Recordkeeping Metadata. Presenting the Results of a Working Meeting

MARGARET HEDSTROM
School of Information, University of Michigan, 550 East University Avenue, 481009-1092 Ann Arbor, MI, USA. E-mail: hedstrom@umich.edu

Goals and Objectives

The main purpose of this report is to present a personal summary of the Recordkeeping Metadata Working Meeting sponsored by the Archiefschool, Nederlands Instituut voor Archiefonderwijs en Onderzoek (Netherlands Institute for Archival Education and Research) and held in The Netherlands on June 5, 6 and 7, 2000.[1] David Wallace has prepared detailed proceedings of the Working Meeting that are published in this issue of Archival Science, together with most of the papers presented at a public seminar Understanding and Preserving Reliable and Authentic Recorded Information in a Digital World – Focus on Metadata which was organized by the Archiefschool after the Working Meeting, on 8 June 2000.

The purpose of the Working Meeting on Recordkeeping Metadata was to bring together a diverse group of archivists, metadata experts, and computer scientists to discuss the current status of metadata projects and initiatives that are oriented toward recordkeeping and long-term preservation of archives. One goal of the Working Meeting was to propose research, standards initiatives, and other projects that might link the recordkeeping community and its interests more closely to other metadata initiatives. The specific goals of the Working Meeting were:

- Position recordkeeping initiatives in context of needs of other disciplines
- Identify what the recordkeeping community shares with others
- Assess existing recordkeeping metadata schema

[1] This report is a revised version of a presentation on the Recordkeeping Metadata Forum, presented at *Understanding and Preserving Reliable and Authentic Recorded Information in a Digital World – Focus on Metadata* on 8 June 2000 in The Hague, organized by the Archiefschool.

- Place metadata in context of business processes and culture
- Develop a common infrastructure for research

Recordkeeping Metadata: Definitions and Relationships to Other Types of Metadata

One accomplishment of the Working Meeting was developing a consensus around the definition of recordkeeping metadata:

> Recordkeeping metadata is "*Structured or semi-structured information which enables the creation, management, and use of records **through time** and **within** and **across domains** in which they are created.*"

Structured metadata provides a fixed schema for organizing metadata. The Dublin Core metadata set is one example of structured metadata. Semi-structured metadata does not require a fixed structure. It is an extensible set of tags that can accommodate unique configurations of metadata. An XML document without a mandatory document type definition (DTD) is one example of semi-structured metadata. Recordkeeping metadata may or may not follow a fixed schema. There was considerable agreement among the participants that it was unlikely that a single fixed schema could be developed or would be accepted to serve all recordkeeping environments.

This definition may leave as many questions unanswered as it addresses, but it is a starting point for identifying what may be unique about recordkeeping metadata compared with other communities that also use metadata for various information and data management purposes. In the process of developing this definition, participants discussed, but did not necessarily resolve, a number of related issues. Metadata is a highly level concept that means "data about data," but this concept has little concrete value or use until defined more specifically by communities that have particular applications for metadata. One of the challenges for the recordkeeping community is to define the particular metadata requirements for recordkeeping and to identify how these differ from what other communities need.

At one point in the Working Meeting, the participants broke into two groups to discuss differences between the records continuum approach and the records life cycle approach. This led to a fruitful discussion about the unique purposes for recordkeeping metadata and the need to exploit aspects of both records continuum and life cycle concepts rather than viewing them as incompatible alternatives. Two important functions of recordkeeping metadata are to support the transfer of records across domains and over time. By across domains, we mean transferring meaningful records from the

domain in which they are created originally to other the business, juridical, social, cultural and/or technological domains. By transferring records across time, we mean preserving records over time in such a way that their integrity, authenticity, and original meaning is maintained.

The recordkeeping community is not unique in either of these requirements. There are many cases where records (and other types of information) are transferred from their original environment to a different domain, and there are many examples of the need to preserve information over time. For example, regulated industries often transfer records to oversight bodies in order to comply with regulatory reporting requirements, while libraries preserve much of their content over time. It is fruitful to look at these two processes as independent, although many records are transferred across domains and over time. Some creators need to retain records for long periods of time for their own use and thus are more concerned with the time dimension than the issue of use by a different community. It is important to keep in mind that records may be transferred across many domains and that what constitutes a different time domain is not well defined.

What may make the recordkeeping community unique is the intersection of interest in cross-domain and over time transfers of records. The records continuum model is especially useful because of its sensitivity to cross domain transfers of records and the life cycle model is valuable for its notions of transfer of responsibility that coincide with points on the records life cycle.

