*Pekka HENTTONEN**

Creating Recordkeeping Metadata

*Pekka Henttonen
Assistant Professor, Doctor of Social Sciences
Department of Information Studies and Interactive Media
University of Tampere, Finland
FI-33014 University of Tampere, Finland
Kanslerinrinne 1, Pinni A 4055, University of Tampere, Tampere, Finland
pekka.henttonen@uta.fi

HENTTONEN, Pekka, Creating Recordkeeping Metadata. Atlanti, Vol. 19, Trieste 2009, pp. 67-76.

Original in English, abstract in English, Italian and Slovenian, summary in English

Records management metadata is an essential tool in managing and keeping authentic, reliable, and retrievable electronic records. The paper discusses the ways to create metadata for records and the findings of the Finnish FinnONTO 2.0 Semantic Web research project in which metadata created in electronic records management systems records has been analyzed. The majority of used metadata elements was about event history. A large portion of SÄHKE metadata scheme remained unused.

HENTTONEN, Pekka, Creare metadati per la conservazione dei documenti. Atlanti, Vol. 19, Trieste 2009, pp. 67-76.

I metadati per la gestione documentale sono uno strumento essenziale nella gestione e conservazione di documenti elettronici autentici, affidabili e rintracciabili. L'articolo discute delle modalità di creazione dei metadati per i documenti e per il progetto di ricerca finlandese FinnONTO 2.0 Semantic Web, nel

Introduction

Recordkeeping metadata is here defined as "structured or semi-structured information, which enables the creation, management, and use of records through time and within and across domains". This is the ISO definition for metadata for managing records. Recordkeeping metadata contains not only information about records, but also about business and records management processes in which records have been created and used, agents involved in processes, and mandates governing processes and agents. Recordkeeping metadata stores information about the history of records, but it also governs how records are managed and used, both at the present time and in the future: for instance, metadata defines access limitations to records as well as the date when the limitations are to be abolished. (International Organization for Standardization, 2007.)

For records professionals (records managers and archivists) recordkeeping metadata is a solution for many problems. Metadata makes it possible to create and keep authentic, reliable, retrievable electronic records that are understandable even outside the original environment. However, creating this rich web of information represents a problem. It is not easy to produce all the metadata that is required in recordkeeping. Manual attribution of metadata makes electronic records management systems (ERMS) cumbersome to use. Appropriate contextualization of records may be hard to achieve without a user input, but it has been claimed that users are reluctant to add metadata to records (Bailey, 2008).

In this paper I examine metadata that is currently created in ERM systems. I also discuss whether metadata creation could happen automatically, or at least, how creating metadata could be made easier for a user. Especially, I discuss the results of our still on-going research in which we have studied characteristics of metadata in records of a Finnish government agency.

Problem of classification

Metadata attribution should be as automatic as is possible to achieve. ISO standard recommends that manual attribution of metadata should as far as possible be done using predefined selection lists (and not open fields which can be populated at will). It also says that sources for automatic attribution of metadata include (International Organization for Standardization, 2007):

- system clocks for data/time,
- network log on or authentication systems for details of individuals and their work units,
- human resource management systems for details of individuals and their work units,
- workflow systems for work process details, business flows, movement or authorizations,
- email systems for receipt/dispatch and transmission details, and
- mapping matadata from the "file properties" of the creating application, or parts of the operating system.

Although this makes it possible to automatically capture some details of record environment, it is not an answer to all the problems in record creation. Especially, the problem is how to define access restrictions, enable systematic disposal of records and link a record to organizational functions and processes. There may be many ways to achieve this, but most (if not all) of them are probably based on record classification in some way: when a record is linked to a classification scheme or put into a "folder" in the classification scheme, the record inherits some of its metadata values from the classification scheme or the folder. Also Finnish electronic records management systems are based on this idea: what type of records are created in organization and in what organizational function is planned in advance. The metadata values governing and access and life cycle depend on both function and record type: when we know (for instance) that the record is an "job application" created in "human resources management" function, we know also how long it should be retained and who has access right to the record.

Automatic classification

However, this still leaves us the problem of classification: how to find out what is the right functional class and record type?[1] In library and information science automatic classification has been a challenging research issue for several decades. Major motivation for this has been the high cost of manual classification (Golub, 2006.) Easier metadata creation would reduce (usually hidden) costs of manual classification of records and to record types and functions also make buraucratic organization more efficent.

