Archival Science (2005) 5: 43–78
DOI 10.1007/s10502-005-9000-4

Towards a 21st Century Metadata Infrastructure Supporting the Creation, Preservation and Use of Trustworthy Records: Developing the InterPARES 2 Metadata Schema Registry★

ANNE GILLILAND[1], NADAV ROUCHE[1], LORI LINDBERG[1] and JOANNE EVANS[2]
[1]*Center for Information as Evidence & Department of Information Studies, University of California, Los Angeles, U.S.A (E-mail: swetland@ucla.edu);* [2]*School of Information Management and Systems, Monash University, Melbourne, Australia*

Abstract. This paper argues that an essential component of electronic recordkeeping needs to be an infrastructure to support the creation, preservation and accessibility over time of trustworthy, understandable metadata. This infrastructure can then also be used to provide specifications and an implementation environment for automated tools to assist archivists in the ongoing management of trustworthy records *and* metadata, and users in the identification, retrieval, and manipulation of those records and metadata. The paper discusses this need in the context of the development by the International research on Permanent Authentic Records in Electronic Systems (InterPARES 2) Description Cross-Domain Group of a metadata schema registry. This registry is a prototype resource designed to assist archivists and records creators in multiple domains in developing and assessing their own and other communities' metadata infrastructures. The paper concludes by identifying two contested issues that are surfaced and how they are being confronted by this work: one of these is a definitional issue that relates to how to delineate the concept of archival description in the face of competing notions of "metadata." The other is the extent to which both the life cycle and continuum worldviews and associated activities can or should be supported, reconciled or even re-thought through the conceptual and analytical approach that is embedded in the metadata schema registry.

★ An abbreviated version of this paper was presented at Smart Metadata and the Archives of the Future: Archives, Memory and Knowledge: International Congress on Archives 2004, Vienna, Austria August 24, 2004. The authors wish to acknowledge the funding support for the InterPARES 2 Project of the United States National Historical Publications and Records Commission, the National Science Foundation, and the Social Sciences and Humanities Research Council of Canada. The authors also wish to acknowledge Professor Sue McKemmish of Monash University and Dr. Richard Marciano who have been instrumental in the development of the InterPARES 2 Metadata Schema Registry.

Introduction

For archivists engaged in preserving and providing access to ever-increasing volumes of electronic records and other born-digital materials, tied to the obvious question of how to preserve these materials and make them available to researchers are several other issues relating to documenting and maintaining the trustworthiness of those materials in and across time and space. What means have creators employed to ensure that their records are reliable? How do archivists ensure that the methods they employ in the acquisition and preservation of those records maintain the level of reliability and the authenticity status of those records that exist when they are accessioned? What kind of conceptual and operational frameworks do models such as the records life cycle and the records continuum provide for addressing issues of trust as they relate to records and to archives and their activities? How are future users to be assured that the preserved digital records they are accessing are trustworthy? How do these users come to understand the metanarratives or realities that the records might reflect or obfuscate without being concerned that any gaps, inaccuracies, low resolution, and other issues that might impede their readings of the record could be artifacts of the archival management of the records rather than the state of the records at the point when they came under archival control?

None of the above questions, of course, are new, no matter what the medium of the records. However, after many years of research and practice with electronic records, many archivists now acknowledge that the development of a rigorous, unambiguously delineated *metadata*[1] infrastructure can serve as an imperative to address such questions, so often swept under the carpet in archival practice, at the same time as providing a powerful means for answering them.[2] What is less frequently acknowledged is that the latter can only be the case if the metadata themselves are trustworthy

[1] The term "metadata," as used in this paper, and as discussed in the paper's concluding section, refers to all types of structured information, including archival description, that is created manually or automatically by recordkeeping systems including metadata that documents the juridical-administrative, business and technical contexts within which records are created; identifies records and delineates how the records behave, their function and use; identifies and describes the relationships within and between records and other information objects; and expresses and supports how records should be managed, and what happens to them over time.

[2] See Gilliland-Swetland, Anne. "Electronic Records Management," *Annual Review of Information Science and Technology*, 2005 (forthcoming).

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, including those generated and managed by records creators (i.e., not just description created by archivists functioning within either a custodial or post-custodial paradigm), must be sufficient, appropriate, understandable, and of high quality. Both the provenance and the version of metadata must be clearly identifiable. Moreover, metadata, which often capture much of the context of records, potentially also offer researchers a rich source of intelligence about records creators and record creation processes. Ideally, metadata should also be machine-processable, since this attribute will help to underpin the future generation and implementation of automated tools for creating and harvesting metadata and for managing, retrieving and manipulating them together with the archival materials to which they relate.

This paper argues, therefore, that archivists must focus much more attention on what metadata are created, preserved, and used, when, by whom, and how. A by-product of doing so, will be the creation of a technical and conceptual infrastructure that will allow entirely new views and technological capabilities to be supported for records and their users. How, however, are the repositories and communities of practice that are involved in the creation, maintenance and dissemination of records to go about developing a metadata infrastructure today that is capable of facilitating the technologies, tools, and user needs of tomorrow? After a brief review of the activities of the InterPARES 2 Description Cross-Domain Group, the paper focuses on the Group's development of a prototype resource – a metadata schema registry and the analytical framework around which it is devised – designed to assist archivists and records creators in any domain in developing and assessing their own and other communities' metadata infrastructures as these relate to concerns of reliability, authenticity, and preservation. The paper concludes by identifying two contested issues that are surfaced and how they are being confronted by this work. One of these is a definitional issue that relates to delineation of the concept of archival description in the face of competing notions of "metadata." The other is the extent to which both the life cycle and continuum worldviews and associated activities can or should be supported, reconciled, or even re-thought through the conceptual and analytical approach embedded in the metadata schema registry.

The need for a metadata infrastructure to support 21st century archival management and services

Recordkeeping metadata have been defined as including "all standardized information that identifies, authenticates, describes, manages and makes accessible, through time and space, documents created in the context of social and business activity."[3] What is distinctive about recordkeeping metadata is the range of ways in which they can capture and explicate salient contexts of records as they move through time, space, systems, and types of use and user. It has become a truism to say that a distinguishing characteristic of archivists among information professionals is their emphasis on context, but in fact, much of what archivists and their users find to be salient about records lies within their contexts, and those contexts are most frequently identified and captured not in the content of the records, but in their metadata.

Context, however, remains a somewhat problematic and intangible concept for those who wish to operationalize it into metadata elements and systems design. One way of addressing this problem is to decompose archives and recordkeeping notions of "context" into types that can then be associated with specific processes and attributes. For example, the InterPARES 1 Project thus identified five different types of contexts as being relevant to the maintenance of authentic records over time: juridical-administrative, provenancial, procedural, documentary, and technological.[4] Metadata documenting these contexts are created throughout the life of the record and are integral to the processes of recordkeeping systems design, creating, registering, scheduling, appraising, preserving, describing, researching, and re-versioning records. It might even be appropriate (and certainly may be necessary for the purposes of operationalizing them within systems design) to decompose these types of context further. For example, perhaps the juridical-administrative type could 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

[3] *Australian Standard AS 4390, 1 1996* (New South Wales: Standards Australia, 1996).

