 b (which stands behind question 504) states, "capture the business rules or other system controls that regulate record creation and management" while 9.3.1 d (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 9.3.1 b differ from the "records management operations" of 9.3.1 d? We assume that 9.3.1 b is about creation, access and use while 9.3.1 d is about activities performed only by records managers, such as preservation actions. Furthermore, since these instruments also largely directed how we crafted the system, some of the concepts in ISO 23081 posed particular challenges. The standard describes that we 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, we considered metadata about a record's "content, context and structure" to be the metadata created 'at record capture.' Any other metadata that we describe is, thus, by definition 'after record capture.' This amounts to isolating the metadata that deals directly with recordkeeping/administration, which appears to be in the spirit of ISO 23081.

4.a.5.i. Findings About the Schemas

As noted earlier, analysing every schema identified as relevant was beyond the scope of this project. However, we did analyse enough from different sectors and of different types to be able to make the following observations:

- Almost no schema analysed, 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 are proving 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;"
- 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 in order 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); and,
- Even if schema were to meet all the requirements, this is unlikely to be the case in

specific implementations/application profiles. The process of completing our 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, ANZLIC 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 a metadata element 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.

Add in some data on how well selected schemas did, including ISO 19115.

We 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.

Some tools, especially those outside the more traditional recordkeeping/archival domain, do not fall neatly into some of MADRAS' classifications. How can we 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, we would be comparing apples to oranges. Also, 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, granularity of content description tends to be higher in non-recordkeeping schemas, while the recordkeeping schemas focus more on context description (as might be expected).

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 to 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.

4.a.6. 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 standalone 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[18]. 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.

4.b. Warrant Database

4.b.1. Scope and Rationale

Description Group 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 we would develop a way of assessing those schemas already developed by different communities against our requirements, and provide them with feedback about how they could be extended or modified to address recordkeeping issues. We then discussed how we could develop persuasive arguments that might lead those communities to respond to our recommendations. We decided that we 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, we felt that we 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 ours upon them.

The literary warrant database built upon the method developed by Wendy Duff as part of the Pittsburgh Electronic Records Project of identifying a warrant for a particular course of action

[18] For each schema or set registered, a set of evaluative reports can be generated that indicate the degree to which the schema meets InterPARES Benchmark and Baseline Requirements or the ISO 23081 metadata requirements (recognizing that users may be interested in addressing either or both sets of requirements), pinpoints in what ways the schema falls short, and provides guidance as to how it could be modified or augmented to meet the requirements.

based upon such things as legal or other juridical mandate, professional best practices, professional literature, and other social sources[19]. In our case, we 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, we conducted a literature review across each focus area in order 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. We designed and set up web-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 Group researchers then analysed 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[20].

4.b.2. Results of Warrant Analysis

The database now contains 177 records that include not only bibliographic information, but 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.

4.c. News Archives Case Study

Although a series of broader case studies were conducted by InterPARES focus groups that included the gathering of data about metadata on behalf of the DCD (discussed below), the Description Group 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 case study surveying this industry would provide important insight into how one specific community is addressing metadata and preservation issues more broadly.

4.c.1. Conducting the Case Study

In recent years there has been a growing awareness that historic news archives in electronic formats are at risk.[21] In the popular media, printed newspapers are frequently described as a

---

[19] See Duff, Wendy M. *The Influence of Warrant on the Acceptance and Credibility of the Functional Requirements for Recordkeeping.* University of Pittsburgh. Ph.D. Dissertation, 1996; ibid. "Warrant and the Definition of Electronic Records: Questions Arising from the Pittsburgh Project." *Archives and Museum Informatics, 11* nos. 3&4 (1997): 223-231; and ibid. "Compiling Warrant in Support of the Functional Requirements for Recordkeeping." *Bulletin of the American Society for Information Science, 23* no. 5 (1997): 12-13.

[20] Available at http://www.interpares.org/ip2/ip2_warrant_db.cfm

[21] Victoria McCargar, "Following the Trail of the Disappearing Data," *Seybold Report* 4, no. 21 (2005).