This resulted in further refinement of domain and time issues. The group agreed that it might be useful to consider how different domains may be more or less amenable to adopting recordkeeping metadata due to their own needs to transfer records across domains or to preserve records over time. Some domains may be "controllable" because there are sufficient laws, incentives or sanctions (sufficient warrant) for recordkeeping authorities or archives to demand compliance with recordkeeping metadata requirements. Others may be potentially controllable because there is sufficient self-interest on the part of records creators to adopt recordkeeping metadata voluntarily. Other domains may be uncontrollable and the recordkeeping community will have to develop alternative strategies for maintaining reliable and authentic records created in those domains. Possible examples of the uncontrollable domain include personal records, situations where the requirements for recordkeeping and accountability are not clear-cut, and emerging organizational forms that operate without a formal organizational hierarchy and few formal rules.

One "time dimension" issue is the question of including recordkeeping metadata during the design phase of systems versus discovering, structuring or adding metadata after records have been created and accumulated. Partici-

pants developed the concepts of "prospective" and "forensic metadata" to distinguish between metadata that might be created as part of recordkeeping processes and metadata that might be restructured or added after the fact. The concepts of the controllable vs. the uncontrolled domains and prospective vs. forensic metadata are not intended to be sharp dichotomies. In the real world, organizations exist along a continuum from the most controllable to the uncontrolled. Likewise, few organizations are likely to anticipate all metadata requirements during the design phase and incorporate them all into the recordkeeping systems. One goal might be to move organizations closer to the controlled end of the continuum and to continue to encourage prospective metadata, recognizing that not all records will be created under these conditions. The recordkeeping community will need different sets of tools to operate effectively in different domains and new tools for capturing and creating forensic metadata (extraction, summarization, data mining, automatic clustering, and emulation).

Related Communities

The participants identified several communities that have complementary or overlapping metadata requirements. There was particular attention to the following communities:
- Libraries
- Museums
- Digital libraries
- E-commerce, E-business, E-government
- Information Intensive Industries
  - Entertainment
  - Regulated Industries
  - Software
- Intellectual Property Rights, etc.,

This is not exclusive list, but it provides examples of domains where there is a core mission or a business interest in maintaining authentic and reliable records. Some of these communities have developed metadata schema that are potentially relevant to recordkeeping for such purposes as resource discovery, authentication, version control, and rights management.

Related Metadata Schema and Standards

Workshop participants discussed and contributed many examples of related metadata developments that are relevant to the recordkeeping community.

We agreed on a rough taxonomy that distinguishes between format-based metadata and function-based metadata. One example of format-based metadata is MPEG-7, but there is much work underway on metadata for other specific types of digital objects. Function-based metadata are designed to facilitate generic functions such as management of intellectual property rights, record keeping, resource discovery, or e-commerce. Dublin Core is an example of metadata designed originally for resource discovery that has not been easily extended to perform other functions. The participants also examined and compared some of the metadata models designed explicitly to support recordkeeping including the Australian SPIRT Model, the draft ISO international records management standard, and some of the templates being developed by the InterPARES project. Several common themes emerged in these discussions. There was considerable concern with defining baseline or minimum requirements for recordkeeping metadata. Metadata are expensive to create, capture and manage and organizations are unlikely to adopt metadata frameworks unless direct benefits can be demonstrated. Another common theme was the need to identify which aspects of existing metadata standards could support recordkeeping and where existing metadata frameworks might be adapted to recordkeeping with minor modifications. There was also an initial effort to identify metadata issues that are unique to recordkeeping and that are not being addressed at all.

"White Spots"

The participants agreed to limit the use of jargon in order to avoid confusion among the non-native English speakers and among the various flavors of English represented among participants. The group nevertheless developed its own new term "white spots" to refer to issues that the community needs to be aware of, but that were not addressed in depth by the participants. The following issues should be incorporated into evaluation of existing metadata projects and/or addressed through research:

- *Scale and scalability*: Given the large and rapidly growing quantity of digital objects as well as the large backlog of paper and other analog records, recordkeeping metadata schema have to scale to accommodate very large quantities of records.
- *Granularity*: Archivists are unclear about the level of granularity necessary for recordkeeping metadata and at which level to apply various metadata schema. There are metadata element sets that apply to individual documents, to aggregations of documents, to part of systems, and to entire systems. There is a need for additional thinking about which metadata elements apply to which objects and which metadata elements

can be inherited from systems, aggregations of documents, or from the recordkeeping context. There is also a need to define a minimal acceptable level of granularity given the issue of scale and scalability discussed above.

- *Persistent linking*: Persistent linking refers to the need to maintain over time the links between metadata and the data that it describes or references. This is a common problem in many communities. It is important for the recordkeeping community to keep in mind that creating adequate metadata is not sufficient if the relationships between metadata and the documents it describes are not also maintained. There are numerous proposals for persistent linking that should be beneficial to the recordkeeping community to address this problem.