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

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 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.[5]

Recordkeeping metadata are created in a variety of ways and by a variety of agents – they may be created manually (as is the case with most archival description) or automatically (as, for example, would be the case with an inverted index of terms culled from a text document). They may also be automatically inferred, derived or harvested from the records and recordkeeping systems themselves, an approach that looks increasingly attractive as systems developers and information professionals of all types become more aware of the burgeoning overhead of metadata creation and management necessary to support the online provision of trustworthy information. They may even be exploited and re-used for purposes for which they were never intended, such as for corporate knowledge mining, developing new institutional market segments, or developing learning objects. In the archival community, research and development activities such as the Archivists' Workbench and PERM Projects of the San Diego Supercomputer Center have begun to explore the development of automated tools for metadata creation and management, as well as for the manipulation of records by end users, and the Clever Recordkeeping Metadata Project described in another article in this issue, which is looking at innovative ways of multi-purposing harvested recordkeeping metadata.[6] All of these developments provide indicators that a rigorous metadata infrastructure can facilitate archivists' traditional activities as well as supporting 21st century archival services such as online retrieval, delivery, certification, compilation and redaction of archival electronic records, and enterprising archival roles and

[5] For example, since almost all extant preservation methods and technologies for electronic records involve some form of reproduction, permission for which has been considerably circumscribed by recent legislation and licensing contracts in many jurisdictions, rights and licensing concerns must now be regarded not only as "use" issues, but also as "preservation" issues.

[6] See McKemmish, Sue and Joanne Evans, "Create Once, Use Many Times: The Clever Use of Recordkeeping Metadata for Multiple Archival Purposes," [need full cite for this issue]; Preserving Electronic Records within an RMA (the PERM Project), available: http://www.sdsc.edu/PERM/; and Create Once, Use Many Times – The Clever Use of Metadata in eGovernment and eBusiness Processes In Networked Environments, available: http://www.sims.monash.edu.au/research/rcrg/research/crm/.

activities that heretofore have been undertaken only in professional domains such as knowledge and digital asset management.

Identifying metadata infrastructure requirements – the work of the InterPARES Description Group

International research on Permanent Authentic Records in Electronic Systems (InterPARES), an international research collaboration involving government archives agencies, academics from Archival Science and several other disciplines, and industry from North America, Australia, Europe and Asia, began examining how to preserve authentic electronic records in 1997. Among its rationales were that (1) the authenticity of electronic records is threatened whenever they are transmitted across *space* (i.e., when sent between persons, systems or applications) or *time* (i.e., either when they are stored offline, or when the hardware or software used to process, communicate, or maintain them is upgraded or replaced); (2) requirements for assessing the authenticity of electronic records that are preserved over the long term are necessary to support the presumption that an electronic record is and continues to be, what it purports to be and has not been modified or corrupted in essential respects; and (3) preservation processes, mechanisms and metadata need to be identified that ensure these requirements are, and continue to be met. The first phase of InterPARES (InterPARES 1) focused mainly on large databases created by government agencies, which account for the largest proportion of electronic records generated by that sector. The second phase of this research, InterPARES 2, commenced in 2001. It continues the work of InterPARES 1, but also incorporates the disciplinary perspectives and activities of communities in the arts and sciences and broadens the focus of the research to include the creation of reliable, as well as the preservation of authentic records. InterPARES 2 also looks in more depth at the implications for records of emergent interactive, experiential and dynamic technologies. The research is divided amongst several domain and focus groups that are examining, primarily through case studies, the nature of the record and conceptualizations of reliability and authenticity in the different sectors or disciplines being investigated, as well as the implications of these findings for archival practice. Four cross-cutting groups, Modeling, Terminology, Policy, and Description, address issues and processes that surface across domains and foci.

As stated earlier, this paper focuses on the activities of the Description Group, and particularly on its development of a metadata schema registry and accompanying analytical framework. Among the main activities of the Description Group are the following:

- Collect data on metadata being identified through case studies being conducted in other InterPARES' groups.
- Develop 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 archival description and other metadata supporting the accuracy, reliability, authenticity and preservation of records.
- Develop and compile a metadata schema registry that describes and analyzes salient features of relevant extant descriptive and other metadata schema and standards.
- Develop and test metadata specifications relating to the activity, entity and data models developed by the Modeling Group which identify the type, source and application of metadata identified in the models, and the existence of relevant metadata schemas.
- Develop specifications for metadata management tools for activities such as automatic metadata creation and extraction.
- Interface with other relevant R&D activities such as ISO 23081 development, the Clever Metadata Project 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.

The notion of a metadata schema registry

In the past few years, several communities, reacting to the proliferation of competing or overlapping metadata schemas, as well as multiple, often incompatible local implementations of those schemas, have begun to develop metadata registries that register schemas and their semantics at the element level and that may also identify local variations (such as application profiles) as well as available, recommended, or supported crosswalks for moving or mapping between schemas.[7] Another area of registry development is the building of the Global Digital Format Registry, which seeks to maintain "persistent, unambiguous bindings between public identifiers for digital representation formats and the syntactic and semantic properties of those formats,"

[7] See for example, the DESIRE (available: http://www.desire.org/), SCHEMAS (available: http://www.schemas-forum.org/registry/), CORES (available: http://cores.dsd.sztaki.hu/), and MEG (available: http://www.ukoln.ac.uk/metadata/education/regproj/) registry projects.

information that "for purposes of long-term preservation of digital objects ... must be sustainable over archival time-spans."[8] The common understanding of a metadata registry is that it is a resource that provides an authoritative and trusted source of information and a resolution service for metadata vocabularies. An existing ISO standard (ISO/IEC 11179) delineates requirements for developing such registries.[9] In most cases, for example, with the Dublin Core Metadata Registry, the registry will contain core elements and their definitions from one or more metadata schemes and versions thereof, accessible through a public interface, together with a capability for users to register local extensions to those schemes. Heery and Wagner, in discussing the development of a metadata registry for the Semantic Web, an endeavor that is particularly concerned with the establishment of trusted sources of metadata, state that "metadata registries essentially provide an index of terms," and identify the goals of such registries as including the provision of a place where "data can be shared and processed by automated tools as well as people." In terms of the W3C Resource Description Framework (RDF) within which the Semantic Web is being developed, the vision is to provide a common approach to declaring schemas in use and supporting environments where, increasingly, users will be automated "agents" navigating, collating and analyzing online information based upon an analysis of the metadata schemas in which that information is encoded.[10]

The over-riding common concern with such developments is the establishment of an authoritative and trusted repository of information about whatever is being registered. Another rising concern has to do with how persistence of identifiers, links and referential integrity over time and version and semantic change can be assured. It is interesting to note that such concerns, arising in the metadata community and addressed toward metadata, have similar resonances to efforts to establish and demonstrate trust in archival repositories, as well as to ensure the integrity of electronic records as they move across time.

The scope of the InterPARES Metadata Schema Registry diverges in some significant respects from other metadata registry initiatives,

[8] Abrams, Stephen L. and David Seaman. "Towards a Global Digital Format Registry," paper presented at the World Library and Information Congress: 69th IFLA General Conference and Council, 1–9 August 2003, Berlin, p. 1–2.

[9] Some aspects of this standard continue to be delineated as experience with developing metadata registries grows, however. See ISO/IEC JTC1 SC32 WG2 Development/Maintenance http://metadata-stds.org/11179/

[10] Heery, Rachel and Harry Wagner. "A Metadata Registry for the Semantic Web." *D-Lib Magazine* 8 no. 2 (May 2002).

however. First of all, it is a resource for the unambiguous registration (and, by implication, identification) of diverse schemas of which some aspects have relevance for the issues with which InterPARES is concerned, namely the reliability, authenticity, and preservation of records. Secondly, the registry does not capture information about schemas at the element level, a proposition that can become rapidly unsustainable, but rather remains at higher level of analysis focusing on element sets within schemas (in other words, it is not attempting to capture detailed information about the elements, attributes, relationships and so forth for each schema registered). Thirdly, the metadata schema registry is intended to provide a service to users, both InterPARES researchers and anyone in the archival or other metadata creating communities, above and beyond schema registration, as an analytical and evaluative tool. In this respect, its purpose is to support the assessment, development and extension of metadata schemas that will support the creation and preservation of trustworthy records. The specific goals of the InterPARES Metadata Schema Registry are as follows:

1. To register unambiguously, relevant metadata schemes and sets; to evaluate each against the Benchmark and Baseline requirements generated by InterPARES 1; and to make recommendations for how each 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.
2. To provide a standardized analytical framework whereby any metadata schema or local application profile could be assessed for its ability to address these needs, that will be incorporated into the draft Recordkeeping Metadata Standard (ISO 23081) being developed by ISO working group (ISO TC46/SC11-WG1).
3. To identify an overall set of metadata requirements that specify what metadata needs to be created, how, and by whom at all points within the Chain of Preservation and Records Continuum Models being developed by the InterPARES2 Modeling Cross-Domain Group.
4. To develop a set of 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.
5. To generate analytical data (for example, the metadata schema registry can capture temporal aspects of schema development as well relationships between schemas, application profiles, and crosswalks).

