JYVÄSKYLÄ STUDIES IN COMPUTING 37

Virpi Lyytikäinen

Contextual and Structural Metadata in Enterprise Document Management

Esitetään Jyväskylän yliopiston informaatioteknologian tiedekunnan suostumuksella
julkisesti tarkastettavaksi yliopiston Agora-rakennuksessa (Ag Aud. 1)
toukokuun 8. päivänä 2004 kello 12.

Academic dissertation to be publicly discussed, by permission of
the the Faculty of Information Technology of the University of Jyväskylä,
in the Building Agora, Ag Aud. 1, on May 8, 2004 at 12 o'clock noon.

UNIVERSITY OF JYVÄSKYLÄ

JYVÄSKYLÄ 2004

Contextual and Structural Metadata in Enterprise Document Management

JYVÄSKYLÄ STUDIES IN COMPUTING 37

Virpi Lyytikäinen

Contextual and Structural Metadata in Enterprise Document Management

UNIVERSITY OF JYVÄSKYLÄ

JYVÄSKYLÄ 2004

Editors
Seppo Puuronen
Department of Computer Science and Information Systems, University of Jyväskylä
Pekka Olsbo, Marja-Leena Tynkkynen
Publishing Unit, University Library of Jyväskylä

URN:ISBN 9513917835
ISBN 951-39-1783-5 (PDF)

ISBN 951-39-1760-6 (nid.)
ISSN 1456-5390

Jyväskylä University Printing House, Jyväskylä
and ER-Paino Ky, Lievestuore 2004

ABSTRACT

Lyytikäinen, Virpi
Contextual and structural metadata in enterprise document management
Jyväskylä: University of Jyväskylä, 2004, 73 p.
(Jyväskylä Studies in Computing
ISSN 1456-5390;37)
ISBN 951-39-1783-5
Finnish summary
Diss.

Documents have a central role in organizations. While the amount of information continually increases, new kinds of methods for managing the documents are needed. Enterprise document management concerns the whole life cycle of documents in organizations, from emergence to disposition, and also development of document management solutions. The utilization of metadata describing documents has been seen as an answer to the problems of finding relevant and avoiding irrelevant information. Previous research on metadata has commonly concentrated on defining appropriate metadata for document instances. This thesis is, however, focused on metadata related to collections of documents. The thesis emphasizes contextual and structural metadata of document collections. Information about the context where documents are produced or used is called contextual metadata. In the thesis, however, contextual metadata refers only to document production context. Structural metadata, on the other hand, describes the logical structure of the documents by document type definitions or schemas. The thesis describes methods and techniques for collecting and using the metadata. The methods and techniques were developed and tested in three projects in three case environments. The results of the thesis show how contextual and structural metadata can be collected and utilized in document analysis, user requirements elicitation and information retrieval. An example of a contextual metadata schema in XML format is also included. In the future, security issues, like access rights, could be included in the metadata definition.

Keywords: metadata, content analysis, document analysis, user requirements analysis

ACM Computing Review Categories

H.1.0. Information Systems: Models and Principles: General

I.7.1 Computing Methodologies: Document and Text Processing, Document and Text Editing
Content analysis
Requirements analysis

K.6.1 Management of Computing and Information Systems: Project and People Management
Systems analysis

Author's address
Virpi Lyytikäinen
University of Jyväskylä
Dept. of Computer Science and Information Systems
P.O. BOX 35 (AgC 519.4), 40014 Jyväskylän yliopisto
Finland. E-mail: lyviau@cc.jyu.fi

Supervisor
Professor, Ph.D., Airi Salminen
Dept. of Computer Science and Information Systems
University of Jyväskylä
E-mail: airi@cs.jyu.fi

Reviewers
Professor, Ph.D., Wendy Duff
Faculty of Information Studies
University of Toronto
140 St. George Street
Toronto, Ontario M5S 3G6, CANADA

Professor, Ph. D., Cecilia Magnusson Sjöberg
Law and Informatics Research Institute
Department of Law
Stockholm University
S -106 91 Stockholm, SWEDEN

Opponent
Associate Professor, Ph. D., Maria A. Wimmer
Institute of Applied Computer Sciences
Division: Business, Administration and Society
University of Linz
Altenbergerstr. 69
A-4040 Linz, AUSTRIA

ACKNOWLEDGEMENTS

The research described in the thesis would have been impossible without the help and support of many persons. First of all, I want to thank my supervisor, Professor Airi Salminen, who through many years has had wisdom and patience to encourage me in this eventful journey. Your faith in me goes beyond my understanding!