Golub (2006) finds four appoaches in automatic subject classification (he regards "classification" synonymous with "categorization" and "clustering" in the broadest meaning of the terms). The

quale i metadati creati nei sistemi di gestione elettronici sono stati analizzati. La maggioranza degli elementi dei metadati utilizzati riguardava la storia. Una larga parte dello schema di metadati SÄHKE è rimasta inutilizzata.

HENTTONEN, Pekka, Ustvarjanje in hramba računalniških mega podatkov. Atlanti, Zv. 19, Trst 2009, str. 67-76.

Upravljanje z računalniškimi mega podatki je zelo pomembno, saj se morajo ohraniti avtentični, zanesljivi in nadomestljivi podatki. V pričujočem prispevku razpravljam o poteh, kako to zagotoviti. Eden od projektov je finska raziskava, ki se imenuje FinnONTO 2.0 Semantic Web, kjer so preverjali in analizirali mega podatke na računalniško zapisanih dokumentih. Večina uporabljenih elementov mega podatkov je nastalo iz zgodovinskih dogodkov, ki so jih vnašali v računalnik, kjer pa je bil izkoriščen samo majhen del SAHKE mega podatkov.

SUMMARY

Recordkeeping metadata enables us to create and keep authentic, reliable, retrievable electronic records that are understandable even outside the original environment. However, creating this rich web of information represents a problem. It is not easy to produce all the metadata that is required in recordkeeping. Manual attribution of metadata makes electronic records management systems (ERMS) cumbersome to use. Metadata attribution should be as automatic as is possible to achieve. Although the source of some metadata values can be the information system itself (for instance, system clock for a datetime), this not an answer to all the problems in record creation. Especially, the problem is how to define access restrictions, enable systematic disposal of records and link a record to organizational functions and processes. There may be many ways to achieve this, but most (if not all) of them are probably based on record classification in some way: when a record is linked to a

[1] In a registry system one has to also link a record to the right "case" of which the transaction creating or using the record is a part. For simplicity I have ignored this problem in the discussion.

classification scheme or put into a "folder" in the classification scheme, the record inherits some of its metadata values from the classification scheme or the folder. However, this still leaves us the problem of classification. In library and information science automatic classification has been a challenging research issue for several decades. There are several methods for automatic subject classification, but as far as I know, they have not been tried on functional classification. Records management is of the areas studied in the Semantic Web 2.0 (FinnONTO 2.0) project. The project has three parts. The first part has already been completed. In it we examined metadata in records of a Finnish government agency. In the other parts of the project we examine, among other things, how the functional classification is used, or could be used. In the completed study we analyzed metadata elements in records. We made statistical analysis to examine what kind of values (unique, non-unique) were given to the elements, how often the element was used, and how equally different values were distributed in records. We also grouped metadata elements to categories defined in the ISO 23081 and explored the usage of optional/mandatory elements to see what kind of metadata was actually created. The first finding was that a substantial number of metadata elements in SÄHKE remained unused. Almost 57 % of all sub-elements remained unused (52 % if we look only sub-elements in main elements with values). The second finding was about the content of the metadata in records. The absolute majority of metadata was about Event history (64.2 % of elements). The next came Description metadata (12.0 %), Use metadata (10.4 %) and Identity metadata (9.4 There was little metadata describing directly the record content. The third result was that optional metadata elements were rarely used. The reason for this remains unclear because it cannot be found by looking at the metadata alone. The last finding was that metadata showed clear patterns: generally metadata values were bipolar. Indirectly, this supports the claim that users are reluctant to add metadata

approaches are defined in the table below.