Two distinct sets of activities underlie the development of the metadata schema registry. The first of these is the development of the analytical framework that is necessary in order to assess the extent to which existing or proposed schemas, applications profiles, and even metadata element sets address the requirements that have been identified by InterPARES and other sources of warrant for the creation and preservation of reliable and trustworthy records. The second is design of the actual registry system that will implement the analytical framework. Both of these are discussed in the following sections.

Analytical framework underlying the metadata schema registry

The primary set of conditions against which metadata schemas registered in the metadata schema registry are measured are the Benchmark and Baseline Requirements that were generated out of the InterPARES 1 Project (see Appendix 1). 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'[11] . 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

[11] 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.

> 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.[12]

The Benchmark and Baseline Requirements, however, were only expressed conceptually, and in narrative form, by InterPARES 1, and have not yet been operationalized for any kind of technological implementation, for example, as a set of logical propositions or production rules. Nor have the requirements been deconstructed in a way that would specify how other processes and metadata might help to meet them. For example, as mentioned earlier in this paper, how might the different types of relevant context be manifested or documented through metadata? Operationalizing these requirements, therefore, in terms of the extent to which they might be met through the development of a rigorous and thorough metadata regime, has been a major aspect of developing the analytical framework.

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), requirements were also derived from an examination of 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).[13] The Description Group has also been working closely through this process with the Technical Committee that is developing ISO/TS 23081-1:2004 Information and Documentation – Records

[12] Report of the Authenticity Task Force of the InterPARES Project, available: http://www.interpares.org.

[13] The Description Group is still working on how such divergences might best be addressed within the analytical framework.

Management Processes – Metadata for Records standard that is currently in review, and the current plan is to incorporate the final analytical framework into the implementation component of the standard. ISO 23081 is being designed as a technical specification that will serve as an extension to ISO 15489 and will help people to understand metadata from a records management and archival perspective. As with InterPARES 2, it does not seek to create a new metadata set specifically for archives and records management, but rather to think through what kinds of metadata might be needed, and how that metadata should be managed over time. The Description Group is also tracking some other standards that might have implications for the analytical framework, such as ISO TC 10 Technical Product Documentation and TC 171 Document Management Applications.

Conducting the analysis

The analysis of any given schema or application profile involves scrutinizing and evaluating it in light of the analytical framework discussed above.[14] It requires knowledge of both the reference instruments and the schema being analyzed. Documentation for the schema, if any, is examined in order to familiarize the analyst with element structure and semantics and provide background as to the schema's conceptual basis. Preparatory work for the analysis involves extracting a summary table of all major structural elements of a schema, including such basic information as element name, description, qualifiers, components, obligation (optional or mandatory), and repeatability.

The primary tool of the analysis process is an Analysis Worksheet, organized to systematically analyze schemas within seven sections:

1. General
2. Recordkeeping – General
3. Recordkeeping – Assessment against ISO 23081
4. Recordkeeping – Assessment against InterPARES Benchmark Requirements
5. Recordkeeping – Assessment against InterPARES Baseline Requirements

[14] For a more detailed discussion of the analytical process, see Evans, Joanne and Lori Lindberg, "Describing and Analyzing the Recordkeeping Capabilities of Metadata Sets," *Proceedings of the International Conference on Dublin Core and Metadata Applications 2004, Shanghai, October 11–14 October 2004* (forthcoming).

6. Recordkeeping – Classification of Recordkeeping Metadata by Purpose
7. General Comments

The General section of the Analysis Worksheet captures basic descriptive information about the schema in the schema's own terminology, including the types of entities and objects it describes, the expected method(s) of metadata capture, the nature of the metadata and the different types of metadata within the schema. The Recordkeeping – General section identifies the specific types of *recordkeeping* entities the schema could be used to describe. The Recordkeeping – Assessment sections evaluate the schema's ability to satisfy the requirements for metadata about records, the benchmark requirements supporting the presumption of authenticity of electronic records and the baseline requirements supporting the production of authentic copies of electronic records, respectively. For purposes of assessing the degree to which a metadata schema satisfies a specific metadata requirement, a scale ranging from None to Comprehensive is used. The scale aims to distinguish whether a metadata schema minimally, adequately, or more comprehensively addresses a particular recordkeeping metadata requirement and is based on a judgment on the degree to which a schema allows for the naming, description, and documentation of relationships amongst recordkeeping entities. The Classification of Recordkeeping Metadata by Purpose section identifies what recordkeeping metadata *purposes* a metadata set meets, and the extent of that support. The General Comments section consists of comments by the analyst about the schema and the analysis in general.

The metadata schemas (and the different versions thereof) being analyzed by researchers in the Description Group are identified in a variety of ways: some have been identified through the case studies undertaken by the InterPARES Focus Groups as well as by specialized studies undertaken by the Description Group; in the course of their work, the Focus and Domain Groups also encounter other metadata schemas that are relevant to their area and pass these along to the Description Group; the Description Group identifies additional metadata schemas, element sets, and guidelines that relate to electronic records management, preservation, resource discovery, digital rights management, and so forth that may have implications for the creation and preservation of reliable and authentic records; and since archival description is probably the pre-eminent type of metadata used to ensure the authenticity of archival records, all the major archival description rules, schemas, and related practices (e.g., ISAD(G)/ISAAR, EAD/EAC/DACS, RAD, and the Australian Series System) are being analyzed.[15]

Some of the issues that have been encountered so far in analysing metadata schemas include, first of all, how to assess "fitness of purpose" of metadata schemas. In other words, how much or how little relevance to they need to have in terms of how they address (or fail to address, when it would be expected that they should) issues of reliability, authenticity, and preservation? Secondly, schemas vary in the amount of documentation available about them, and this can hamper the analysis effort. A third issue is how to quantify the degree to which a schema meets a particular requirement.

An example of analysis: encoded archival description (EAD)

For EAD, the initial summary table extraction identified a metadata set with 146 elements. Of the 146, only eight are mandatory. Thirty-six elements, some with qualifiers, are the minimum recommended by the authors of the schema for its primary use, the online representation of archival finding aids. The general section of the analysis worksheet identifies EAD as a metadata set that provides markup of descriptive elements contained in archival inventories and registers, describing an archival collection and digital surrogates of items in the collection. The Society of American Archivists, in conjunction with the Library of Congress, maintains EAD, which has been adopted as a professional descriptive standard in the United States. The schema encapsulates the entity(-ies) it describes and its expected method of metadata creation is manual, created by experts; in this case,

[15] For example, identified schemas include the Metadata Encoding and Transmission Standard (1.2) and Metadata Encoding and Transmission Standard (1.3) ;the Australian Recordkeeping Metadata Schema (1.00) ; the New South Wales Recordkeeping Metadata Standard (1.0); the Recordkeeping Metadata Standard for Commonwealth Agencies (1.0); the South Australian Recordkeeping Metadata Standard (2.4); the VERS Metadata Scheme (2); the Record-Keeping Metadata Requirements for the Government of Canada (January 2001); the Arizona Electronic Recordkeeping Systems (ERS) Guidelines – IV Functional Requirements for Recordkeeping Systems (2.0); the Minnesota Recordkeeping Metadata Standard (1.2); the PERM Preservation Attributes (December 2002); GILS, ISO 82045-2 Document Management Metadata;, CEDARS metadata specification for preservation, MARC; Digital Rights Metadata – 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; PREMIS metadata set (forthcoming), NISO Z39.87-2002 AIM 20-2002 Data Dictionary – Technical Metadata for Still Images, Metadata for Images in XML (MIX); and a range of geospatial metadata standards.

archivists. (Some automation has been introduced into the process by various implementers of the schema, but it does not support a particular automated method.)