I am also very grateful for the reviewers of the thesis, Professor Wendy Duff (University of Toronto) and Professor Cecilia Magnusson Sjöberg (Stockholm University). Their comments and suggestions encouraged me to finish the writing of the thesis. I owe a lot to Professor Pasi Tyrväinen for his innovative ideas and support. Thank you for keeping your office open for me!

My colleagues at the University of Jyväskylä have provided very pleasurable moments in work and also out of office. Pasi Tiitinen has been my closest colleague, with whom I have had a pleasure to work in many projects, write several articles, see interesting places, and share my thoughts on various subjects. Thank you for your friendship! Also, I want to thank my dear colleague and friend, Ph.D. Anne Honkaranta for her co-operation, example, and great support during the years. Our discussions while shopping and playing badminton have meant a lot to me! In addition, Riitta Kuisma, Marketta Niemelä, Antti Lehtinen, Tero Päivärinta, Timo Riepponen, Kalevi Manninen, Rauno Veijola, and Matti Järvenpää have been wonderful companions during the years.

Many people in collaborative organizations have enabled the research. Olli Mustajärvi from the Finnish Parliament has given a lot of insightful comments and remarks. Jean-Luc Vidick, Laurent Mercier from Atos S.A., and Filip Evenepoel from Katholieke Universiteit Leuven, Belgium, offered their opinions into problems of retrieving legal documents from heterogeneous databases. Special thanks are owed to Vesa Pylvänäinen from the Pentecostal Church in Jyväskylä for his time, energy and courage to answer difficult questions!

The funding of the research has come from numerous sources, including the Finnish Parliament, Telematics Application Programme of the European Commission, Finnish Technology Development Centre, and other project participants. The Department of Computer Science and Information Systems has provided space, equipment and a pleasant environment for the research.

Without friends outside the academic atmosphere my life might have been quite unvaried. Thank you, Anneli, Irene, Kirsti, Mirja, Pirkko and many other friends, choirs and committees for showing me different aspects of life. I also want thank my parents, Hilkka and Jukka, for many kinds of support in my life.

Finally, my deepest gratitude is owed to Jesus, "in whom all the treasures of wisdom and knowledge are hidden" (Col. 2:3). I pray that He may teach me to number my days aright, that I may gain a heart of wisdom (Psalm 90:12).

Virpi Lyytikäinen
Jyväskylä 20.4.2004

FIGURES

TABLES

CONTENTS OF THE INTRODUCTION

LIST OF INCLUDED ARTICLES

1 Salminen, A., Lyytikäinen, V., & Tiitinen, P. 2000. Putting documents into their work context in document analysis. *Information Processing & Management* 36 (4), 623-641.

2 Tiitinen, P., Lyytikäinen, V., Päivärinta, T., & Salminen, A. 2000. User needs for electronic document management in public administration: a study of two cases. In H.R. Hansen, M. Bichler, & H. Mahrer (Eds.), Proceedings of ECIS 2000, European Conference on Information Systems, Volume 2, Wien: Wirtschaftsuniversität Wien, 1144-1151.

3 Salminen, A., Tiitinen, P., & Lyytikäinen, V. 1999. Usability evaluation of a structured document archive. In R.H. Sprague, Jr. (Ed.), Proceedings of the Thirty-Second Hawaii International Conference on System Sciences (file ddhfu06.pdf at CD-ROM). Los Alamitos, CA: IEEE Computer Society.

4 Salminen, A., Lyytikäinen, V., Tiitinen, P., & Mustajärvi, O. 2001. Experiences of SGML standardization: The case of the Finnish legislative documents. In R.H. Sprague, Jr. (Ed.), Proceedings of the Thirty-Fourth Hawaii International Conference on System Sciences (file etegv01.pdf at CD-ROM). Los Alamitos, CA: IEEE Computer Society.

5 Lyytikäinen, V., Tiitinen, P., Salminen, A., Mercier, L., & Vidick, J.-L. 2000. Visualizing legal systems for information retrieval. In M. Khosrowpour (Ed.) Challenges of Information Technology Management in the 21st Century, Proceedings of 2000 Information Resources Management Association International Conference, Hersley, PA: Idea Group Publishing, 245-249.

