Describing and analyzing the recordkeeping capabilities of metadata sets

Joanne Evans
School of Information Management and Systems
Monash University, Melbourne, Australia
Tel. 61-3-9903 2467
Mail: jeeva2@student.monash.edu.au

Lori Lindberg
Graduate School of Education and Information Studies
Department of Information Studies
University of California, Los Angeles
Tel. 415-902-8919
Mail msarchiv@ucla.edu

Abstract: This paper describes initiatives in the archives and records management communities to address the metadata required to ensure the creation and preservation of authentic, accurate and reliable records across systems and through time. In particular, it discusses the development by the InterPARES 2 project of a metadata schema registry to describe, analyze, and identify potential extensions of key metadata standards and sets, including Dublin Core, in terms of addressing recordkeeping requirements. **Keywords**: recordkeeping metadata, metadata schema registry, metadata standardization, Dublin Core.

1   Introduction

Recordkeeping – the process of 'making and maintaining complete, accurate and reliable evidence of business transactions in the form of recorded information'[1] – enables individuals, businesses and governments to be accountable for their actions. A key to ensuring the authenticity, integrity and reliability of records as evidence of business and social activities is the capture and efficient management of metadata relating to the content, structure and context of records creation and use. The maintenance and management of this recordkeeping metadata has become an issue of critical importance for archivists and records managers of late, particularly concerning electronic records. This renewed attention to the roles recordkeeping metadata plays in the management of records across systems and through time and to facilitate resource discovery is the driver behind a number of initiatives including the development of an ISO standard, *ISO 23081: Metadata for Records and Records Management Processes.*

While on the one hand digital technologies pose a threat to the creation and management of records, on the other hand they open up an exciting array of possibilities relating to the ways in which recordkeepers may be able to capture and utilize metadata. Freed from the constraints of the paper world, the digital environment holds out the potential to source metadata automatically and to re-purpose it in a multitude of ways that may not only enable archivists and records managers to be more effective, but may also contribute to improved operational efficiency and accountability enterprise-wide.[2] Before that potential can be realized, however, it is necessary to come to some understanding of:-

a)   What is the recordkeeping metadata that needs to be captured?

b)   Is it already being captured in existing metadata sets and standards, and if so, where and how?

2   Recordkeeping metadata

Recordkeeping metadata has been defined to 'include 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] A number of initiatives have been undertaken in order to understand its nature and purpose.[4]

In 1998, the Records Continuum Research Group at Monash University undertook a project (known as the SPIRT Recordkeeping Metadata Project) to 'comprehensively specify and codify recordkeeping metadata'[5] The project used modeling techniques to develop

modeling techniques to develop conceptual models 'of records in their business and socio-legal context'. These models identified recordkeeping entities and relationships amongst entities for which metadata should be captured. Iterative conceptual mapping of best practice recordkeeping and other related metadata sets, along with literary warrant analysis of recordkeeping metadata requirements, helped to elucidate the element requirements for these entities and relationships. The major deliverable of the project was the Australian Recordkeeping Metadata Schema[6] that has since been used as the framework for the development of a number of recordkeeping metadata standards in particular jurisdictions[7]. The SPIRT conceptual models are also being utilized in the development of *ISO 23081 Metadata for Records and Records Management Processes*. This standard aims to act as a guide to understanding and implementing the metadata requirements of *ISO 15489 International Standard on Records Management*, by laying out a framework and principles for creating, managing and using records management metadata, exploring implementation issues and evaluating existing metadata initiatives against *ISO 15489* requirements.[8] A derivation of the SPIRT models provides the framework for the types of metadata to be captured and managed in records systems in accordance with *ISO 15489*.

In its work between 1999 and 2001 the International Research on Permanent Authentic Records in Electronic Systems (InterPARES) Project investigated requirements for assessing and maintaining the authenticity of electronic records, and identified a set of benchmark requirements to support the presumption of the authenticity of electronic records and a set of baseline requirements for the production of authentic copies of electronic records. 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'.[9] 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 can be the presumption of the authenticity of the electronic records held within it.

records held within it.

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.

3 Recordkeeping metadata in existing schemas

While these initiatives have sought to identify the nature of recordkeeping metadata, the questions of whether, where, and how recordkeeping metadata is being captured in existing metadata sets is currently being addressed by the Description Cross Domain of the InterPARES 2 Project. Building on findings from the initial InterPARES project, InterPARES 2 is a collaborative international research initiative that is investigating issues of authenticity, reliability and accuracy of records created in digital environments resulting from artistic, scientific and government activities[10]. Within this project, the Description Cross Domain is investigating the role of metadata schemas and standards in records creation, control, maintenance, appraisal, preservation and use in both traditional and emerging digital and web-based environments within the three focus areas. In conjunction with this work, the research is exploring specifications for automated tools to be used in metadata creation and handling.[11] In order to undertake these investigations the research team is building a database to register and describe metadata schemas in a standardized, authoritative manner and to assess their recordkeeping and archival capabilities. While there are existing directories of metadata schemas available on the Web, none fully meet the needs of this research. In particular, InterPARES 2 is interested in what descriptive data is needed to manage metadata schemas through time and across domains. A version of a metadata schema is a record of a metadata structure at a particular time. Hence in order for information objects to be understandable through time, metadata about their metadata, i.e. the structure and semantics of elements, must also be maintained through time. Capture and maintenance of this 'meta' metadata is essential, particularly to support

particularly to support digital preservation and metadata re-use.[12]