Approach	Description
Text categorization (supervised learning)	A machine-learning approach, in which also information retrieval methods are applied. It consists of three main parts: categorizing a number of documents to pre-defined categories, learning the characteristics of those documents, and categorizing new documents. In the machine-learning terminology, text categorization is known as supervised learning, since the process is "supervised" by learning categories' characteristics from manually categorized documents.
Document clustering	An information-retrieval approach. Unlike text categorization, it does not involve pre-defined categories or training documents and is thus called unsupervised. In this approach the clusters and, to a limited degree, relationships between clusters are derived automatically from the documents to be clustered, and the documents are subsequently assigned to those clusters.
Document classification	A library science approach. It involves an intellectually created controlled vocabulary (such as classification schemes), into classes of which documents are classified. The algorithm typically compares terms extracted from the text to be classified, to terms from the controlled vocabulary (string-to-string matching).
Mixed approach	Sometimes methods from text categorization or document clustering are used together with controlled vocabularies

Table 1. Approaches of automated subject classification (adapted from Golub, 2006)

Highly developed algorithms can be used to classify documents in an organization (e.g. Hou & Lin, 2006) However, as far as I know, there are no studies in which the methods of automated *subject* classification are applied automated *functional* classification. The difference is clear. The subject, what record "talks about", may have no direct relationship with the function in which the record is used or created. For instance, almost anything can be invoiced and, consequently, an invoice may contain terms from different areas of life. Sill every invoice belongs to the same functional category. Another example is a legislative process: a law about health care of conscripted personnel would be categorized under "health care" or "national defence" in subject classification, and not to "legislation", which would be the appropriate functional class. A third example: a love letter is produced as evidence in court - there probably are no terms in the document which might reveal its real functional context.

Hence, the difference between and subject and functional classification seems significant, but its practical implications for automated classification are unclear. Perhaps best result might be achieved by combining methods of automated subject classification with other approaches, like genre classification (about genre classification, see e.g. Kim & Ross, 2007). Genre classification might also provide tools for recognizing different record types.

FinnONTO 2.0 Research Project

Semantic Web 2.0 (FinnONTO 2.0) 2008-2010 is a national continuation project based on the results of the National Semantic Web Ontology Project in Finland (FinnONTO 2003-2007). The general goal of this large project is to combine benefits and synergy of Web 2.0 and semantic web technologies and demonstrate the results in various semantic web portals and applications.

The project uses ontologies to achive this goal. An ontology is "a *formal*, *explicit* specification of of a shared *conceptualization*". Here "conceptualization" refers to an abstract model of some phenomenon in the world which identidies the relevant concepts of that phenomenon. "Expilicit" means that the type of the concepts used and the constraints on their use are explicitly defined. "Formal" refers to the fact that that the ontology should be machine readable. (Fensel, 2001, p. 11).

The project has built national portals based on ontologies (for instance, see Finnish Culture and History portal, http://www.kulttuurisampo.fi/). For this purpose controlled vocabularies have been ontologized. Unfortunately, controlled vocabularies have not been generally utilized in Finnish records and archives management. The project focus has largely been on subject-based description of web documents and materials in libraries and museums. Libraries and museums were involved already in the first FInnONTO project, whereas management of records and archives is a new area.

The FinnONTO 2.0 project Is lead by the Semantic Computing Research Group (SeCo) of the Helsinki University of Technology[2]. However, the project has also independent sub-projects at the Department of Information Studies and Interactive Media in the University of Tampere. One of these projects has focused on records management. It has three parts:

1. Analyzing what metadata is actually created in recordkeeping (versus what elements are defined in the metadata scheme) and what patterns the metadata exhibits. This is needed to understand both the process of creation and services that might be built on the existing metadata.
2. Exploring how record creation and use is related to organizational units and their functions.
3. Studying how ontologies might facilitate use of common functional classification scheme of Finnish municipalities.

The third part of the project will be completed by the end of this year. The common classification scheme is fairly new and was

[2] For more information about the SeCo and its projects, see http://www.seco.tkk.fi

published only this year. Up to now every municipality has had a functional classification scheme of its own, although the functions are everywhere the same. In the project we examine current municipal classifications and try to build an ontology which allows a user to find the right class in the common classification scheme also with terms not employed in the common classification. In this we plan to use Search Ontology Editor for Concept-based Information Retrieval Interface (better known shortly as SHOE or SHOE4CIRI). The SHOE ontology editor has been developed at the department. Until now it has been used in cross-lingual information retrieval, for instance. (For the three level architecture behind the SHOE - concepts, linguistic and string level - see Järvelin & Kekäläinen, 2001; Suomela & Kekäläinen, 2006.)