6 Lyytikäinen, V. 2003. Analysing requirements for content management. In O. Camp, J. Filipe, S. Hammoudi & M. Piattini (Eds.), Proceedings of the 5th International Conference on Enterprise Information Systems, Vol. 3, Angers, France, April 23-26, 2003, Portugal: Escola Superior de Tecnologia do Instituto Politécnico de Setúbal, 104-111. Also published in O. Camp, J. Filipe, S. Hammoudi & M. Piattini (Eds.), Enterprise Information Systems V. Kluwer Academic Publishers B.V.

7 Honkaranta, A., & Lyytikäinen, V. 2003. Operationalizing a genre-based method for content analysis: A case of a church. In W. Abramowicz & G. Klein (Eds.), Proceedings of the 6th International Conference on Business Information Systems. Software Engineering Track. Colorado Springs, 4-6 June, 2003. Poland: Department of Management Information Systems at the Poznan University of Economics, 108-116.

INTRODUCTION

1 METADATA AND DOCUMENT MANAGEMENT

The amount of textual and multimedia information in enterprises is growing all the time. The major portion of the information is stored in documents (Sprague, 1995). In this thesis, the term *document* is defined according to Sprague (1995, pp 32) as "… a set of information pertaining to topic, structured for human comprehension, represented by a variety of symbols, stored and handled as a unit". *Document management* in turn refers to definition, creation, storage, organization, transmission, retrieval, manipulation, updating, and disposition of documents. Document management in many enterprises can be seen as a crucial task, because in many business areas, like e-government and e-business in general, the work itself is quite document-centric. Also in more traditional areas, like manufacturing, different kinds of documents form a huge pool of information. The research on *enterprise document management* seeks to find solutions for managing documents across and between enterprises, and emphasizes the connection of document management to the business processes of enterprises.

Methods and techniques to manage documents are more and more needed. Especially, new kinds of retrieval methods relying on metadata related to the documents should be developed. Contextual metadata describing the context where documents have been created could be useful in finding the needed information. My research has concentrated on utilizing metadata concerning collections of documents of the enterprises. The term 'enterprise' is understood widely, to cover organizations both in private, public, and so called third sector. The case organizations of the thesis, however, represent public and third sector organizations.

In the following sections the most central terms of this thesis are further discussed. At first, the notion of metadata is studied in more detail, and then concepts related to document management are described. The following sections also give short introductions to previous literature of the area.

1.1 What Is Metadata?

The most well-know definition for the metadata is "data about data". It has also been said that metadata characterizes documents for discovery and use (Murphy, 1998). In order to understand more thoroughly the purpose and meaning of the word, however, a more profound definition is needed. Gilliland-Swetland (1998, pp 1) has defined metadata as "the sum total of what one can say about any information object at any level of aggregation". The information object in question can be anything that can be addressed and manipulated by a human or a system as an entity. For the purpose of the thesis, the definition of Gilliland-Swetland serves as a starting point.

According to the definition metadata is quite a broad concept. Metadata can, for example, identify and describe an information object, document the object's behavior, functions and use, and describe relationships to other objects. Figure 1 models concepts relating to metadata according to Bearman, et al. (1999). The figure is a simplification of a model produced in cooperation by Dublin Core metadata community and INDECS/DOI community of authors, rights holders and publishers. In the model, the term information resource is used instead of information object. According to the model, there can be various kinds of information resources, which have relations to each other. Each information resource is about some subject. Actions are performed in order to transform information resources. Actions take place in a specific time and place. Agents that can be persons, corporate bodies, or instruments, have a role in performing the actions. A new draft ISO metadata standard for records management includes in their concepts also an entity called 'mandates', which represents all regulations involved with the management of an information resource

