A Metadata Schema Registry for the Registration and Analysis of Recordkeeping and Preservation Metadata

Anne Gilliland-Swetland
University of California, Los Angeles, California, USA
and
Sue McKemmish
Monash University, Melbourne, Australia

Introduction

After many years of research and practice, most archivists now acknowledge that one key, if not *the* key to the creation, management, preservation and use of electronic records (that is, those records that are created and/or maintained in electronic form as evidence of business or personal activity) is metadata. The term "metadata," as used in this paper, is based upon how it has recently been applied in the archives and recordkeeping community to refer to all types of structured information, including archival description, which is created manually or automatically and captured by recordkeeping and archival systems. Such metadata serves to document the juridical-administrative, business and technical contexts within which records are created, managed and used; identify records; delineate how the records behave, their function and use; identify and describe the relationships within and between records and other information objects and the ways in which these relationships evolve over time; express and support how records should be managed, and what should happen to them as they age (e.g. destruction or preservation requirements); and provide audit trails of recordkeeping processes.

Many metadata schemas have been developed and applied in recent years by archivists and other professionals engaged in electronic recordkeeping and the preservation and archiving of digital materials, for purposes such as records management, collection description, digitization of items for online access, and digital rights management. However, there remains a need to focus much more attention on the creation, management, preservation and use of metadata that is demonstrably trustworthy, and that is sufficient, appropriate, and of high enough quality to demonstrate the continued authenticity of the electronic records or archived digital materials to which it relates. Moreover, trusted metadata in and of itself can provide end users with an additional valuable information resource. What is required is a delineation, not only of what metadata needs to be created, but also how its integrity is to be guaranteed over time, how much of it needs to be preserved and what eliminated, and, in each case, when, by whom, and how. The benefits of such attention are not only that the integrity of metadata as it is created and accrued across time, space and activity, is ensured, but also that a technical and descriptive metadata infrastructure will be developed that could underpin the development of automated metadata management and manipulation tools to better support activities ranging from current electronic recordkeeping to collection management and the creation of entirely new views of archived materials.

This paper discusses the development by the Description Group of the InterPARES2 (International research on Permanent Authentic Records in Electronic Systems) Project of an XML-based prototype metadata schema registry and analytical framework for the identification, registration, and analysis of existing and prospective metadata schemas, sets, and application profiles relevant to electronic recordkeeping and digital preservation. InterPARES is an international multi-disciplinary research collaboration emanating out of the archival community that has been working since 1999 to devise new models, methods and automated tools for ensuring the creation of reliable, and preservation of authentic electronic records. The second phase of this project, InterPARES2, which is due to be completed in 2006, integrates the disciplinary perspectives and concerns of the scientific and digital arts communities, as well as those of e-government, and is focusing in particular on the preservation of records generated by emergent interactive, experiential and dynamic systems and processes.

Goals of the Metadata Schema Registry

The aims of the metadata schema registry differ from those of most metadata registries currently under development, both conceptually and in terms of content. Firstly, the registry comprises information on multiple metadata schemas and element sets that have been identified as having some relevance for the creation, management, preservation and use of trustworthy electronic records rather than serving

as an authoritative source of information about a single schema. Secondly, the registry only captures sufficient information about these schemas and element sets to be able to identify them definitively and distinguish between each variant version. It does not exhaustively register or describe at the element level. Thirdly, the registry contains descriptive information not only on all versions of specific metadata schemas and element sets, but also on existing crosswalks and other mappings between them, and, in some cases, also on local application profiles. Fourthly, the registry also implements an analytical framework based upon a set of requirements for creating reliable and preserving authentic electronic records (discussed below). By incorporating the analytical component, the registry can also serve as a tool to assist users in the identification of appropriate schemas and element sets that will address their specific needs for creating, managing, preserving and using trustworthy records. InterPARES2 researchers, and subsequently any other interested parties, will be able to evaluate existing schemas and element sets as well as to register and assess the extent to which their local application profiles and proposed schemas and element sets measure up against the requirements embedded in the analytical component of the registry. Fifthly, the registry also contains recommendations, derived from the analyses conducted by InterPARES2 researchers, on how each registered version of a schema, element set, or application profile might need to be extended or otherwise revised in order to address the reliability, authenticity and preservation needs of records created within the domain, community, sector, or institution to which they pertain.