A key purpose of the registry is to identify and describe 'salient' features of metadata schemas in order to act as a single point of discovery. There is a vast array of initiatives underway to define metadata schemas for different information objects, from differing perspectives, and for differing purposes. The instantiation process has also revealed a multitude of ways of describing and presenting information about these metadata schemas, often with key data seemingly buried in ancillary documentation. In some cases, the specification documentation assumes an understanding of the purpose, scope and perspective of the schema, as well as the development context, and launches straight into the description of metadata elements. While this is acceptable if one has already made the decision to use a particular schema, it is not particularly helpful when trying to assess whether a particular schema exists to meet one's needs. Thus, the registry is seen to have a broader audience than just the InterPARES2 research team in extracting descriptive metadata about metadata schemas into a standard structure to aid in their discovery, understanding and use.

In forming the metadata registry and other descriptive initiatives, the team has developed a structure for the registry's records as 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 effected;
- 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;
- 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;
- Administration – data elements for the administration of the schema registry.[13]

Instantiations have been used to refine the descriptive requirements and the data structure of elements. A key outcome of the process has been the need for flexibility to capture the myriad ways of presenting and describing information about metadata schemas. For example, in the Provenance section the original intent was to capture developer, owner, and publisher information. When it came to populating the DTD with examples, however, it was quickly discovered that a more flexible structure was needed to cope with the variety of potential relationships between agents and schemas.

4   Analyzing recordkeeping capabilities

One of the key purposes of the database is to assess the recordkeeping and archival capabilities of extant metadata sets. This is an important issue for archivists and records managers. Initiatives like the *Australian Recordkeeping Metadata Schema* and *ISO 23081* assume that much of the metadata required for recordkeeping is available in the business systems in which records are created and kept. Indeed, the presence of such metadata can make these systems into recordkeeping systems even when they were not designed as such. At issue, then, is whether the metadata can be inextricably linked to the record to which it pertains and that the authenticity, accuracy and reliability of the metadata can be maintained, in order to maintain the authenticity, accuracy and reliability of the record itself.

Along with other metadata communities, the recordkeeping profession is aware that automatic capture of metadata is the only sustainable method of metadata creation. Automatically capturing metadata as part of the process to which it pertains is not only efficient, but leads to better quality metadata as its creation is an integral part of the process rather than a post-hoc costly add-on. Hence the need for the InterPARES Description Cross

for the InterPARES Description Cross Domain to assess existing metadata sets against recordkeeping requirements to identify what recordkeeping metadata is being captured and what is not. An understanding of this will enable archivists and records managers to strategically address the disparities, whether that be by engaging with other metadata communities as appropriate to address recordkeeping requirements within their schemas, or in our own recordkeeping metadata schemas providing the lead and/or the mechanisms to fill the gaps. The database thus has the potential to be utilized as a tool to discover the gaps that exist and to find other metadata sets that might help to fill those gaps.

Having established why it is desirable to analyze the recordkeeping capabilities of metadata sets, the next question is how can such a determination be made? The answer is to use the 'best practice' warrant of instruments like *ISO 23081* and the *InterPARES Benchmark and Baseline Requirements.* In conjunction with the registry database we have established an analysis process that provides an in-depth examination of metadata schemas and evaluates them as to their recordkeeping metadata capabilities. First, the recordkeeping entities that a particular schema can be used to describe are established, followed by the identification of elements that meet the metadata needs expressed in *ISO 23081* and the *InterPARES Benchmark and Baseline Requirements.* As part of the process it is important to assess the degree to which a schema meets a requirement. When analyzing a complex schema with a multi-entity and relationship framework, such as the Australian Recordkeeping Metadata Schema, certain metadata requirements can be represented to varying degrees either as separate linked entities or by 'in-place' elements[14] Thus, there is a need to assess whether a particular metadata schema minimally or more comprehensively addresses a particular requirement (such as a schema which simply identifies recordkeeping entities versus a schema that identifies, describes and allows for relationships among recordkeeping entities).