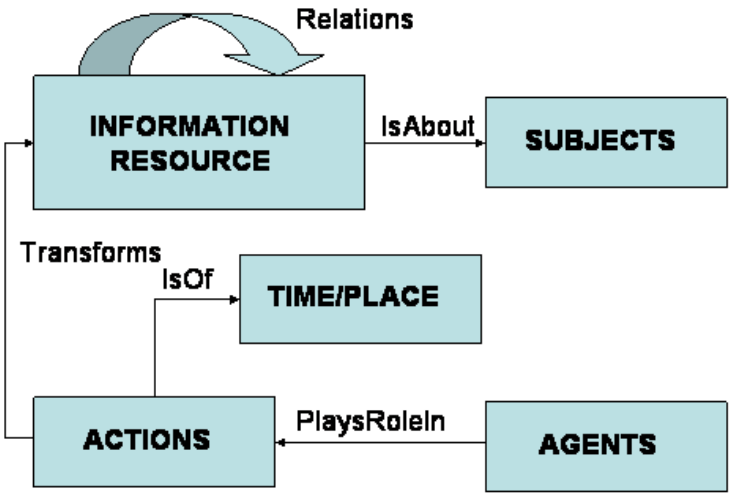

FIGURE 1 Concepts Relating to Metadata

(ISO TC 46/SC 11/WG 1, 2003). Open Archives Information System Reference Model (OAIS) provides, in addition to the above described concepts, also functionalities of a system dealing with archival metadata. Producer, consumer, and management are roles to be supported with predefined functionalities like preservation planning, archival storage, data management, ingest, access, and administration (Consultative Committee for Space Data Systems, 2002).

In the following, a review of metadata categorizations in previous literature is given. Also, the roles, models, and schemas for metadata are examined.

1.1.1 Metadata categorizations

In order to understand varying kinds of uses of metadata, research literature provides categorizations for the concept. Different authors have used different kinds of categories, depending on their interests. The following descriptions of the metadata categories should give quite a broad view of key aspects of the functionality of metadata.

Gilliland-Swetland (1998) breaks the concept of metadata into five categories: *administrative*, *descriptive*, *preservation*, *use*, and *technical* metadata. Administrative metadata is used for managing and administering information resources. Examples of such metadata are location and acquisition information. Descriptive metadata is used to describe or identify information resources. Cataloging records and specialized indexes are examples of this kind of metadata. Preservation metadata relates to preservation management of information resources, for example, documentation of physical condition of resources. The use category is related to the level and type of use of information resources, like use and user tracking, or content re-use and multi-versioning information. Technical metadata describes system's functions or metadata behavior. Examples of this kind of metadata are hardware and software documentation, or authentication and security data.

Gilliland-Swetland (1998) also gives attributes for metadata. Each attribute describes different aspects of metadata. The *source* of metadata can be either internal, i.e. generated by a creating agent at the time when the information object is first created, or external, in which case the metadata is created later (see also, Saarela, 1999). The *method for metadata creation* can be automatic, produced by computer (e.g., Han et al., 2003; Soo et al., 2003) or manual, produced by humans[1] (e.g. with Dublin Core metadata, see for example, Murphy, 1998). Metadata can be either lay or expert of its *nature*, depending on the expertise of the persons creating the metadata. *Status* of the metadata can be static, dynamic, long-term, or short-term. Metadata can either be unstructured or structured conforming to a predictable structure. The *semantics* of the metadata can be expressed either with standardized vocabulary or by uncontrolled metadata, like free-text notes. The last attribute characterizing metadata is *level*, which can range from collections to individual information objects.

---

[1] A third class of metadata creation also exists: semi-automatic metadata creation (Jokela et al., 2000; Rödig et al., 2003)

A more technical view to categorizing metadata for individual information objects is adopted in Jokela (2001) and Pöyry et al. (2002). There are three main categories defined: *structural*, *control*, and *descriptive*. Structural metadata describes the format of the information object, for example, video, audio, and graphics formats, or compression data. Control metadata is meant for controlling the flow of content of information objects, for example it can be used to determine whether the information object is ready for the next step in its production process. The last category, descriptive metadata, can be further divided into subcategories, *contextual* and content-based, *semantic* metadata. Contextual metadata in this vocabulary describes the environment and conditions, where the information object is created, for example, geospatial information, timing information, or information about the equipment used in production process. Semantic metadata, in turn, describes what the content of the information object actually means (Jokela et al., 2000).

Murphy (1998) offers two frameworks for categorizing metadata structures. The first one is adopted from Dempsey (1996). The framework places different metadata structures into continuum based on their relative richness. Relative richness depends on four areas: amount of manual effort in creation, amount of specialization, level of description, and external sources for metadata content. According to the framework, terse metadata have a small range of information on a very large number of networked information objects. Rich metadata structures, in turn, have a large amount of data on fewer information objects. The second framework by Murphy (1998) categorizes metadata structures by logical and physical separability. Metadata is logically separable if it is not determinable directly of the full content of the document. Physical separability means that a metadata structure has to be able to persist when the object it represents no longer exists.