Records management metadata in real life

The first and second part of our study are based on analysis of electronic records of a Finnish government agency. In the second part, which is not finished yet, we are studying how functions in the functional classification scheme are distributed to organizational units and how employees have used the ERMS in the light of metadata. For this every employee in the agency delivering the electronic records was mapped to a unit in the organization, and every function in the functional classification scheme to a unit or units. The questions (among others) are how predictable is the use of ERMS and how large portion of the classification is in use by different units and persons at different levels in the hierarchy. The answers may help to build more intelligent systems that assist the user by making enlightened guesses about what the user is likely to need. At the moment we are analyzing the results. They should be ready for publishing by the end of the year.

The first part of the project has been completed. The results will be published in an article (Kettunen & Henttonen, 2009) which is now in the review process. Hence, I give here only a summary of them and the methods used.

Like in the second part, in this part we also examined metadata in records of a Finnish government agency. Except for two records series - which were excluded because they contained sensitive or classified information - the set included all the electronic records which were received or created by the agency and captured in its ERMS in “cases” (see below) opened during the time period of 30.9.2005 - 31.12.2007. Altogether there were 7252 records of permanent or non-permanent value in 67 record series.

Metadata in the records studied complies with Finnish SÄHKE metadata specification, which is one of the current jurisdiction-specific specifications for ERMS. Finnish public authorities are required to use SÄHKE-compatible electronic records management systems if they want to keep records with permanent value solely in electronic format (without printing them to paper or microfilm) and later transfer them to the custody of the National Archives Service. (Henttonen, 2009).

Although SÄHKE is a national specification and not used outside the country, the metadata does not differ basically from records management metadata in general as defined in ISO 23081: there is information about records and aggregates of records, business processes in which records have been created and used, and agents involved in the processes (International Organization for Standardization, 2007). Besides metadata specification, SÄHKE sets functional requirements to electronic records management systems. It also defines a metadata scheme for an XML-transfer file that can be used to transfer records together with their metadata from an agency to National Archives Service of Finland.

SÄHKE has over 120 metadata elements, many of which can be used at several levels in archival hierarchy (Figure 1, below). Some metadata elements are used only in a transfer of records to National Archives Service. Altogether there are about 280 possible metadata element and entity combinations that may get a value. Thus, the SÄHKE metadata structure is quite complex.

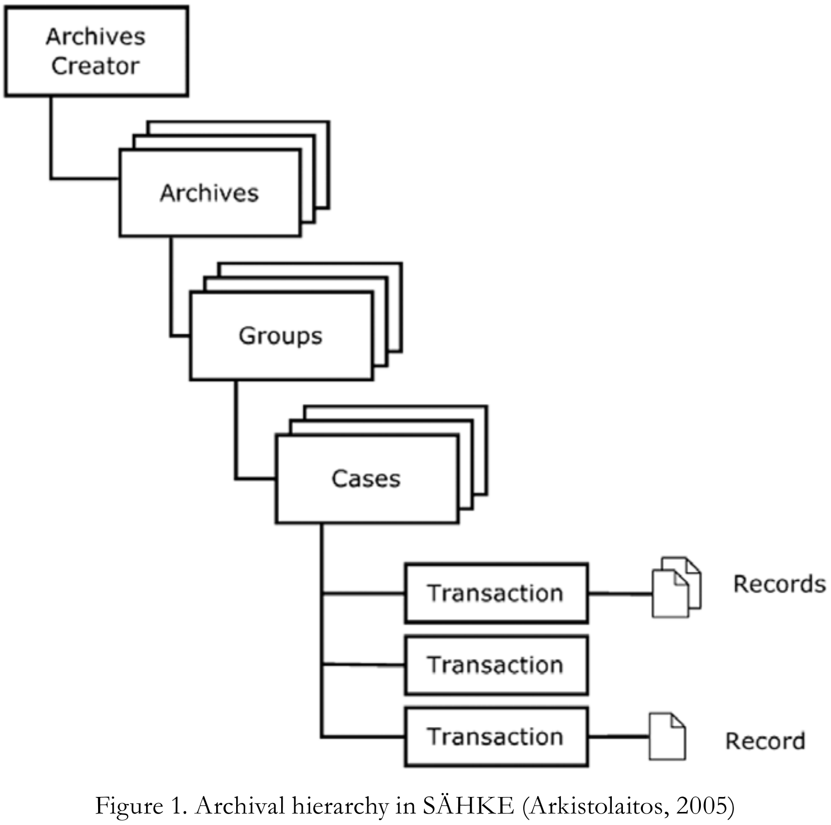

Figure 1. Archival hierarchy in SÄHKE (Arkistolaitos, 2005)