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.[22]

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.[23] 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 (see APPENDIX B). After receiving the appropriate permissions for human-subjects testing through UCLA, the survey was uploaded to a professional interface at the Surveymonkey website 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 80 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

---

[22] Bernard F. Reilly, Jr., "Knowledge Biodiversity: The Perilous Economic of World News Heritage Materials," in *ACRL Twelfth National Conference* (Minneapolis, MN: Association of College and Research Libraries, 2005).

[23] Helen R. Tibbo, "Primarily History in America: How U.S. Historians Search for Primary Materials at the Dawn of the Digital Age," *American Archivist* 66, no. 1 (2003): 9-50.

Special Library Association's 2006 annual conference in Baltimore[24].

4.c.2. Discussion of the Case Study Results

Despite the advances in digital preservation research in the last 10 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 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 we 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 77 participants who started the survey, only 28 – 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 2 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 28 responses, only five answered that they had not. Twenty-one of the remaining 23 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.[25] 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 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

[24] Victoria McCargar's Powerpoint slides and audio from the SLA Conference are posted on the News Division's web site at http://www.ibiblio.org/slanews/conferences/sla2006/#wed.
[25] Victoria McCargar, "The Heart of Darkness: A Foray into Aging JPEGs," *Seybold Report* 5, no. 22 (2006).

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 fulfill a legal request, either one arising from the newspaper's own activities or those between third parties.[26]

- 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.[27] 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 web pages in many of the archives as the impact of multimedia publishing matures. However, few controls are in place. 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 catalog hardware and software metadata in their archives, while only 15% record the operating system and 7% record the necessary

[26] Personal conversation between McCargar and Virginia Everett, news director of the *Atlanta Journal-Constitution,* May 9, 2006, in Atlanta.
[27] Photographers' archiving practices are highly idiosyncratic; see Jessica Bushey, *Survey of Record-Keeping Practices of Photographers Using Digital Photography* (Vancouver, British Columbia: Interpares 2, 2005), PDF.

peripherals even though all of these elements are specified in PREMIS.[28] These numbers cannot be extrapolated across all news archives since only 26 respondents of the original 77 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*[29] 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 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.[30]

4.c.3. 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: "We are so busy creating digital archives we are not paying attention to the problems we 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.[31]

---

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

[29] *New York Times Co. v. Tasini et al.* (00-201) 533 U.S. 483 (2001) 206 F.3d 161, affirmed June 25, 2001.

[30] See http://www.loc.gov/section108/. McCargar contributed a public comment on behalf of news archivists at http://www.loc.gov/section108/docs/McCargar.pdf

[31] 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 http://www.crl.edu/content.asp?l1=13&l2=58&l3=162.

4.d. Metadata Specification Model

The premise underlying the work of the DCD is that detailed trustworthy metadata is key to ensuring the creation of reliable and preservation of authentic records and other entities in electronic systems argues for is 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 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 it refers.

One goal of the metadata specification model was to identify an overall set of metadata requirements that specify what metadata needs to be created, from which sources, how, and by whom, at which points within both the Chain of Preservation (life cycle) and the Business Process (records continuum) Models being developed by the IP2 Modeling Cross-Domain Group, 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 and 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.

4.d.i. Actions Taken and Products Created

Description Group 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 Process 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 (see Appendix C). 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 is defined as those categories of metadata carried forward to be used as evidence for archival description. 137 different metadata assertions were identified (i.e., different instances of types of metadata), 16 types of assertions were identified. 2 cut across all stages of the lifecycle, 1 cut across 2 stages, and the other 15 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 of specific implementations, and by actual system building. When researchers start to work on the development of the metadata specification model for the Business Process Model, it is anticipated that we 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 life cycle. 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.

4.e. InterPARES 2 Case Study Data Analysis

4.e.1. Actions Taken

As part of the case studies undertaken by the InterPARES 2 focus groups, several questions were included in the case study protocols that potentially addressed issues of concern to the Description Cross Domain Group:

- 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 systems 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 schema 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 this metadata (institutional convention, professional body, international standard, individual practice, etc.?)