Structural metadata is the metadata category in focus in the study of Dushay (2002). Other categories mentioned are *descriptive* metadata (Dublin Core records, as an example), and *administrative* metadata pertaining digital file creation and storage, rights management, etc. Structural metadata in Dushay's study is meant for mapping relationships among the components of information objects. The mapping can be done either by assigning labels or a hierarchy. Both options can also be used simultaneously.

Böhm and Rakow (1994) have defined six categories for metadata for multimedia documents: metadata for the *representation of media types*, *content-descriptive* metadata, metadata for *content classification*, metadata for *document composition*, metadata for *document history*, and metadata for *document location*. All these relate to multimedia documents. Böhm and Rakow also give an example, *statistical metadata*, of metadata for collections of multimedia documents.

Boll et al. (1998) also view the notion of metadata from multimedia perspective. Figure 2 illustrates their metadata classification with examples. The metadata classification is adapted from Kashyap and Sheth (1997). In this classification, metadata is first divided into two main classes: *content-independent* metadata and *content-dependent* metadata. The latter is then further divided into

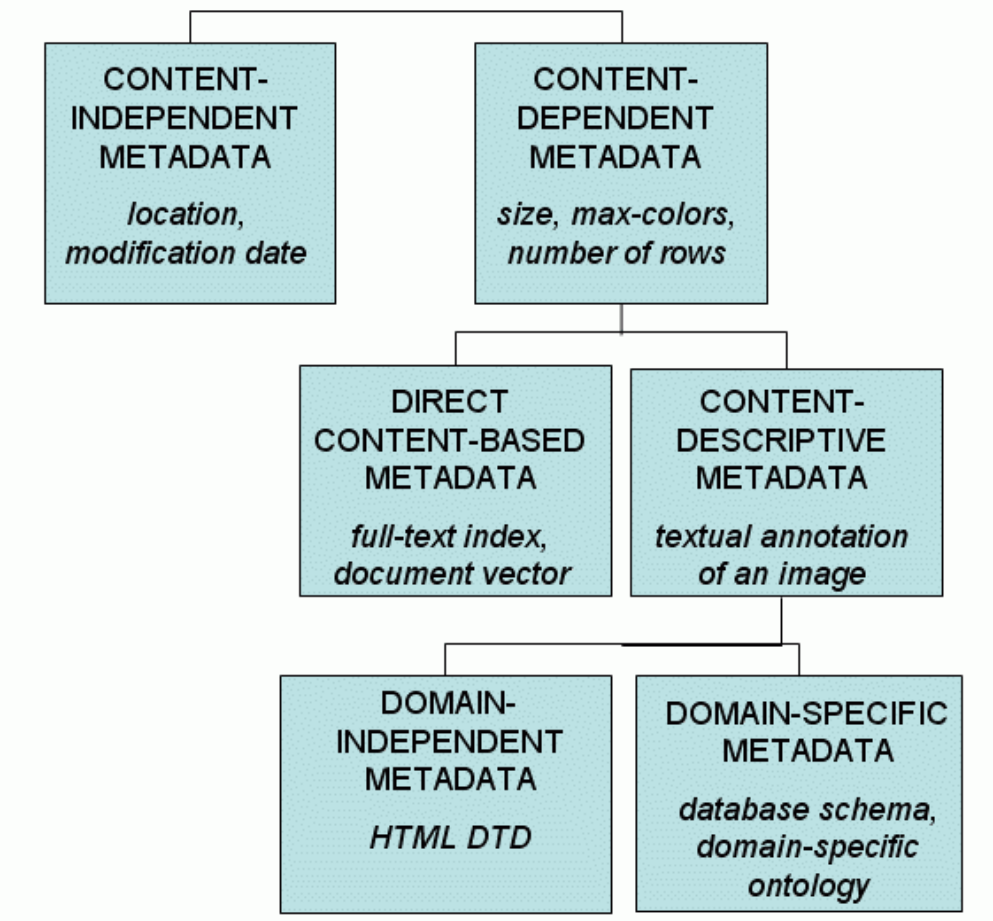

FIGURE 2 A Metadata Classification of Boll, et al. (1998)

two: *direct content-based* metadata and *content-descriptive* metadata. Content descriptive metadata can be either *domain-independent* or *domain-specific*.