Two important secondary aims underlie the building of the metadata schema registry. The first of these is to use what is learned as the basis for developing specifications for metadata management tools to be used in activities such as automatic metadata creation and extraction. The second is to feed the outcomes of this research to other relevant research and development activities such as the Clever Recordkeeping Metadata Project[1] and the development by the San Diego Supercomputer Center as part of its Persistent Archives Technology of metadata tools for the automated creation, harvesting, and end-user manipulation of metadata.

Developing the Analytical Framework

The analytical framework assesses how and the extent to which individual entries within the registry address a set of requirements for the creation, management, preservation and use of trustworthy records that has been derived from several sources considered by the researchers to be relevant. The framework, however, is designed to exist as a standalone tool as well as one that generates data to be ingested into the metadata schema registry, and InterPARES2 researchers are working in concert with the developers of the ISO 23081 Recordkeeping Metadata Standard (ISO TC46/SC11-WG1) to incorporate the framework into the implementation section of the standard. The Standard itself also provides a major source of the requirements used in the analytical framework to assess metadata schemas. This assessment framework enables statements to be made about how far any particular metadata schema complies with the requirements of the Standard.

Other sources used in the development of the framework were the result of previous empirical research, and others were existing archival and recordkeeping standards, several of which themselves emanated out of the previous research. The first of these sources is a set of Benchmark Requirements for the Creation of Reliable Electronic Records and Baseline Requirements for the Preservation of Authentic Electronic Records that were developed by the InterPARES1 Project based on a diplomatic analysis of requirements for authentic electronic records coupled with an analysis of data collected through detailed case studies of existing electronic recordkeeping systems (primarily databases and electronic document management systems) in bureaucratic settings in North America, Europe and China. 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. The Benchmark Requirements are, therefore, addressed to those who are responsible for the creation and management of active electronic records, whereas the Baseline Requirements are addressed to those, often (but not always) the archivists, who are responsible for the preservation of electronic records of long-term value. In traditional archival practice, these two aspects (i.e., active recordkeeping and archival preservation) have been viewed as distinct and separate activities undertaken by different agents (known as the life cycle approach). An alternative view, however, is presented by continuum theory, whereby recordkeeping activities are not necessarily attached to specific phases of a record's existence, but are integrated throughout its life span.[2,3] The implication for the development of the analytical framework and also for the metadata schema registry of the existence of both of these worldviews is that it is very important to identify the agents, processes, and points in the existence of a record that are relevant for the creation, management preservation, use and elimination of metadata in different contexts.[4]

Other sources used in the development of the analytical framework were the ISO 15489 Information and Documentation -- Records Management Standard (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).

A key issue encountered by the researchers in developing the analytical framework was that the InterPARES1 requirements were expressed as a set of narrative, conceptual requirements, rather than as production rules (see

Table 1 for examples). A process of operationalising the requirements in terms of what they implied for metadata elements, values, identification of responsible agents and related metadata creation and management procedures needed to be undertaken. It was also necessary to make decisions as to how to reconcile any inconsistencies that existed, either internally or between these and some of the other requirements expressed in the other standards used.

Table 1. Examples of Benchmark (A) and Baseline Requirements (B) (4).

To support a presumption of authenticity the preserver must obtain evidence that:	
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*
The preserver should be able to demonstrate that:	
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

This latter point raises an important concern for the developers of the framework and the registry, namely transparency. The framework and registry, while they are being developed in part to be made available as tools to be used by the professional communities involved in electronic recordkeeping and digital archiving, they are, first and foremost, research instruments. As the previous discussion has indicated, researchers have had to make judgment calls about how to translate conceptual requirements into operational ones, and, in the case of the registry, to decide which elements and combinations thereof in particular schemas and element sets, might meet those requirements and to what extent. A hallmark of metadata registry development is the notion of a trusted resource. InterPARES2 Description Group researchers, therefore, have been careful to include in the design of their tools, components that disclose to other users the basis upon which particular decisions were made in assessing and codifying both conceptual requirements and the metadata resources being analysed.