5 An example of analysis: assessing the Dublin Core

The analysis process examines a metadata schema and evaluates its recordkeeping capabilities by mapping the schema against recordkeeping and archival requirements as expressed in instruments like *ISO 23081* and the *InterPARES Benchmark and Baseline*

InterPARES Benchmark and Baseline requirements. Documentation for the schema 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. For Dublin Core, this work was relatively easy and straightforward, as the Core is a simple flat schema with a small number of elements, all optional and repeatable.

The primary tool of the analysis process is an analysis worksheet, organized to systematically analyze schemas within 7 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
6. Recordkeeping – Classification of Purpose of Recordkeeping Metadata
7. General Comments

Post-analysis, the worksheet is saved and linked to the Documentation category of the registry record for the schema. Analysis results are compiled into a summary document that is also linked to the metadata schema registry record. It is hoped that this summary document will serve as a basis for structured presentation of recordkeeping capabilities of particular schemas within the database itself.

The analysis of Dublin Core reveals significant limitations as to its applicability for recordkeeping metadata. Designed primarily for resource discovery on the Web, the metadata elements provided in Dublin Core do not meet the requirements for managing a resource as a record through time. Its simplicity, as well as the obvious fact that it was never designed for such functions, naturally limits its use in this way. Recordkeeping metadata functions in a manner much more complex than simply helping users discover records, thus there are necessary elements of recordkeeping metadata related to resource management that are not included in Dublin Core. Elements that package such information as a record's function, its level of aggregation, its location, disposition and

mandate, as well as various histories of use, management, and preservation are essential for the management of records in active use as well as those placed in archival custody.

The extensibility and flexibility of the Core, however, means it can be extended to enable recordkeeping metadata creation and capture. This has already been accomplished successfully through such metadata standards developments as the Australian AGLS Metadata Standard[15] and its subsequent influence on the design of the Recordkeeping Metadata Standard for Commonwealth Agencies,[16] as well as U.S. recordkeeping metadata standards such as the Minnesota Recordkeeping Metadata Standard.[17]

6   Current Status

The registry is in its very early stages of development with an exploratory prototype created to allow for instantiations. These examples can then be used to refine the specification and develop the understandings of the research team, in terms of technological issues and the realities of the domain under investigation. As already noted, instantiations have revealed great diversity in the presentation of information about metadata sets. Of most concern is the lack of infrastructure for global persistent identifiers and the variable nature of the metadata associated with their web publication. One would assume that publishers of metadata sets have an appreciation of the value of metadata for information discovery and would be using such schema as Dublin Core in an exemplary manner, especially as the initial instantiations are being drawn from information management and recordkeeping communities. However it has been surprising to see in some cases the lack of Dublin Core metadata and in others the poor quality of the Dublin Core content. Wendy Duff notes in a recent article that studies are revealing 'little progress…in the use of metadata standards and that existing metadata tags are *ad hoc.*'[18] It will be interesting to see how this plays out when looking at metadata sets from the artistic and scientific domains. The development of the registry is proceeding using recordkeeping metadata standards. Assessing the recordkeeping capabilities requires a detailed understanding of both the metadata set and the instruments against which it is to be measured. The recordkeeping domain represents the field of expertise of the analysts in the research team and hence seems the logical point from which

the logical point from which to develop and validate the analysis process. In addition the analysis of recordkeeping metadata standards can set the benchmark for the measure of the degree to which a metadata set meets a particular requirement.

There was also an opportunity to test the usefulness of the analysis process in feeding into an Australian initiative to develop a recordkeeping metadata standard under the banner of the IT/21 Records Management Committee of Standards Australia. The analysis of the recordkeeping metadata standards developed by Australian archival institutions identified defects which can be addressed in the production of a national standard. The process also highlighted the need for such an initiative as it illustrated variations between the standards that could make metadata interchange amongst them problematic. Recordkeeping professionals must lead by example in fostering compatibility between metadata sets so that metadata can be re-positioned and re-purposed with ease.

7   Future Work

Ahead lies the task of translating the registry and the analysis process from a prototype into a production version. Initially to be utilized by researchers in the InterPARES project, one of the research aims is to release it to a wider audience in order to foster efforts for standardization and compatibility between metadata sets and standards. With appropriate search and retrieval interfaces, the database can act as a single point of discovery for metadata developers encouraging the sharing of structures, where appropriate, in order to reduce the current proliferation and improve compatibility amongst standards.

The current structure allows for the description and analysis of metadata standards but will need extension to deal with application profiles, where metadata is sourced from different schemas in a particular implementation. The pilot instantiations are also revealing a proliferation of encoding schemes, such as controlled vocabularies from which element values are sourced or standardized data representations for elements (e.g. *ISO 8601 Data elements and interchange formats - Information interchange - Representation of dates and times*). It is indicating the need to analyze relationships amongst these schemes as part of the aim of fostering greater data compatibility

8 Conclusion

By identifying, describing and analyzing the archival and recordkeeping capabilities of existing metadata standards and sets, the database can play an important role in helping us understand commonalities and diferences between them. From a recordkeeping perspective it can identify what recordkeeping metadata is being addressed in existing sets and what is not, and hence form the basis for strategic collaborations with other metadata communities in order to address issues of authenticity and integrity In order to harness the power of metadata in digital environments it is vital that we look at synergies and harmonization whilst maintaining the necessary diversity It is hoped that this database will become part of the framework to address such issues.