Metadata for electronic document collections have been especially discussed by Hill, et al. (1999). They categorize collection metadata into two classes: *inherent metadata*, i.e. information that can be derived by computers from the content of the collection, and *contextual metadata*, i.e. information supplied by the collection provider that cannot be derived from the content. Examples of contextual metadata of a collection are title, responsible party, scope and purpose, type of collection, and update frequency. Hill, et al. (1999) define the use of collection metadata to be four folded:

- *collection identification* for the software trying to access the collection,
- *network discovery*, for search agents to know what the collection contains,
- *user documentation* about the collection and the digital library around it,
- *management* of the collection for providing information about where the content of the collection is stored.

1.1.2 Roles of metadata

There are several purposes for which metadata can be used in different phases and tasks during the management of the information objects. Examples of such are creation of information objects, information retrieval, archiving and long-term preservation, and digital rights management. Metadata can also serve as evidence, for example, for legal or audit purposes (Bearman & Sochats, 1996; McKemmish & Acland, 1999). In the following, the different kinds of roles for metadata are shortly described.

In creating information objects, documents especially, the metadata about the structure of the documents can be used as an aid. When the structure of the documents is known, text processing systems can assist the author in producing the document content so that the result can be further manipulated with information systems (e.g., Leinonen & Penttonen, 1998; Fahrenholz-Mann, 1999).

Formal metadata descriptions have been in use in libraries for decades in organizing collections and assisting finding and retrieving relevant information. The metadata in libraries includes classifications, indexes, abstracts, and catalog records (Gilliland-Swetland, 1998). One well-known format of the catalog records is MARC (Machine-Readable Cataloging format). Also the structure of the documents has been an important aid in information retrieval even before electronic publication era. If the users are familiar with the structure of the documents, they can use the knowledge to scan the documents and find the relevant piece of information (Gilliland-Swetland, 1998). For example, if they know the structure of a meeting memorandum, they can easily find the names of the participants of that meeting. Nowadays, information retrieval takes place also outside libraries, due to Internet. Therefore, new kinds of retrieval aids have been designed. For example, integration of metadata databases and full-text databases has also been considered as one solution for assisting users in their information retrieval tasks (Deniman et al., 2003). Profound review of the information retrieval on the World Wide Web environment can be found in the study of Kobayashi and Takeda (2000).

In archiving and long-term preservation, the context of the archived information object assists identifying and preserving their evidential value of records over time. It is essential to be able to authenticate the objects, analyze them, and interpret their meaning. Therefore, the metadata used in archival science includes accession records, finding aids, and catalog records for describing the archived content (Gilliland-Swetland, 1998). Metadata definitions and implementations in the domain of archiving and long-term preservation have been discussed, for example in (Rothenberg, 1995; Bearman, 1999; Duff, 2001; Rödig et al., 2003).

The management of digital rights is quite a new area of metadata utilization (Luoma et al., 2003). The emergence of electronic publication has created a situation where traditional means for managing the intellectual property rights are not satisfactory. The new situation is much more complex than in the era of printed books. Now, different kinds of content are mixed together, and their

rights are traded, or sometimes given freely, in complex ways (Bearman et al., 1999).

1.1.3 Standards for metadata

For describing and defining metadata there are several internationally developed standards on different levels. Since the amount of standardization efforts and standards themselves is increasing all the time, only examples of standards are given here.

On the most detailed level, each domain has developed its own standard to suit for their special needs. Examples of such standards are

- Australian Recordkeeping Metadata Schema (RKMS) for archiving purposes (Recordkeeping Metadata Standard for Commonwealth Agencies, 1999),
- International standard for records management (ISO 15489-1, 2001),
- Open Digital Rights Language (ODRL) for defining rights for different kinds of use of electronic material (Iannella, 2002),
- Learning Object Metadata (LOM) for Web-based education (Hodgins, 2002), and
- Governmental Markup Language (GovML) for describing governmental data and metadata (Wimmer & Tambouris, 2002; Kavadias & Tambouris, 2003).