Henttonen (2009) notes that SÄHKE requirements are mostly about the core electronic recordkeeping functionality: life cycle management, access rights, retention and disposal. Consequently, metadata scheme in SÄHKE has elements important from this perspective. SÄHKE does not address questions of collaborative working, digital rights management, workflow, offline and remote working, email or integrating ERMS with content or document management.

In addition to core records management metadata, SAHKE has metadata elements which are needed in a registry filing system. SÄHKE is a specification for an ERMS in Finnish public administration. Registry filing systems have been common in the country since the 18th century. Hence, a registry is an integral part of a Finnish public sector ERMS. In a registry system, incoming and outgoing letters are registered when they enter/leave the organization. By registration a record is linked to a process. Management and citizens may use registries to follow what takes place inside administration. A "case", which is shown in the archival hierarchy (Figure 1, above), is an administrative process with a definite beginning and an end. The registry tells what transactions have taken place in a case and what records have been created in the transactions. Registry information is a part of SÄHKE metadata. Also transactions without a record created or received are described in metadata. An example of this kind of transaction is marking the case "closed". (Henttonen, 2009; for registry filing systems, see Stephens, 1995).

Metadata values in the record set came from different sources. Some values were given by users (for example, document title) during the record creation and use; others were filled in automatically by the system either without user input (e.g., date of transaction) or according to user selection (for example, document type). In some cases (like access restrictions) metadata element received a default value from the metadata associated with the organization's business classification scheme in the ERMS and the user had the possibility to change it before accepting the value.

In the study we analyzed what metadata elements were used and what not. We made statistical analysis to examine what kind of values (unique, non-unique) were given to the elements, how often the element was used, and how equally different values were distributed in records. We also grouped metadata elements to categories defined in ISO 23081-2 (International Organization for Standardization, 2007) and explored the usage of optional/mandatory elements to see how degree of optionality affects to metadata.

The first finding was that a substantial number of metadata elements in SÄHKE remained unused. The name of a metadata element has typically two parts (for instance, *Document.Title*) although in some cases the main element does not actually have a sub element. In these cases it was interpreted that the main element has one, anonymous sub-element. The table below shows that SÄHKE has 162 elements divided into 22 main elements. Almost 57 % of all sub-elements remained unused (52 % if we look only sub-elements in main elements with values).

Main-element	**Sub-element**	**N**	**Percentages**	
Used (12)	Unused	72	44.4 %	51,8 %
	Used	67	41.4 %	48,2 %
Unused (10)	Unused	23	14.2 %	
	Total	162	100.0 %	100,0 %

The second finding was about the content of the metadata in records. Altogether there were 690 048 metadata values in the record set. When we compared the metadata to the classification of metadata in the ISO 23081-2 (International Organization for Standardization, 2007), we found out that the absolute majority of metadata was about Event history (64.2 % of elements). The next came Description metadata (12.0 %), Use metadata (10.4 %) and Identity metadata (9.4 %). Four percent of elements contained Event plan metadata. There was no Relation metadata in the records. In a registry system there is perhaps more event history metadata than in non-registry system. Still, the amount of event history metadata is quite high.

There was little metadata describing directly the record content. When both *Subject* and *Abstract* elements were empty, document title was the primary descriptive element at the record level. On the other hand, the metadata makes an indirect, contextual approach possible. Bearman and Lytle (1985–86) have regarded contextual, provenance-based approach superior to subject- and content-based methods. However, collecting and studying the user's perspective has not been common in archival world (Coats, 2004). So far there are only few studies about archival provenance-based information systems (for instance,Fachry, Kamps, & Zhang, 2008; Prom, 2004).

The third result was that optional metadata elements were rarely used. It is understandable that mandatory elements are used, because system vendors and government agencies try to comply with the specification. However, it is less obvious why optional parts of the metadata specification were neglected. There are many possible reasons for this: 1) the record creating agency had decided that there is no business need for these elements, 2) users systematically skipped the elements when they added metadata to records, 3) the system vendor had not implemented the elements in the system. From the metadata alone it is no possible to say why optional metadata element values are missing.