References

1. *Australian Standard AS 4390.1 1996 Records Management Part 1: General*, Standards Australia, New South Wales, 1996

2. Adrian Cunningham, 'Six Degrees of Separation: Australian Metadata Initiatives and their Relationships with International Standards', *Archival Science*, vol. 1, no. 3, 2001, pp. 271-283.

3. Sue McKemmish, Glenda Acland, Nigel Ward, and Barbara Reed, 'Describing Records in Context in the Continuum: the Australian Recordkeeping Metadata Schema', *Archivaria*, vol. 48, Fall 1999, p. 3-43.

4. John Roberts and Andrew Wilson, *International Recordkeeping / Records Management Metadata Initiatives: Report and Recommendations for DC Advisory Board*, DC-Gov Working Group, Dublin Core Metadata Initiative, 23 April 2002, http://dublincore.org/groups/government/DCMIResourceManagement.html

5. *Archival metadata standards for managing and accessing information resources in networked environments over time for government, social and cultural purposes;* CIs Sue McKemmish and Anne Pederson, PI Steve Stuckey; funded by an Australian Research Council SPIRT Grant and industry partners State Records NSW, National Archives of Australia, Queensland State Archives, Australian Council of Archives and the RMAA, http://www.sims.monash.edu.au/research/rcrg/research/spirt/about.html, 1998.

6. Sue McKemmish, Glenda Acland, Kate Cumming, Barbara Reed and Nigel Ward, Australian Recordkeeping Metadata Schema (RKMS), Version 1.0, 31 May 2000, http://www.sims.monash.edu.au/research/rcrg/research/spirt/deliverables.html

7. Standards derived from the Australian RKMS include: National Archives of Australia, *Recordkeeping Metadata Standard for Commonwealth Agencies*, http://www.naa.gov.au/recordkeeping/control/rkms/contents.html, 1999; State Archives Authority of New South Wales, *New South Wales Recordkeeping Metadata Standard*, http://www.records.nsw.gov.au/publicsector/erk/metadata/rkmetadata.htm, 2001;

r/erk/metadata/rkmetadata.htm, 2001; and State Records of South Australia, *South Australian Recordkeeping Metadata Standard*, http://www.archives.sa.gov.au/publications/standards.htm, 2003.

8. *ISO/TS 23081-1:2004 Information and documentation - Records management processes - Metadata for records -- Part 1: Principles*, International Standard Organization, 2004.

9. 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

10. See the InterPARES 2 Project website at http://www.interpares.org/ip2/ip2_index.cfm.

11. See the Description Cross Domain page at http://www.interpares.org/ip2/ip2_description.cfm.

12. Michael Day, 'Integrating Metadata Schema Registries with Digital Preservation Systems to Support Interoperability: a Proposal', in *DC 2003: Supporting Communities of Discourse and Practice - Metadata Research & Applications, Seattle, Washington (USA), September 28-October 2*, 2003, http://www.siderean.com/dc2003/101_paper38.pdf.

13. The initial requirements for the metadata elements of the registry were developed by Anne Gilliland-Swetland, Hans Hofman, Sue McKemmish and Bill Underwood, as an expert focus group within the Description Cross Domain Research Team of InterPARES2. These requirements were iteratively refined to form the basis for the design and development of the initial prototype by Nadav Rouche and Joanne Evans.

14. See 'Redundancy and Robustness of Descriptions' within Records Continuum Research Group, 'Technical Introduction' in *SPIRT Recordkeeping Metadata Standards Project*, http://www.sims.monash.edu.au/research/rcrg/research/spirt/deliver/techintro.html

rg/research/spirt/deliver/techintro.html 1998.

15. 'The AGLS Metadata Standard is a set of 19 descriptive elements which government departments and agencies can use to improve the visibility and accessibility of their services and information over the Internet.' See http://www.naa.gov.au/recordkeeping/gov_online/agls/summary.html

16. See Note 7 for access to this standard.

17. The Minnesota Recordkeeping Metadata Standard comprises 20 elements and a number of sub elements. It is mapped to Dublin Core as part of its documentation. See http://www.mnhs.org/preserve/records/metamrms.html

18. Wendy M. Duff, 'The Dublin Core and Its Limitations' *DigiCULT.Info*, December 2003. http://www.digicult.info/downloads/dc_info_issue6_december_20031.pdf