Results of the recordkeeping – general analysis reveal that some elements of the schema are able to minimally address recordkeeping metadata describing essential recordkeeping entity classes such as Agents (those responsible for creating, controlling and managing records) and some Record Objects (particularly Record Aggregations and Collective Archives), but other entity classes such as Business Rules, Policies or Mandates, Business Activities or Processes, and Records Management/Recordkeeping Business Processes are either not supported or minimally supported. Evaluating the schema against the ISO recordkeeping standards, it does minimally or adequately satisfy a number of the requirements for basic metadata about records, such as structural and storage metadata, accessibility metadata, and security metadata. However, metadata about business rules, policies or mandates, particular agent metadata, business process metadata and other important metadata requirements such as views of records management metadata (business views, records management views, and use views) are not supported. Analysis of the schema against the InterPARES Benchmark and Baseline requirements exposes even further limitations of the schema for recordkeeping.

Metadata schema registry design considerations and structure

As discussed above, the metadata schema registry's primary purpose and application is to 'act as a data collection and analysis tool to support comparative studies of descriptive schemas.'[16] In this capacity, the registry needs to support and to ease the tasks of registering, describing, and comparing schemas, and to support the analysis and conclusion-making processes of InterPARES researchers and others using the registry for their own purposes. Thus, the metadata registry schema structure and data formatting need to be conducive to the streamlining of data entry, and also to enable complex manipulation and discovery of analytical data. Since a central notion of a metadata registry is that it engenders trust, it is essential that the metadata schema registry have not only quality control but also integrity. This means that it must support a discussion and/or declaration of how analytical values were arrived at and disclosure of instances where a

[16] Evans, Joanne and Rouche, Nadav "Development of Metadata Schema Registry", version 1.3, revised April 27th 2004.

value or deconstruction is contested. The implications of this for the design of the registry are that the processes whereby, and the principles upon which assessments have been made about individual schemas or element sets are disclosed and openly available to any user who wishes to review and comment upon them.

The development of the metadata schema registry has employed an iterative design process, whereby prototypes of components of the registry have been developed and tested against pilot analyses of selected metadata schemas that were also used in the development of the analytical framework, and then refined. In order to ensure efficient workflow, it was decided that analysis of schemas would proceed even before the registry prototype was completed, and would be integrated at that point. Precisely how and to what extent that integration should be achieved raises issues related to both technical capabilities and integrity and transparency. It has, therefore, remained one of the design questions that needs to be addressed during the development (see further discussion below). In terms of the technical development of the registry, it was also decided that the back and front ends would be developed separately. Priority was given to back-end development in order to create the element hierarchy and identify attributes and mappings to the relevant value spaces so that analyzed schemas could be integrated as expeditiously as possible. The front end is still in the process of being developed and is anticipated to take longer to build and refine, in part because some user evaluation may be required, and in part because it is being asked to support several complex interactive functions. Realistically, however, since the entire registry design is a circular feedback process and since it helps to visualize a system and its functionalities and not just to work from abstract conceptual functions, the designers have gone back and forth between interface, hierarchy, requirements, and system development.

Fields

The metadata registry schema prototype currently includes approximately 120 fields organized hierarchically using XML encoding. The first level of the hierarchy comprises 11 elements: Registration, Identification, Accessibility, Rights, Provenance, Description, Analysis, Documentation, Relationships, Administration, and a general Note element. These elements are further broken down into sub-elements, going three levels deep (i.e., up to sub-sub-elements). In the early phase of development, the hierarchy also included a fourth level, but the iterative process of development of the schema structure quickly

Table I. Metadata registry schema hierarchy and assigned values sample

Hierarchy			Controlled vocabularies
Elements	Sub-elements	Sub-sub-elements	Assigned values
Accessibility			
	Hardware		
			Authoring
			Viewing
			Validating
			Retrieval
			Manipulation
			Parsing
	Software	Software type	Operating system

revealed that using the elements structure to capture data at that level of granularity was overly rigid. While the top three levels of the hierarchy reached a fairly stable form early on in the development process, the fourth level kept changing substantially as additional fields were constantly added, changed or deleted. In order to allow for a stable hierarchical structure, the use of controlled vocabularies was introduced to replace the fourth level of the hierarchy, provide more flexibility in design, and allow for seamless and continuous updating. By using controlled vocabularies as XML-assigned values to third-level elements, it became possible to make changes to assigned values without affecting the hierarchy structure, thus enabling researchers to make changes easily and continuously to the controlled vocabularies at later stages of development. This strategy also allowed for changes in a functional system to be easily implemented, without having to change database tables or the underlying XML schema structure (see Table I). The design process also has had to take into account the ISO/IEC 1179 Information Technology – Metadata Registry (MDR) standard which provides guidance for formulation of data definitions and the naming and identification of administered items.[17]

The resulting hierarchy has been turned into an XML DTD which groups elements into the following categories:

- Registration – data elements to register metadata schema into the registry, such as: registration number, date and action officer;

[17] ISO/IEC 1179 Information Technology – Metadata Registries (MDR). Available: http://metadata-stds.org/1179/

- Identification – data elements to identify and distinguish metadata schema, such as: title, unique global identifier, version, and publication statements;
- Description – data elements to capture the purpose, scope, and jurisdiction of a metadata schema, including the types of entities and objects the schema describes. This category is further subdivided into a description section, which describes the metadata schema from the viewpoint of the providers of the schema; and an Analysis section, which describes the schema from a records and archival management perspective;
- Rights – data elements to capture intellectual property rights associated with the use of a metadata schema;
- Provenance – data elements to capture organizations or other bodies/agents associated with the development, publication and maintenance of a metadata schema;
- Documentation – data elements for capturing citations to the documentation of a metadata schema, such as specifications or guidelines;
- Relationships – data elements to capture relationships amongst metadata schema and to other classification schemes;
- Accessibility – data elements to capture information relating to the accessibility of a schema, e.g. hardware and software requirements and character encoding;
- Analysis – data elements for capturing the results of analysis of a metadata schema against recordkeeping requirements;
- Administration – data elements for the administration of the schema registry.

Analysis element

The analysis element of the metadata registry schema currently includes 15 sub-elements, which represent only a fraction of the analysis tool developed by researchers. The complete analysis tool, which forms the basis for the analytical framework, is currently in the form of a worksheet comprising approximately 40 questions and associated sub-questions or comments (Table II). One reason for this discrepancy is that in its current form, the questions on the worksheet are too long and wordy to fit into a concise element structure, and some additional work of labeling and restructuring still needs to be conducted in order to achieve that goal. However, this is even more a reflection of the analysis process itself being something that cannot easily be reduced to simple elements and values (that is, by necessity

Table II. Analysis worksheet, question 3.8

Questions	Comments
ISO 23081 Section 6.3 Perspectives of records management metadata	Indicate what entities/elements capture context, content, structure and appearance initially and through time
Does the schema capture metadata relating to the initial context, content, structure and appearance? Does it allow for the capture of context, content, structure and appearance metadata through time?	Indicate degree using none, minimal, adequate and comprehensive qualifiers

the questions are long and wordy and in some cases the answers may be too). Hence the summary document becomes a key in informing the metadata registry structure. Additionally, in terms of workflow, using a local document seems to be more efficient than entering the analysis data directly into a remote database. A number of strategies can be applied in order to resolve this discrepancy.