Metadata issues also could surface in a more general manner in the course of the case study. APPENDIX D contains the summaries of the data collected in the Focus 1 and 2 case studies as these relate to metadata and description concerns, The summaries were generated by researchers at UBC.

4.e.2. Data Analysis

The data analysis sought to identify through the answers to the above 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 targeted recordkeeping issues; and to identify the extent to which real-world implementation of metadata measured up to the ideal of the metadata requirements delineated in the Analytical Framework. We anticipated that we would see more, and more rigorous metadata implementation in the government and scientific rather than the artistic area.

In the artistic focus, only two case studies uncovered use of metadata standards, and none of these were standards developed specifically for recordkeeping, archival or preservation functions: CS09 (Animation) cites use of several common bibliographic description and resource discovery metadata schemas--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). CS09 (WGBH) also cites use of Dublin Core and LCSH, as well as an industry schema, the Public Broadcasting Core (PBCore). CS03 Horizon Zero 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 standard and the IMS Learning Resource Metadata specification. In terms of overall metadata implementation, none of the artistic focus case studies indicated conscious attempts to apply metadata, beyond a few efforts to establish filenaming conventions, largely for retrieval purposes, some version control, and in some cases, rudimentary tracking of file check-in or out or file archiving. Roeder report, pp. 4, 7, 8, 10, 11, 12, 13, 15.

In the scientific focus, CS 14, Archaeological Records in a Geographic Information Systems 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 Metadata Standard. 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 with the exception of CS06 the Cybercartographic Atlas of Antarctica. CS 06 does outline important metadata elements that should be present and where these should be located. … contains a much more detailed account of metadata considerations … ,. Indeed, each of the four scientific case studies, which are all primarily geospatial in nature, indicate that a rich level of metadata is created or could be created (expensive) and that there is clearly an overall concern with metadata quality control. However, scientific metadata standards need to explicitly address archival and preservation as well as data quality and lineage requirements.

In the government focus, CS19 uses OWL to represent semantics within metadata schemas.

4.e.3. Conclusions Drawn from the Case Studies

Overall, the findings of the case studies are disappointing, but not surprising, and it would be interesting 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. While the scientific case studies indicate an awareness of the need for metadata and the role it can play in ensuring the accuracy and long-term usability of digital materials that is absent from the artistic case studies, no set of case studies indicated any real consciousness of the overall role of *recordkeeping* metadata in their activities. In the artistic 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. An additional concern that was raised in the Scientific Focus is that despite element-rich complex metadata schemas being developed in areas such as the geospatial domain, there is little incentive or resources actually to create that metadata. If we wish to influence these communities and persuade them to add even more elements to their schemas, then we 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.

5. OVERALL RESULTS

The work of the InterPARES 2 Description Cross Domain Group 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.

How policy is covered – rich and oppositional schemas for IP, as well as policy issues raised – proprietariness of schemas work against schema registration and analysis as well as preservation. Should metadata creation, management and preservation be required for grant-funded projects?

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, 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 catalogers 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 relates. 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 works and needs to work. The development of MADRAS established an ideal against which existing or draft metadata schemas and sets can be assessed[32]. 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 studies and the news archive case study, where there was little evidence, especially in the artistic focus, of any attempt to implement recordkeeping metadata at all. Although the trend in information management is toward the creation of leaner metadata, we believe that it is important to contemplate how to change the dynamics of metadata depreciation and minimalisation 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 plays 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 DCD's work arises not only with the differing scopes and viewpoints of the metadata schemas that have been registered and analysed 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

[32] Although it should be noted that as it is, it is difficult to perform a sophisticated interpretation of the analysis when we 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.

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 operationalised as tools. The Description Group 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 DCD 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 Process Model, with the dimensionality afforded by its four axes of identity, evidentiality, transactionality and recordkeeping entity,[33] necessitated that the DCD take a more complex view of metadata and archival description than might have been needed if it had looked only at supporting a Life Cycle Model.