- *Relationships to other communities*: Not surprisingly, the participants spent considerable time discussing the many types of relationships between the recordkeeping community and many other communities. Among the most salient relationships are relationships between the recordkeeping community and records creators, relationships with other communities that are developing and using metadata for related purposes, and relationships with standards bodies. One suggestion for identifying and cultivating relationships with the most relevant representatives of these communities is to distinguish between problems that only the recordkeeping community is likely to solve, problems that others may solve in ways that are beneficial to recordkeeping, and problems where the recordkeeping perspective contributes solutions to larger problems. There was a general consensus that the time dimension and the interest in moving records across domains are primary considerations for the recordkeeping community.

- *Tools to create, capture, and manage metadata automatically*: There is a need to utilize tools for automatic extraction, automatic classification, and data mining as an integral part of any metadata implementation whether prospective or forensic. The recordkeeping community may need to modify existing tools or develop new ones to limit the amount of human intervention necessary for metadata capture and management.

- *Users*: Users have been missing from most metadata initiatives. There are two particularly important reasons for including users in the development of recordkeeping metadata. Studies from the field of computer-supported cooperative work (CSCW) consistently show that users do not comply with system requirements unless they see the benefits of doing so. For example, if system designers require users to add metadata to documents, the recordkeeping community will benefit from understanding what benefits users see in these approaches and how new

requirements alter existing work routines. Second, part of the justification for recordkeeping metadata is its value or added value for future users of records. Assumptions about users' needs and preferences rarely are based on actual data about user needs or preferences.

- *Unstructured environments*: Assumptions about recordkeeping in formal organizations that operate in environments with clear mandates and accountability requirements may not be valid in less structured environments. The recordkeeping community needs to test its proposals in a variety of organizational settings. Strategies that are effective in formal, structured environments may not work well for less structured organizations or for personal records.

- *Maintaining metadata through time*: Most attention to date has been focused on developing metadata strategies and encouraging records creators and records managers to adopt them. If recordkeeping metadata becomes one of the foundations for creating and keeping authentic and reliable records, then maintaining metadata will also be an important recordkeeping function. In addition to the problem of persistent linking, metadata strategies should include version control and must consider metadata maintenance if records are migrated to new systems.

The Way Forward

Research issues

The final discussions at the Working Meeting focused on research and on other measures to move the recordkeeping metadata agenda forward. Although the participants did not develop a formal research agenda, several important areas for research were identified.

- *Research on the social, cultural, policy domain of creation*: There is considerable need for research on the incentives, motivations, and other drivers that might encourage or discourage records creators from meeting recordkeeping metadata requirements. Such research needs to address different social, cultural, business and policy domains because drivers will vary across domains. There is a particular need for research on recordkeeping in unstructured and seemingly uncontrollable domains.

- *Research in the domain of use*: The participants proposed several areas for research around the problems of knowledge transfer across domains. What may be explicit knowledge in one environment can become invisible when records are transferred and used in a different domain.

Fruitful areas for research include developing definitions of useful interfaces between domains, developing methods to use metadata to capture and maintain contextual information about records and records creators, and finding ways to make knowledge that is tacit in one domain explicit in a different domain.

- *Technical issues*: Two technical issues emerged as important research areas that may be unique to the recordkeeping community. There is a need to investigate, refine or develop technical methods for post-hoc metadata discovery in order to implement that concept of forensic metadata. A second technical area is research into methods for maintaining metadata over time.

Next steps

The participants proposed several actions to expand and continue work on recordkeeping metadata.

- *Formation of the Archiving Metadata Forum*. There was considerable interest in using the Working Meeting as the basis for forming a broader Archiving Metadata Forum. Such a Forum would need institutional support for hosting a listserve or other communication tools and a web site. The Forum could also publicize its presence by writing a summary of the proceedings for journals or newsletters that are widely read by people who are conducting research or developing metadata schema and standards. *D-Lib Magazine* was suggested as one venue for such an article.

- *Web Site at the Archief School*. The Archief School in the Hague agreed to set up a web site for the Forum. That site is now available at http://www.archiefschool.nl/amf/

- *Influence ISO Record Keeping Metadata Standard*. The ISO is in the process of reviewing a proposed standard for Record Keeping Metadata. Several of the participants are in involved in the development and review of this standard. There are opportunities to influence the content of the standard and to support its adoption and implementation.

- *White Paper or Draft Research Agenda*. There was considerable discussion of research issues but insufficient time in the Working Meeting to develop a formal research agenda or white paper on research issues. Many of the participants are involved in research or in the development of metadata schema. A common research agenda or framework would facilitate more collaboration among participants and help to distinguish between relevant research underway outside the record keeping community and research questions that are not being addressed. This

could be followed by the development of research proposals and projects to fill gaps in current research.

- *Meeting Next Year*. The participants supported the idea of organizing another meeting in 2001.