Strategy 1 – Formatting the complete analysis worksheet to fit into the schema registry element structure

From a discovery and retrieval perspective, it would be beneficial to be able to perform complex searching and manipulation on the analysis data, such as retrieving all schema records that contain specific values within one or more analysis fields; producing graphs using quantifiable data across records; and comparing the analysis work produced by different researchers pertaining to the same schema to ensure inter-coder reliability. Using this strategy, all the analysis data would be properly formatted and available for data entry, search and retrieval. However, beyond the non-trivial task of adapting the analysis worksheet questions to the element structure, additional difficulty lies in creating an interface and workflow that supports the analysis process and data entry as opposed to making it more cumbersome.

Strategy 2 – Automatic ingestion of the complete analysis worksheet without any formatting

As an alternate solution, instead of formatting the analysis worksheet in order to fit into the element structure and supporting data entry, an automatic ingestion mechanism of completed worksheets could be put in place without making any changes to the worksheet structure and labeling so that the data is available as is for manipulation by

Table III. Analysis summary worksheet, section 'Analytical framework underlying the metadata schema registry': recordkeeping – assessment against InterPARES Baseline Requirements

B1: Controls over records transfer, maintenance and reproduction B2: Documentation of reproduction process and its effects B3: Archival description	[Yes, no, possible]	[Statement]

researchers. This strategy would be less work-intensive since it would not require the Analysis Worksheet to be redesigned. However, for that same reason, its current data format and labeling would be less conducive to search and manipulation, since it was not designed with that purpose in mind. Additionally, this strategy would not provide for a data entry mechanism, and the researchers would continue to use the worksheet to enter data.

Strategy 3 – Formatting a partial analysis worksheet to fit into the schema registry element structure

In addition to the Analysis Worksheet, an Analysis Summary was also devised, which is basically a partial analysis worksheet. It consists of seven sections and 40 sub-sections (Table III), which could be translated fairly easily into a hierarchical element structure and be integrated as such into the registry schema structure. For example, Section 'Analytical framework underlying the metadata schema registry' (Recordkeeping – Assessment against InterPARES Baseline Requirements) could be translated into three sub-elements (Control, Documentation, Archival Description) of the Analysis element in the registry.

At this point, in the development of the metadata schema registry, an assessment is being conducted in order to determine which solution to adopt. While adapting the Analysis Worksheet to an efficient data entry format is the most comprehensive strategy, it is also the most work-intensive.[18] Automatic ingestion of the completed worksheet requires significantly less investment, but it would also be less efficient in terms of the ability to search and manipulate the data. Finally, solely integrating the Summary Worksheet would be a fairly simple task, but it would also mean that only a part of the data is available for manipulation. In this case, the completed Analysis Worksheet would still be available as a standalone document linked to the corresponding schema record.

[18] Gilliland-Swetland, Anne. "Report of the Description Group International Team Meeting", Dublin, June 2004.

Information retrieval and report

In order to support the analysis and conclusion-making processes, some types of advanced search functionalities have been identified. Users should be able to confine their search to specific elements or combinations of elements in order to retrieve schemas according to their queries. Such elements would need to be identified for their usefulness. For example, searching by version of schemas, by relationships between schemas, or by schema classification, would be quite useful. Additionally, users should be able to select terms from the controlled vocabularies developed for the schema to ease the selection of query terms, and retrieve records if a specific field is empty in order to be able to conduct queries such as "retrieve all schemas that do not have rights management elements". For workflow management and administration purposes, users should be able to retrieve schemas according to their workflow status (such as "completed", "reviewed" and "published"), as a well as retrieve schemas by the action officer who registered the schema. Finally, there needs to be a publicly available Frequently Asked Questions (FAQ) section and a user comment or feedback mechanism.

Beyond generating reports on the extent to, and ways in which a given schema or application profile, or combination of schemas meet the requirements identified in the registry, or range of other potentially interesting data and analyses could be generated from the registry. For example, metadata schemas are in many ways actualizations of the perspectives of different communities of practice, and the registry might provide a tool for analysing those perspectives.[19] Moreover, metadata schemas also encode conceptualizations of core concepts such as authenticity that may differ from those of archivists and might be similarly examined. The registry can also demonstrate when issues that are of concern to the archival community have become sufficiently important to other communities that they are addressed through their metadata schemas, either explicitly, or implicitly.

Development of a metadata specification model

One additional role that the metadata schema registry is playing within InterPARES 2 is to provide input into an overall set of metadata

[19] Etienne Wenger defines communities of practice as groups of people who share a concern or a passion for something they do and who interact regularly to learn how to do it better. See Wenger, Etienne. *Communities of Practice* (Cambridge: Cambridge University Press, 1999).

requirements (not, however, a metadata set) that specify what metadata needs to be created, how, and by whom at all points within the two sets of activity and entity models that are being developed by the InterPARES 2 Modeling Cross-Domain Group. These models are designed to reflect both the records life cycle and continuum approaches to the MoReq and it will be illuminating to see whether one overall set of metadata requirements will meet the needs of both approaches. Based upon these requirements, and how they map onto the two sets of models "walkthroughs" of selected case studies conducted by InterPARES 2 Focus Groups against the models, a set of specifications will be drawn up around which automated tools can be developed that can be used to assist with the creation, capture, management, preservation, and end-user manipulation of essential metadata for active and preserved records. The draft specifications will be tested against metadata scenarios to be developed by the Description group and then refined accordingly.

Surfacing contested issues

Much of this paper has been addressing, explicitly or implicitly, ideas about trust. There is considerable rhetoric in the archival literature about archival repositories serving as "trusted repositories," whether they function in a custodial or a non-custodial paradigm. Luciana Duranti writes of this function that:

> "... acceptance into custody [i.e., into the archival repository] is more than a declaration of authenticity. It is taking responsibility for preserving that authenticity, and it requires taking the appropriate measures for guaranteeing that authenticity will never be questioned, measures that go much beyond physical security."[20]

It is interesting that those communities who are now developing metadata registries also cite the creation of trusted repositories as being an over-arching goal. It must be recognized that both declarations arise, at least in part, out of the problems of controlling integrity across time, as well as ideas about how these repositories may be able to assure such control by operating outside of the immediate interests and day-to-day operational concerns of the creators of either records or metadata.

At points, this paper has raised the need, in terms of ensuring trust in the metadata schema registry, for there to be transparency for

[20] Duranti, Luciana. "Archives as a Place," *Archives and Manuscripts* 24 no. 2 (1996): 247.

external users regarding these processes and understandings. While any discussion of data analysis and systems design, by the time they are written up for publication, can sound quite neat and clean, the processes underlying those activities demand that any differences of interpretation and underlying ambiguities in semantics, theoretical understandings, and so forth, be disclosed and resolved. In the course of the development of the metadata schema registry and analytical framework, two contested spaces have been surfaced and this paper will conclude with a discussion of the dimensions of these contested spaces in more general terms, and then how they are being addressed in the context of the work of the InterPARES Description Group as described above.