The activity models developed in InterPARES 1 were based on a life cycle 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 DCD 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 life cycle 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 recordkeepers 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 life cycle 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 to be addressed.[34]

6. AREAS FOR FUTURE RESEARCH AND DEVELOPMENT

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

Can metadata associated with the creation and active use of records ever contribute to

[33] Upward, Frank, "Structuring the Records Continuum Part One: Post-custodial Principles and Properties," *Archives and Manuscripts* 24 no. 2 (Nov 1996): 268-85; Upward, Frank, "Structuring the Records Continuum Part Two: Structuration Theory and Recordkeeping,"*Archives and Manuscripts* 25 no. 1 (May 1997): 10-35.

[34] 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 life cycle model, but upon further scrutiny could equally well support a continuum approach.

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,[35] that certain types of metadata, created while the records to which they relate are active, could be captured or analysed 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. It 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).

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 is 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

[35] See San Diego Supercomputer Center, Archivists' Workbench Project Summary. Available: http://www.sdsc.edu/NARA/Publications/nhprc_summary.pdf

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 labor-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 unimplementable for archivists.

In terms of development work, we 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. We recognise 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, we 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 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, our finding is 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 we 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 our 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) in order 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 us 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, we 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 B. News Archive Survey Instrument

News Archives Survey questions (survey was conducted August-October, 2005, on the Survey Monkey website. Survey may be viewed at http://www.surveymonkey.com/MySurveys.asp?Rnd=0.4796411

Contact questions

Contact Name: _____

Name of Institution: _____
_____

Type of institution: _____
_____
_____

Address: _____
_____
_____
_____

Phone (landline): ( )_____
Phone (cell): ( )_____
Fax: ( )_____
e-mail: _____

Would you be willing to participate in a follow-up interview?
□ Yes □ No

Institutional environment

Type of news archives
□ Newspaper
□ Broadcast
□ Other (describe): _____

If newspaper, what is the circulation?
□ < 20,000
□ 20,000-99,999
□ 100,000-249,999
□ 250,000-499,999
□ 500,000-749,999
□ > 750,000

If a broadcast property, what is the size of your audience? ____
_____
_____

What is the size of your editorial staff?
□ < 20
□ 20-100
□ 101-250
□ 251-500
□ 500-1,000
□ > 1,000

Is the archives part of a larger department within your institution (e.g., the editorial library)?
□ Yes □ No

To what extent is your senior newsroom management committed to archival policy?
□ Very committed
□ Somewhat committed
□ Not committed at all
□ Don't know

Professionalism

Are there specific job competencies required for the archivist, such as professional education levels for archives management or staff (e.g., MLS or MLIS, archives certification)?
□ Yes □ No

If yes, what are they? _____
_____
_____

Are continuing education opportunities made available to archives staff?
□ Yes □ No

How familiar is archives staff with software used by the newsroom to create objects that go into the archives? (E.g., software such as Photoshop, Adobe Illustrator, DreamWeaver, pagination system.)
□ Very familiar
□ Somewhat familiar
□ Not familiar at all
□ Don't know

Budget

How are the archives funded? _____
_____
_____
_____

Have you noticed any trends in the budget as it relates to the archives in the past five years? _____

_____

_____

_____

What departments or other factors compete with the archives in the budget (e.g., serials budget)? _____

_____

_____

_____

Is any of the archives' budget earmarked for preservation?

□ yes, part of the archives' budget is reserved for preservation activities
□ no, but there is a separate budget for preservation activities
□ no, there is no budget for preservation activities

Use of archives

In general, how are the archives used?

□ News research
□ Corporate or business-unit research
□ Sales to vendors
□ Sales to the public
□ Outside researchers
□ Other (describe) _____

_____

_____

_____

Is there any recognizable pattern to how items in the archives are reused?

_____

_____

_____

How important is it to maintain in your archives access to files in their original, native software format, as opposed to a more "generic" format like PDF or EPS (e.g., native Photoshop or Adobe Illustrator)?

□ very important
□ somewhat important
□ not important
□ Don't know

What year did you start digitally archiving objects of the following types?