The last finding was that metadata showed clear patterns. In other words, generally speaking metadata values were bipolar: they tend to be either always unique or never unique, always given or never given, either very evenly or unevenly distributed . This may suggest systematic recordkeeping processes and possibly minimal human intervention in metadata creation. You probably would get a result like if users prefer not to input metadata (if they have a choice) and they also tend to accept default values as such: some (perhaps unique) values are generated in the system and occur always, others are left to user consideration and are never given. Hence, indirectly, this supports the claim that users are reluctant to add metadata.

Conclusions

The project findings suggest that electronic records management systems and archival systems have to deal with a very large amount of event history metadata. At the same time there is little metadata about the record content. Because there is still little experience of using electronic records outside their original context, it remains to be seen how well the metadata is capable for satisfying various user needs. We need more research especially on the role of event history metadata. When do the users need it? How much does one need it?

It is interesting that a large portion of the metadata scheme remained unused. We need more research to understand the reasons for this. From records of one agency one should also not draw too far-reaching conclusions about what metadata is created in recordkeeping.

The findings suggest that users seldom key in records management metadata. If this is true, it may open ways to facilitate metadata creation, once we understand better record processes. For instance, in a process the person making a draft letter and the person approving it are likely to be generally the same. Consequently, the record process, and what metadata is required, may be predictable and repeating to some extent. In future we should work to see whether there are regularities of this kind in metadata creation.

Bibliography

Arkistolaitos (2005). SÄHKE-määritykset. Osa I. Abstrakti mallintaminen [SÄHKE specification. Part I. Abstract model], Available from http://www.narc.fi/sahke/Aineisto/SAHKE-abstrakti-V2-koko.pdf

Bailey, S.: *Managing the crowd. Rethinking records management for the web 2.0 world*. London: Facet, 2008.

Bearman, D., & Lytle, R. H.: *The power of the principle of provenance*. «Archivaria», 21(1985-86), pp. 15-27.

Coats, L. R.: *Users of EAD finding aids: Who are they and are they satisfied?* «Journal of Archival Organization», 2(2004), N. 3, pp. 25–39.

Fachry, K. N., Kamps, J., & Zhang, J.:*Access to archival material in context.* Paper presented at the Information Interaction in Context. Second International Symposium on Information Interaction in Context, IIiX 2008, London, UK, 14–17 October 2008.

Fensel, D.: *Ontologies: A Silver Bullet for Knowledge Management and Electronic Commerce*. Berlin, Heidelberg, New York: Springer, 2001.

Golub, K.: *Automated subject classification of textual web documents*. «Journal of Documentation», 62(2006), N. 3, pp. 350–371.

Henttonen, P.: *A comparison of MoReq and SÄHKE metadata and functional requirements*. «Records Management Journal», 19(2009), pp. 26–36.

Hou, J.-L., & Lin, F.-H.: *A hierarchical classification mechanism for organization document management*. «International Journal of Advanced Manufacturing Technology», Vol. 28 (March 2006, Issue 3/4), 417-427.

International Organization for Standardization: *ISO 23081-2. Information and documentation. Records management processes. Metadata for records. Part 2. Conceptual and implementation issues*: ISO, 2007.

Järvelin, K., & Kekäläinen, J.: Expansion tool: concept-based query expansion and construction. «Information Retrieval», 4(2001), pp. 231–255.

Kettunen, K., & Henttonen, P.: *Missing in action? Content of records management metadata in real life* [Unpublished manuscript], 2009.

Kim, Y., & Ross, S. (2007). *"The Naming of Cats": Automated Genre Classification*. «International Journal of Digital Curation», 2(2007), N. 1.

Prom, C. J.: *User interactions with electronic finding aids in a controlled setting*. «The American Archivist», 67(2004), pp. 234–268.

Stephens, D. O.: *The registry: the world's most predominant recordkeeping system*. «Records Management Quarterly», 1995, N. 1, pp. 64–66.

Suomela, S., & Kekäläinen, J.: *User evaluation of ontology as query construction tool*. «Information Retrieval», 9(2006), pp. 455–475.