Metadata vs. archival description

The first of these contested areas relates to differing definitions and their underlying theoretical conceptualizations. It is reflected in the nomenclature of the Description Group itself and relates to whether archival description is merely one type of the several forms of metadata required to ensure the preservation and intellectual accessibility of authentic archival records (the perspective ascribed to by these authors), or whether archival description, as generated by archivists, performs a distinct descriptive function that makes it the primary means by which the continued authenticity of archival records can or should be assured.[21] Duranti discusses the role of archival description in the following terms:

> "The identification of the documents, the assignment to them of an intellectual and physical place in the whole of the authentic documents, that is, their location and description in context, by freezing and perpetuating their interrelationships, ensure that possible tampering will be easy to identify."[22]

In other words, where and when should the emphasis fall, upon *metadata* created continuously across the life of the record, or upon *archival description* that occurs once the record has been transferred into the intellectual control (and, usually, physical custody) of the archivist?

Metadata, a term first used in geospatial and data processing communities to refer to "data about data" (essentially salient information

[21] See, for example, MacNeil, Heather, "Metadata Strategies and Archival Description: Comparing Apples and Oranges," *Archivaria* 39, 1995: 11–21.

[22] Duranti, ibid. 247.

about scientific data that was not the data itself), was largely appropriated in the 1990s by the bibliographic community and narrowed to refer to value-added information about a resource that was contributed primarily by catalogers. Since then, the term has been used increasingly to refer to any kind of descriptive or resource discovery information, whether manually or automatically created. In this respect, recordkeeping metadata, as expressed through models such as RKMS include event and process-related metadata and has a purview that is actually quite a bit broader than how metadata is conceptualized in many other communities. At the same time, however, those who would argue that archival description is not metadata take a different view from either perspective, positing that metadata accrues to the record during the processes of its creation and active life and that archival description is both the manual process and the product that occurs when a record is transferred into archival custody. From this viewpoint, archival description not only creates a value-added description that documents the various contexts of the archival record, but is the mechanism through which the record is intellectually incorporated into the archival bond and also performs an important role in ensuring the continued authenticity of that record.

While it is beyond the scope of this particular paper to go into this issue in depth, and it certainly could be argued that realistically, both positions are not that far apart conceptually if not terminologically, two things are certain. In the future, time and cost concerns as well as new technological capabilities will ensure that even archival description may be created, at least partially, by automated means, likely including through harvesting and re-purposing metadata created by others prior to the records coming into archival custody. Equally certain is that if a more comprehensive and rigorously delineated metadata infrastructure (including archival description) is integral to maintaining and demonstrating the trustworthiness of the records as well as to developing archival roles and services of the future, as is argued in this paper, there is a concomitant need for overt integrity control and transparency regarding archival metadata. What the latter would require is a more critical and reflexive approach on the part of archivists in terms of declaring and analyzing the assumptions that are embedded within their metadata schemas, elements, crosswalks, application profiles, descriptive standards and best practice guidelines. Archivists must be cognizant that the accession records, finding aids, and use records they typically create today are not only all part of the archival metadata for the records to which they relate, but they are also records in their own rights. The scrutiny, therefore, that archivists give to the records and recordkeeping metadata of others in order to assess and validate their management and reliability, they must also give to their own (hopefully, in the future, with the assistance of the analytical framework and metadata schema registry discussed in this paper). Archivists must also ponder the implications of metadata and metadata schemas in terms of the metanarratives they have encoded within them – the assumptions that they unconsciously bring to bear in everything they do. How might these facilitate the silencing or liberation of voices in the records and their creating communities in terms of addressing submerged contexts, tacit narratives and counter-narratives as well as what is overt on the face of the record? Archivists must be prepared to act if they identify ways in which their metadata are excluding voices and narratives. Creating unambiguous metadata may also necessitate recognizing who might be considered as co-creators of the records-perhaps through actions as obvious as documenting co-creating roles for collaborators in a shared workspace, supporting notions of parallel provenance,[23] or as politically charged as acknowledging the co-creation, or at least co-ownership of the subjects of records. Consider, for example, the cases of indigenous peoples for whom government records may be the only available evidence of government programs or entitlements.

As for the metadata schema registry and analytical framework, by drawing upon requirements that had been published in a variety of archival contexts, it was necessary for the Description Group to be receptive to any information that lay outside the records themselves might be necessary to ensure their continued trustworthiness as they move through time and space. Hence, a range of metadata schemas and application profiles are being analyzed, some of which overtly relate to archival description, and others to metadata conceived in other terms that may have element sets that explicitly or implicitly address the concerns of InterPARES, among them description.

Records life cycle vs. records continuum models

The second issue surfaced by the Description Group's work regards the extent to which both the life cycle and continuum views on

[23] Recent presentations by Chris Hurley have been exploring the parameters of parallel provenance as a potential archival construct. See for example, Chris Hurley, "Documentation, Parallel Provenance and Archival Issues of Concern," Seminar on Archives and Collective Memory: Challenges and Issues in the Pluralised Archival Role, The Recordkeeping Institute and School of Information Management and Systems, Monash University, August 2004; and Chris Hurley, Workshop on Relationships in Records, Center for Information as Evidence, University of California, Los Angeles, August 2004.

archival roles can be respected or even reconciled through the analytical approach embedded in the metadata schema registry. 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. In the absence of hard data to establish the lack of viability of any perspective or practice, it would seem to be arrogant to attempt to vanquish any perspective or practice. That appears to leave two alternatives – one being the toleration and supporting of more than one approach, the other being an attempt to reconcile approaches that appear at first, and maybe even second glance, to be irreconcilable.

The Description Group is, in many ways, attempting to straddle both of these alternatives, although, in the interests of full disclosure, it needs to be said that the work of the group has turned out to be situated closer to the conceptualizations of recordkeeping metadata expressed by RKMS and supported by continuum thinking. This has been an inevitable consequence of taking the position that many different types of metadata are needed to satisfy even the InterPARES Benchmark and Baseline Requirements, and that archival description is indeed one of these types of metadata. Moreover, having made a conscious decision to assess the metadata implications of both of the dominant existing models, the relative extensiveness of the Continuum model, with the dimensionality afforded by its four axes of identity, evidentiality, transactionality and recordkeeping entity, necessitated that the Description Group 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* (rather than the archivist), but were still predicated upon the physical transfer of records into an archival repository. However, the Description Group 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.[24]

Working together with the InterPARES2 Modeling Group, which has been developing activity and entity models for both the Records Life Cycle ("Chain-of-Preservation") and a Continuum approaches to managing records, the Description Group is identifying the points when metadata are or should be created within each set of models. However, it then intends to look at both sets of outcomes side-by-side with the intent of seeing whether or not they really are mutually exclusive, and whether or not it is viable to fold them into a unified metadata specification model that could be used within any existing or future recordkeeping approach.

Appendix 1. Benchmark Requirements Supporting the Presumption of Authenticity of Electronic Records[25]

Preamble

The benchmark requirements are the conditions that serve as a basis for the preserver's assessment of the authenticity of the creator's electronic records. Satisfaction of these benchmark requirements will enable the pres erver to infer a record's authenticity on the basis of the manner in which the records have been created, handled, and maintained by the creator.

Within the benchmark requirements, Requirement A.1 identifies the core information about an electronic record – the immediate context of its creation and the manner in which it has been handled and maintained – that establishes the record's identity and lays a foundation for demonstrating its integrity. Requirements A.2–A.8 identify the kinds of procedural controls over the record's creation, handling, and maintenance that support a presumption of its integrity.

[24] 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.