Text _____
Photography _____
Information graphics _____

Video _____
Web pages _____
Other type (describe) _____
_____
_____
_____

Policy

Do you have a written preservation or other policy that determines what digital materials you take into your archives and for how long?
□ Yes (please answer questions 1a-1d, below)
□ No (please answer questions 1e-1f, below)

If yes, which of the following are specified? Check all that apply.
□ what is accessioned (taken in)
□ willingness to maintain materials in problematic, hard-to -handle formats
□ the length of time access will be guaranteed for nonstandard formats

If yes, how often is the policy reviewed to provide for new technologies?
_____
_____

If yes, if your institution has a preservation policy, does it address records in electronic form?
□ Yes □ No

If no, are unwritten preservation procedures and practices familiar to all staff members?
□ Yes □ No

If no, are you planning to introduce a written policy in the future?
□ Yes □ No

If no, how do you handle disputes arising from archival practices?
_____
_____
_____
_____

Technology

How are digital objects and metadata structured in the archives?
□ single objects with integrated metadata
□ single objects with metadata stored separately
□ compound objects with integrated metadata
□ compound objects with metadata stored separately

Who is responsible for hardware support?

□ archives/library staff
□ information technology (IT) department
□ other: _____

Who is responsible for software support?
□ archives/library staff
□ information technology (IT) department
□ other: _____

Is library software upgraded and synchronized with newsroom software upgrades?
□ Yes □ No

Are software licenses documented and maintained?
□ Yes □ No

5a. If yes, how? _____
_____
_____

Are back versions of software and older models of hardware kept and maintained?
□ both
□ just software
□ just hardware
□ neither

How and to what extent are you monitoring technological changes over time (e.g., software compatibility, sequential upgrades)?
_____
_____
_____

How and where is the data backed up? _____
_____
_____
_____

What features does your archive system include? Check all that apply.
□ file server with proprietary database off the shelf
□ file server with proprietary database that has been customized
□ homegrown database
□ loose CD-ROMS or DVDs
□ CD-ROMs in a jukebox
□ other (describe): _____
_____
_____

Is hardware subject to a regular upgrade schedule?
□ Yes □ No

Are upgrades to hardware and software formally documented?
□ Yes □ No

What kind of files do you maintain? Check all that apply.
□ ASCII text
□ text in proprietary format like Word
□ images in JPEG, TIFF, etc.
□ images in formats other than JPEG
□ graphics files in Illustrator, Freehand, Canvas
□ web video formats (e.g., Digital Beta, Final Cut)
□ spreadsheets
□ GIS databases
□ databases created in FileMaker Pro or MS Access
□ other databases (describe): _____
_____
_____
□ other (describe): _____
_____
_____

Are certain types of software that you use machine-dependent (e.g., software available only for Macintosh)? ____
_____
_____

Are certain types of software that you use dependent on other types of software (e.g., Quark or Photoshop extensions)?
_____
_____
_____

How are any existing software dependencies reflected in your archives' metadata?
_____
_____
_____
_____

Are you archiving web pages?
□ Yes □ No

If yes, how? _____
_____
_____

Are you archiving information graphics
□ Yes □ No

If yes, how? _____
_____
_____

Are you archiving news video?
□ Yes □ No

If yes, how? _____
_____
_____

Metadata

Does your metadata incorporate the IPTC standard (sometimes known as the "IPTC header")?
Yes No

In general, how much customization/variation have you introduced to the IPTC standard? (For example, have you added fields, eliminated fields, reused fields for purposes other than what is labeled)? _____
_____
_____

Do you have a corporate metadata standard used by more than one business unit?
_____
_____
_____

If yes, has this required you to change how you would ordinarily work?
_____
_____
_____

Do you use any metadata standards (e.g., IPTC, MIX, JPEG2000, etc.)?
□ Yes □ No

If yes, which? _____
_____

Which of the following does your archives' metadata scheme include? Check all that apply.
software name
version information

hardware requirements
operating system requirements (e.g., PC, Mac, Unix, Linux)
peripherals required for rendering
persistent identifier (a unique identifier that doesn't change)
rights information
encryption keys
other (describe): _____