Developing the Metadata Schema Registry

The metadata schema registry has been developed through a staged iterative process, with the back and front ends being developed separately. This decision allowed for a pilot analysis of selected schemas to identify, name and refine registry elements, attributes, values and capabilities and the relationships between them, in compliance with the ISO/IEC 1179 Information Technology – Metadata Registry (MDR) standard's guidelines regarding the naming of registry elements and attributes, that could then be built into the XML schema being designed for the registry. It also allowed for schema analysis to move ahead while the registry was still being developed and tested. The front end of the registry has been the last to be developed and tested and is still in prototype stage.

The metadata registry schema prototype currently includes approximately 120 fields organized hierarchically. The first level of the hierarchy comprises eleven elements: Registration, Identification, Accessibility, Rights, Provenance, Description, Analysis, Documentation, Relationships, Administration, and a general Note element. These elements are further broken down into sub- and sub-sub-elements. The Analysis element of the metadata registry schema currently includes 15 sub-elements, but will eventually contain the complete analysis tool, comprising approximately 40 questions and associated sub-questions or comments.

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

Conclusion

The communities involved in digital imaging, digital archiving (as it relates to the ongoing management of library materials and scientific data), and electronic recordkeeping often function in very distinct environments, developing different approaches based upon different needs and contingencies. Metadata, however, is an aspect that is integral to the activities in each area. Moreover, in an era that is increasingly concerned with notions of reliability, authenticity and other hallmarks of trustworthiness, the ongoing creation, management, preservation and use of trusted metadata is a concern that is surfacing across all communities. While the work of InterPARES2 is primarily directed toward the preservation of authentic electronic *records*, it is relevant to all of these related communities for several reasons. For one thing, records and other types of digital materials often do not exist in mutually exclusive management environments and are often subject to the application of common metadata practices. Since the demonstration of the continued authenticity of electronic records is generally a *sine qua non,* requirements established by archivists and implemented through tools such as those discussed in this paper set the bar at its highest level for ensuring that archived digital materials cannot be challenged on the basis of their trustworthiness, and might assist any community in measuring up and developing its own resources and practices. Finally, until recently, the archival community has lacked in software tools developed to address its specific needs and has been reticent about engaging in its own development activities. The metadata schema registry and analytical framework, together with any specifications for metadata management and manipulation tools that emerge from this research, represent one major step toward addressing this lack.

References

1. Create Once., Use Many Times: The Clever Use of Metadata , http://www.sims.monash.edu.au/research/rcrg/research/crm/index.html
2. Frank Upward, Archives and Manuscripts 24, 2 (1996).
3. Frank Upward, Archives and Manuscripts 25, 1 (1997).
4. Anne Gilliland-Swetland et al., Archival Science (submitted).
5. InterPARES 1 Project, The Long-term Preservation of Authentic Electronic Records: Findings of the InterPARES Project, 2003. Available: http://www.interpares.org/book/index.cfm

Acknowledgements

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. Financial support for the project has also been provided as part of the Create Once, Use Many Times - The Clever Use of Metadata in eGovernment and eBusiness Recordkeeping Processes in Networked Environments Australian Resource Council (ARC) Linkage Grant in conjunction with Monash University, National Archives of Australia, State Records Authority of New South Wales and the Australian Society of Archivist's Committee on Descriptive Standards. The authors also wish to acknowledge Joanne Evans of Monash University, Lori Lindberg and Nadav Rouche of the University of California, Los Angeles, Hans Hofman of the National Archives of the Netherlands, and Dr. Richard Marciano of the San Diego Supercomputer Center who have been instrumental in the development of the InterPARES 2 Metadata Schema Registry.

Biographies

Anne Gilliland-Swetland is Associate Professor and Director of the Center for Information as Evidence and the Archival Studies specialization in the Department of Information Studies at the University of California, Los Angeles. Her current research includes the InterPARES2, Museums and the Online Archives of California Evaluation (MOACII), and Clever Recordkeeping Metadata Projects. She has an MA (Trinity College Dublin) and PhD (University of Michigan), and is a Fellow of the Society of American Archivists.

Sue McKemmish is Professor of Archival Systems and directs the postgraduate teaching program in records and archives in the School of Information Management and Systems at Monash University. Her current research includes the Breast Cancer Knowledge Online, Clever Recordkeeping Metadata, and InterPARES2 Projects. She has published extensively on the role of recordkeeping in society, records continuum theory and practice, recordkeeping metadata, and archival systems. She has a PhD (Monash), and is a Laureate of the Australian Society of Archivists.