[25] Available: http://www.interpares.org

Benchmark Requirements (Requirement Set A)

	To support a presumption of authenticity the preserver must obtain evidence that:
Requirement A.1: Expression of Record Attributes and Linkage to Record	The value of the following attributes are explicitly expressed and inextricably linked to every record. These attributes can be distinguished into categories, the first concerning the identity of records, and the second concerning the integrity of records. *A.1.a* Identity of the record: *A.1.a.i* Names of the persons concurring in the formation of the record, that is: ●name of author[26] ●name of writer[27] (if different from the author) ●name of originator[28] (if different from name of author or writer) ●name of addressee[29] *A.1.a.ii* Name of action or matter *A.1.a.iii* Date(s) of creation and transmission, that is: ●chronological date[30] ●received date[31] ●archival date[32] transmission date(s)[33]

[26] The name of the physical or juridical person having the authority and capacity to issue the record or in whose name or by whose command the record has been issued.

[27] The name of the physical or juridical person having the authority and capacity to articulate the content of the record.

[28] The name of the physical or juridical person assigned the electronic address in which the record has been generated and/or sent.

[29] The name of the physical or juridical person(s) to whom the record is directed or for whom the record is intended.

[30] The date, and possibly the time, of compilation of a record included in the record by the author or the electronic system on the author's behalf.

[31] The date, and possibly the time, when a record is received by the addressee.

[32] The date, and possibly the time, when a record is officially incorporated into the creator's records.

[33] The date and time when a record leaves the space in which it was generated.

	To support a presumption of authenticity the preserver must obtainevidence that:	
	A.1.a.iv	Expression of archival bond[34] (e.g., classification code, file identifier)
	A.1.a.v	Indication of attachments
	A.1.b	Integrity of the record:
	A.1.b.i	Name of handling office[35]
	A.1.b.ii	Name of office of primary responsibility[36] (if different from handlingoffice)
	A.1.b.iii	Indication of types of annotations added to the record[37]
	A.1.b.iv	Indication of technical modifications;[38]
Requirement A.2: Access Privileges	The creator has defined and effectively implemented access privileges concerning the creation, modification, annotation, relocation, and destruction of records;	
Requirement A.3: Protective Procedures: Loss and Corruption of Records	The creator has established and effectively implemented procedures to prevent, discover, and correct loss or corruption of records;	

[34] The archival bond is the relationship that links each record, incrementally, to the previous and subsequent ones and to all those participate in the same activity. It is originary (i.e., it comes into existence when a record is made or received and set aside), necessary (i.e., it exists for every record), and determined (i.e., it is characterized by the purpose of the record).

[35] The office (or officer) formally competent for carrying out the action to which the record relates or for the matter to which the record pertains.

[36] The office (or officer) given the formal competence for maintaining the authoritative record, that is, the record considered by the creator to be its official record.

[37] Annotations are additions made to a record after it has been completed. Therefore, they are not considered elements of the record's documentary form.

[38] Technical modifications are any changes in the digital components of the record as defined by the Preservation Task Force. Such modifications would include any changes in the way any elements of the record are digitally encoded and changes in the methods (software) applied to reproduce the record from the stored digital components; that is, any changes that might raise questions as to whether the reproduced record is the same as it would have been before the technical modification. The indication of modifications might refer to additional documentation external to the record that explains in more detail the nature of those modifications.

	To support a presumption of authenticity the preserver must obtain evidence that:
Requirement A.4: Protective Procedures: Media and Technology	The creator has established and effectively implemented procedures to guarantee the continuing identity and integrity of records against media deterioration and across technological change;
Requirement A.5: Establishment of Documentary Forms	The creator has established the documentary forms of records associated with each procedure either according to the requirements of the juridical system or those of the creator;
Requirement A.6: Authentication of Records	If authentication is required by the juridical system or the needs of the organization, the creator has established specific rules regarding which records must be authenticated, by whom, and the means of authentication;
Requirement A.7: Identification of Authoritative Record	If multiple copies of the same record exist, the creator has established procedures that identify which record is authoritative;
Requirement A.8: Removal and Transfer of Relevant Documentation	If there is a transition of records from active status to semi-active and inactive status, which involves the removal of records from the electronic system, the creator has established and effectively implemented procedures determining what documentation has to be removed and transferred to the preserver along with the records.

Appendix 2. Baseline Requirements Supporting the Production of Authentic Copies of Electronic Records[39]

Preamble

The baseline requirements outline the minimum conditions necessary to enable the preserver to attest to the authenticity of copies of inactive electronic records.

[39] Available: http://www.interpares.org

Baseline Requirements (Requirement Set B)

	The preserver should be able to demonstrate that:
Requirement B.1: Controls over Records Transfer, Maintenance, and Reproduction	The procedures and system(s) used to transfer records to the archival institution or program; maintain them; and reproduce them embody adequate and effective controls to guarantee the records' identity and integrity, and specifically that **B.1.a** Unbroken custody of the records is maintained; **B.1.b** Security and control procedures are implemented and monitored; and **B.1.c** The content of the record and any required annotations and elements of documentary form remain unchanged after reproduction.
Requirement B.2: Documentation of Reproduction Process and its Effects	The activity of reproduction has been documented, and this documentation includes ***B.2.a*** The date of the records' reproduction and the name of the responsible person; **B.2.b** The relationship between the records acquired from the creator and the copies produced by the preserver; **B.2.c** The impact of the reproduction process on their form, content, accessibility and use; and **B.2.d** In those cases where a copy of a record is known not to fully and faithfully reproduce the elements expressing its identity and integrity, such information has been documented by the preserver, and this documentation is readily accessible to the user;
Requirement B.3: Archival Description	The archival description of the fonds containing the electronic records includes – in addition to information about the records' juridical-administrative, provenancial, procedural, and documentary contexts – information about changes the electronic records of the creator have undergone since they were first created.

Commentary on the Benchmark Requirements Supporting the Presumption of Authenticity of Electronic Records

The assessment of the authenticity of the creator's records takes place as part of the appraisal process. That process and the role of the benchmark requirements within it are described in more detail in the "Appraisal Task Force Report." This assessment should be verified when the records are transferred to the preserver's custody.

A.1. Expression of Record Attributes and Linkage to Record

The presumption of a record's authenticity is strengthened by knowledge of certain basic facts about it. The attributes identified in this requirement embody those facts. The requirement that the attributes be expressed explicitly and linked inextricably[40] to the record during its life, and carried forward with it over time and space, reflects the task force's belief that such expression and linkage provide a strong foundation on which to establish a record's identity and demonstrate its integrity. The case studies undertaken as part of the work of the task force revealed very little consistency in the way the attributes that specifically establish the identity of a record are captured and expressed from one electronic system to another. In certain systems, some attributes were explicitly mentioned on the face of the record; in others they could be found in a wide range of metadata linked to the record or they were simply implicit in one or more of the record's contexts. In many cases, certain attributes (e.g., the expression of the archival bond) were not captured at all. The task force's concern is that, in the absence of a precise and explicit statement of the basic facts concerning a record's identity and integrity, it will be necessary for the preserver to acquire enormous, and otherwise unnecessary, quantities of data and documentation simply to establish those facts.

The link between the record and the attributes listed in Requirement A.1 is viewed by the task force as a *conceptual* rather than a *physical* one, and the requirement could be satisfied in different ways, depending on the nature of the electronic system in which the record resides. For example, in ERMS, this requirement is usually met through the creation of a record profile.[41] In other types of systems, the requirement could be fulfilled through a topic map. A topic map expresses the characteristics (i.e., *topics*) of subjects (e.g., records or record attributes) and the relationships between and among them.

When a record is exported from the live system, migrated in a system update, or transferred to the preserver, the attributes should be linked to

[40] For the purposes of this requirement, inextricable means incapable of being disentangled or untied, and link means a connecting structure.

[41] If the attribute values contained in the profile are also expressed independently as entries in a register of all records made or received by the creator, then, in addition to establishing the identity and supporting the inference of the integrity of the record, they would corroborate such identity and strengthen the inference of integrity.

the record and available to the user. When pulling together the data prior to export, the creator should also ensure that the data captured are the right data. For example, in the case of distribution lists, the creator must ensure that if the recipients specified on "List A" were changed at some point in the active life of records, the accurate "List A: Version 1" is exported with the records associated with the first version, and that the second version is sent forward with those records sent to recipients on "List A: Version 2."