What categories of metadata are used by your archives? Check all that apply.
rights and permissions
bibliographic (e.g., publication date/part/page)
reuse of object (e.g., publication history)
technical (e.g., information about system, software, hardware used in creating document)
relationships to other objects (e.g., picture or graphic to story, story to sidebar)
other (describe): _____

How do you indicate relationships between objects? _____
_____
_____
_____

Are your records indexed by subject?
☐ Yes ☐ No

Is a controlled vocabulary (sometimes called "taxonomy") used in your subject indexing?
☐ Yes ☐ No

If yes, from what source did you get the controlled vocabulary?
_____
_____

How are indexers trained?_____
_____
_____

How are your vocabularies managed? _____
_____
_____

How is metadata gathered? Check all that apply.
automatically exported from production systems without human intervention
manually entered
automatically exported from production system with manual enhancement
other (describe): _____
_____
_____

Digital Preservation

What kind of files do you maintain? Check all that apply.
ASCII text
text in proprietary format like Word
images in JPEG, TIFF, etc.
images in formats other than JPEG
graphics files in Illustrator, Freehand, Canvas
spreadsheets
GIS databases
MS Access, FileMaker Pro or other databases
other (list): _____
_____
_____
_____

How do you *currently* ensure continued access to content stored digitally? Check all that apply.
batch migration of aging or obsolete files to newer software version
migration of individual files to new format as needed
restricting ingestion to limited list of supportable formats
transferring data to new storage medium of the same type
transferring data to new storage medium of a different type
keeping up with all recommended software upgrades
keeping old hardware/software available
keeping unreadable files in anticipation of a later translation solution
recording technical metadata about hardware, operating systems and formats to enable future preservation activities
engaging in regular metadata evaluation and development
normalizing text files to ASCII or Unicode
storing images as TIFFs
converting one or more different formats to PDF
printing to paper
microform
other (describe): _____
_____
_____
_____

How did you arrive at your current digital preservation strategies?
_____
_____
_____

How are you documenting your current digital preservation activities?
_____

_____
_____

What are your plans for maintaining access to content stored digitally *in the future (e.g., beyond a 10-year horizon)*? Check all that apply.
batch migration of aging or obsolete files to newer software version (forward migration)
migration of individual files to new format as needed
specially written software to emulate older operating systems
transferring data to new storage medium of the same type
transferring data to new storage medium of a different type
keeping up with all recommended software upgrades
keeping old hardware/software available (digital archaeology)
engaging in regular metadata evaluation and development
recording technical metadata about hardware, operating systems and formats to enable future preservation activities
restricting ingestion to limited list of supportable formats
normalizing textual files to ASCII or Unicode
storing images as TIFFs
converting to PDF
printing to paper
microfilming
other (describe): _____
_____
_____
_____

How did you arrive at your future digital preservation strategies?
_____
_____
_____

In any of your previous preservation activities (e.g., upgrading software, moving to new storage media, moving to a new software product, etc., did you experience any loss of data or metadata, or otherwise compromise the archives)? Please describe.
_____
_____
_____
_____
_____

What issues do you have or anticipate having in migrating your digital materials?
_____
_____
_____
_____
_____

Are you actively working with a vendor on digital preservation (for reasons other than disaster recovery)? Please identify which vendors you use.

_____
_____
_____

What is the current level of archives staff familiarity with current digital preservation issues, research and solutions? ____

_____
_____
_____

What do you see as the most pressing issue in digital preservation in general?

_____
_____
_____
_____
_____

Copyright

Have you encountered copyright problems with your legacy files in the wake of the Tasini case affecting repurposing of freelance material in digital form?
□ Yes □ No

If yes, what are the issues? _____
_____
_____
_____

If yes, approximately how much of your collection is affected?
□ less than 10%
□ as much as 25%
□ between 25% and 50%
□ more than 50%
Comment: _____
_____
_____
_____

How long have you had a written contract in effect covering reuse, repurposing and/or archiving of freelancers' material?
□ prior to 1985
□ since about 1990
□ since about 1995
□ between 1995-2000

□ since 2001
□ no contract exists
Comment: _____
_____
_____

Do you archive all published material digitally as a record of publication, regardless of its copyright status?
□ Yes □ No

If no, what do you do with material for which you do not hold copyright?
_____
_____
_____
_____

Which department handles copyright issues in your organization?
□ legal department
□ finance department
□ newsroom
□ publisher's office
□ other (describe): _____
_____

Do you have a dedicated rights management software module as part of your archiving system?
□ Yes □ No

What rights metadata do you keep with each file? Check all that apply.
□ copyright holder of object
□ source of object, if not copyright holder
□ contract status of creator, if freelance
□ restrictions or limitations on reuse
□ restrictions or limitations on archiving
□ dates associated with the contract (e.g., expiration, reversion)
□ other (describe): _____
_____

Are rights metadata terms assigned according to a controlled vocabulary (authority file, standard set of terms) or free text?
□ controlled vocabulary
□ free text
□ both

Comments

Please use this space for anything else you would like to add to your response.

APPENDIX D. Focus 1 and 2 Case Study Data Relating to Metadata

Focus 1. Artistic Activities

General information regarding metadata

CS01 *Arbo*	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. (32) 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. (51)
CS02 *Stelarc*	The report states that there are no formal recordkeeping practices and thus, no metadata are recorded. The materials are arranged according to his performance and publicity needs, see question 4(d).
CS03 *Horizon Zero*	The report states that the organization of the files pertaining to each issue of Horizon Zero is ad hoc, and is generally organized by the issue for which they were created. (5) These files are accessible through a shared space which can be navigated using tracking software that organizes the posting into threads. (7) These tracker entries are saved using an archival function implemented in the tracker software (Mantis 0.18.0A4). (7)
CS09 *Animation*	The final 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. (14)
CS09 *Altair 4*	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. In order to use them, it is necessary to know the filename, path, and approximate date of production. (6)
CS09 *WGBH*	N/A.
CS10 *Danube Exodus*	Filenaming 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. (5)

Metadata information in the 23 questions:

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

CS01 *Arbo*	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. (34)
CS02 *Stelarc*	The digital entities are identified under project titles, event series, and biographical content on the web site. (8)
CS03 *Horizon Zero*	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. (6)
CS09 *Animation*	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 we 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 *Altair 4*	These conventions comprise the folder with project name/file object name/number of version and the last version file object name/final version. (4)
CS09 *WGBH*	Current: Yes, and the unique identifier links the catalog 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. (6)
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. (4)
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. (4)

18b. 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.)?

CS01 *Arbo*	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. (48)
CS02 *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. (12)
CS03 *Horizon Zero*	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. (9).
CS09 *Animation*	Another database, built on FileMaker Pro and called ArchiveWorks, is used for tracking physical pieces of artwork that are not digital.
CS09 *Altair 4*	No information provided.
CS09 *WGBH*	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 subject has 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. (9)
CS13 *Obsessed Again*	None. (9)
CS15 *Waking Dream*	Not applicable. (7)

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

CS01 *Arbo*	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. (48)
CS02 *Stelarc*	Yes. Links are also present to make collaborators' websites and other relevant internet locations accessible. If general links

	become obsolete the webmaster will keep them on the website as dead links. If important links become obsolete new links will be set up to make that information accessible. (12)
CS03 *Horizon Zero*	Yes. (9)
CS09 *Animation*	Yes, access is maintained for all relevant digital entities and their metadata. Everything in the system that can be opened can be downloaded.
CS09 *Altair 4*	No information provided.
CS09 *WGBH*	Current: No, the analog/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. (9)
CS15 *Waking Dream*	Not applicable. (7)

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*	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. (48)
CS02 *Stelarc*	No, the webmaster does not keep a record of specific updates to the website. The report states that the metadata are unknown. (12)
CS03 *Horizon Zero*	The report states there are no recordkeeping system. (9)
CS09 *Animation*	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 *Altair 4*	No information provided.
CS09 *WGBH*	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. (9)
CS15 *Waking Dream*	No metadata is consciously captured. (7)