A.2. Access Privileges
Defining access privileges means assigning responsibility for the creation, modification, annotation, relocation, and destruction of records on the basis of competence, which is the authority and capacity to carry out an administrative action. Implementing access privileges means conferring exclusive capability to exercise such responsibility. In electronic systems, access privileges are usually articulated in tables of user profiles. Effective implementation of access privileges involves the monitoring of access through an audit trail that records every interaction that an officer has with each record (with the possible exception of viewing the record). If the access privileges are not embedded within the electronic system but are based on an external security system (such as the exclusive assignment of keys to a location), the effective implementation of access privileges will involve monitoring the security system.

A.3. Protective Procedures: Loss and Corruption of Records
Procedures to protect records against loss or corruption include: prescribing regular back-up copies of records and their attributes; maintaining a system back-up that includes system programs, operating system files, etc.; maintaining an audit trail of additions and changes to records since the last periodic back-up; ensuring that, following any system failure, the back-up and recovery procedures will automatically guarantee that all complete updates (records and any control information such as indexes required to access the records) contained in the audit trail are reflected in the rebuilt files and also guarantee that any incomplete operation is backed up. The capability should be provided to rebuild forward from any back-up copy, using the back-up copy and all subsequent audit trails.

A.4. Protective Procedures: Media and Technology
Procedures to counteract media fragility and technological obsolescence include: planning upgrades to the organization's technology base; ensuring the ability to retrieve, access, and use stored records when components of the electronic system are changed; refreshing the records by regularly moving them from one storage medium to another; and migrating records from an obsolescent technology to a new technology.

A.5. Establishment of Documentary Forms

The documentary form of a record may be determined in connection to a specific administrative procedure, or in connection to a specific phase(s) within a procedure. The documentary form may be prescribed by business process and work-flow control technology, where each step in an administrative procedure is identified by specific record forms. If a creator customizes a specific application, such as an electronic mail application, to carry certain fields, the customized form becomes, by default, the required documentary form. It is understood that the creator, acting either on the basis of its own needs or the requirements of the juridical system, not an individual officer, establishes the required documentary form(s) of records.

When the creator establishes the documentary form in connection to a procedure, or to specific phases of a procedure, it is understood that this includes the determination of the intrinsic and extrinsic elements of form[42] that will allow for the maintenance of the authenticity of the record. Because, generally speaking, that determination will vary from one form of a record to another, and from one creator to another, it is not possible to predetermine or generalize the relevance of specific intrinsic and extrinsic elements of documentary form in relation to authenticity.

A.6. Authentication of Records

In common usage, to authenticate means to prove or serve to prove the authenticity of something. More specifically, the term implies establishing genuineness by adducing legal or official documents or expert opinion. For the purposes of the benchmark requirements, authentication is understood to be a declaration of a record's authenticity at a specific point in time by a juridical person entrusted with the authority to make such declaration. It takes the form of an authoritative statement (which may be in the form of words or symbols) that is added to or inserted in the record attesting that the record is authentic.[43] The requirement may be met by linking the authentication of specific types of records to business procedures and assigning responsibility to a specific office or officer for authentication.

The authentication of copies differs from the validation of the process of reproduction of the digital components of the records. The latter process occurs every time the records of the creator are moved from one medium to another or migrated from one technology to another.

[42] The extrinsic and intrinsic elements of form are defined and explained in the Authenticity Task Force's *Template for Analysis*, Appendix 1 <j app01>.

[43] The meaning of authentication as it is used by the Authenticity Task Force in this report is broader than its meaning in public key infrastructure (PKI) applications. In such applications, authentication is restricted to proving identity and public key ownership over a communication network.

A.7. Identification of Authoritative Record
An authoritative record is a record that is considered by the creator to be its official record and is usually subject to procedural controls that are not required for other copies. The identification of authoritative records corresponds to the designation of an office of primary responsibility as one of the components of a record retention schedule. The Office of Primary Responsibility is the office given the formal competence for maintaining the authoritative (that is, official) records belonging to a given class within an integrated classification scheme and retention schedule. The purpose of designating an Office of Primary Responsibility for each class of record is to reduce duplication and to designate accountability for records.

It is understood that in certain circumstances there may be multiple authoritative copies of records, depending on the purpose for which the record is created.

A.8. Removal and Transfer of Relevant Documentation
This requirement implies that the creator needs to carry forward with the removed records all the information that is necessary to establish the identity and demonstrate the integrity of those records, as well as the information necessary to place the records in their relevant contexts.

Commentary on the Baseline Requirements Supporting the Production of Authentic Copies of Electronic Records

The establishment and implementation of the baseline requirements take place as part of the function of managing preservation. The preservation function and the role of the baseline requirements within it are described in more detail in the "Preservation Task Force Report."

B.1. Controls over Records Transfer, Maintenance, and Reproduction
The controls over the transfer of electronic records to archival custody include establishing, implementing, and monitoring procedures for registering the records' transfer; verifying the authority for transfer; examining the records to determine whether they correspond to the records that are designated in the terms and conditions governing their transfer; and accessioning the records.

As part of the transfer process, the assessment of the authenticity of the creator's records, which has taken place as part of the appraisal process, should be verified. This includes verifying that the attributes relating to the records' identity and integrity have been carried forward with them (Requirement A.1), along with any relevant documentation (Requirement A.8).

The controls over the maintenance of electronic records once they have been transferred to archival custody are similar to several of the ones enumerated in the benchmark requirements. For example, the preserver should establish access privileges concerning the access, use, and reproduction of records (Requirement A.2); establish procedures to prevent, discover, and correct loss or corruption of records (Requirement A.3), as well as procedures to guarantee the continuing identity and integrity of records against media deterioration and across technological change (Requirement A.4). Once established, the privileges and procedures should be effectively implemented and regularly monitored. If authentication of the records is required, the preserver should establish specific rules regarding who is authorized to authenticate them and the means of authentication that will be used (Requirement A.6).

The controls over the reproduction of records include establishing, implementing, and monitoring reproduction procedures that are capable of ensuring that the content of the record is not changed in the course of reproduction.

B.2. Documentation of Reproduction Process and its Effects
Documenting the reproduction process and its effects is an essential means of demonstrating that the reproduction process is transparent (i.e., free from pretence or deceit). Such transparency is necessary to the effective fulfillment of the preserver's role as a trusted custodian of the records. Documenting the reproduction process and its effects is also important for the users of records since the history of reproduction is an essential part of the history of the record itself. Documentation of the process and its effects provides users of the records with a critical tool for assessing and interpreting the records.

B.3. Archival Description
Traditionally it has been a function of archival description to authenticate the records and perpetuate their administrative and documentary relationships. With electronic records, this function becomes critical. Once the records no longer exist except as authentic copies, the archival description is the primary source of information about the history of the record, that is, its various reproductions and the changes to the record that have resulted from them. While it is true that the documentation of each reproduction of the record copies[44] may be preserved, the archival description summarizes the history of all the reproductions, thereby obviating the need to preserve all the documentation for each and every reproduction. In this respect, the description constitutes a collective attestation of the authenticity of the records and their relationships in the context of the fonds to which the records belong. This is different from a certificate of authenticity, which attests to the authenticity of individual records. The importance of this collective attestation is that it authenticates and perpetuates the relationships between and among records within the same fonds.

[44] Although, technically, every reproduction of a record that follows its acquisition by the preserver is an authentic copy, it is the only record that exists and, therefore, should normally be referred to as "the record" rather than as "the copy."