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*	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. (51)
CS02 *Stelarc*	This is unknown.
CS03 *Horizon Zero*	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. (10)
CS09 *Animation*	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 *Altair 4*	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. (7)

CS09 *WGBH*	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. (11)
CS13 *Obsessed Again*	There are no descriptive or other metadata schemas or standards currently being used. (10)
CS15 *Waking Dream*	No descriptive or metadata standards are currently being used. There is no recordkeeping system being used. (8)

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*	Arbo does not use any descriptive or metadata standards. The report states that the "grey-zones" list information, thus, are not standardized. (51)
CS02 *Stelarc*	The report states that it is likely individual practice by Stelarc and his webmaster that are the sources for any descriptive standards. (14)
CS03 *Horizon Zero*	The CanCore standard is derived form 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. (11)
CS09 *Animation*	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 *Altair 4*	The final report states that only material that is archived are then governed by international standards. (14)

CS09 *WGBH*	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. (11)
CS13 *Obsessed Again*	No such schema or standards are employed. (11)
CS15 *Waking Dream*	Not applicable. (8)

Focus 2. Scientific Activities

General information regarding metadata

CS06 *CyberCartographic Atlas of Antarctica*	The final report has yet to be submitted, thus the answers to the 23 questions have been taken from the research proposal and other interim reports. The research proposal states that CyberCartographic Atlas of Canada may be the first project of its kind to incorporate at the very beginning of the project the process of preserving the entire life cycle of its data. The project is expected to develop processes, methods, tools and guidelines for archiving and explore the possibility of expanding metadata tools for this purpose.
CS08 *NASA*	No report available.
CS14 *Archaeological Records in a Geographical Information System*	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. (17)
CS19 *Electronic Engineering and Manufacturing*	No report available.
CS26 *MOST Satellite)*	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?

CS14 *Archaeological Records in a Geographical Information System*	The 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. (16)
CS26 *MOST Satellite)*	The key elements are mainly textual, but there are graphic elements as well. (14)

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

CS14 *Archaeological Records in a Geographical Information System*	Digital entities are identified through file naming conventions. Aggregations of files within certain folders can also create an associative identity of their own. (17)
CS26 *MOST Satellite)*	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. (15)

18b. 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.)?

CS14 *Archaeological Records in a Geographical Information System*	The 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. (23)
CS26 *MOST Satellite)*	The report states that there is no formal capture system in place, beyond the tools within Microsoft Windows. (18)

18d. Does the record keeping systems provide ready access to all relevant digital entities and related metadata?

CS14 *Archaeological Records in a Geographical Information System*	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. (24)
CS26 *MOST Satellite)*	The report reveals that it is possible to access all digital entities via Windows Explorer. However, does not mention how it provides access to the metadata prescribed by the MOST researchers. (18)

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*	Interim reports state that the following are integral components of spatial metadata: lineage, positional accuracy, attribute and thematic accuracy, completeness, logical consistency, semantic accuracy, and temporal information.
CS14 *Archaeological Records in a Geographical Information System*	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. (24)
CS26 *MOST Satellite)*	The report states that there is no audit trail. (18)

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*	A preliminary report states that standard metadata catalogues are captured as data lineage. In addition, each scientific domain is governed by their particular data quality standards, measures and assurances and these are included in the metadata.
CS14 *Archaeological Records in a Geographical Information System*	The 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, website, database, etc.) the data was retrieved. (25)

	In addition, time tagging of georeferenced information is part of the documentation of the processes of creating online digital maps, models and georeferenced visualizations.
CS26 *MOST Satellite)*	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 refers 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. (19)

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*	Metadata standards for geospatial information are well developed; however, the research proposal identifies that there are insufficient multimedia metadata standards that exist.
CS14 *Archaeological Records in a Geographical Information System*	Within ArcCatalogue, the user could create, manage and edit metadata based on the Federal Geographic Data Committie (FGDC) Content Standards for Digital Geospatial Metadata or the ISO 19115 Metadata Standard. This metadata would be stored in XML. (25)
CS26 *MOST Satellite)*	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. (19)

[i] 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*, InterPARES, September 2002, http://www.interpares.org/book/index.htm.