Many of the detailed metadata standards are based on, or provide mappings to a more general, cross-domain metadata standard, Dublin Core Metadata Element Set (DC) (DublinCore, 2003). DC was developed by the library community for resource discovery on the World Wide Web. DC identifies a set of 15 metadata elements to be used for any domain on the Web. The metadata elements of DC include, for example, title, creator, subject, and description. The elements are optional, repeatable, and can be modified and extended for local needs (Murphy, 1998). Therefore, DC metadata set can be used in multiple languages (Baker, 1998).

Resource Description Framework, RDF (Lassila & Swick, 1999) is meant for encoding metadata in various application domains (Saarela, 1999). RDF's basic concepts are *resource*, *property*, and *statement*. Resources are anything that can be addressed via Uniform Resource Identifier, URI (Berners-Lee et al., 1998). A property, in turn, is a specific aspect, characteristic, attribute, or relation used to describe a resource. A statement is formed, when a specific resource and a named property plus a value for that property are expressed together. These three parts of a statement are also called *subject*, *predicate*, and *object*, respectively (Lassila & Swick, 1999). RDF has been used to express different metadata standards, for example Dublin Core (Kokkelink & Schwänzl, 2002).

1.1.4 Summary and positioning of the thesis

The purpose for gathering the metadata also affects the need to define and categorize different kinds of metadata. In this thesis quite a broad definition of general metadata by Gilliland-Swetland (1998, pp 1) is adopted:

> Metadata is the sum total of what one can say about any information object at any level of aggregation.

However, the study described in the thesis has not covered quite everything one could say about all the information objects with whatever granularity. The scope of the thesis is metadata related to collections of documents, a research area not so deeply covered by earlier research. The term 'collection' in this context means a set of related documents. From all the possible things that one could say about document collections, the thesis discusses only issues relating to the context and structure of the documents in collection, i.e. contextual and structural metadata. The context in the thesis covers the relationships between document types, document production processes as well as the roles of the actors producing the documents. Issues relating to technology used for document production are purposefully left out of the scope, even though some scholars include them to the area of contextual metadata. Within an individual document collection instance, the metadata discussed in the thesis is all the time focused on class level, not on individual instances of documents, processes or actors. Structural metadata in the thesis is defined to describe the logical structure of the documents by document type definitions or schemas. Relationships between individual document instances are not of interest here.

Table 1 includes all the metadata categorizations presented above as a summary. Categories that match with the scope of this thesis are marked with bolding. Contextual and structural metadata of the thesis fits into the descriptive category of Gilliland-Swetland (1998), because the purpose of defining the context and structure is to describe document collections. Also according to Jokela's (2001) categorization the focus of the thesis fits most conveniently into descriptive category, although here the contextual metadata does not describe the technical issues defined by Jokela et al. (2000). The structural metadata of the thesis is not defined similarly to Jokela (2001) because here the focus is not on presentation formats, but in logical structures of documents. From the first category presented by Murphy (1998) the contextual and structural metadata can be said to be quite rich, since its creation requires a lot of manual effort and expertise. According to the second framework by Murphy (1998) the contextual metadata here is both logically and physically separated, since the content of the metadata cannot be derived from the document collection nor it is its existence dependable on the existence of the collection. Structural metadata, in turn, can also be logically integrated, since sometimes it is possible to derive the document type definition from a set of document instances (Ahonen, 1996). Structural metadata of Dushay (2002) is not exactly the same as the one in the thesis, since here structural metadata describes logical structures of the docu-

TABLE 1 Summary of metadata categorizations

	Gilliland-Swetland (1998)	Jokela (2001)	Murphy 1 (Murphy 1998)	Murphy 2 (Murphy 1998)	Dushay (2002)	Böhm & Rakow (1994)	Boll et al. (1998)	Hill, et al. (1999)
Categories	Administrative **Descriptive** Preservation Technical Use	Structural Contol **Descriptive** Contextual Semantic	**Rich** Terse	Logically integrated **Logically separable** Physically integrated **Physically separable**	Administrative **Descriptive** Structural	Representation of media types **Content-descriptive** Content classification Document composition Document history Document location	Content-independent Content-dependent Direct content-based Content-descriptive Domain-independent **Domain specific**	Inherent **Contextual**
Purpose	General	Individual multimedia documents	General	General	General	Multimedia documents	Multimedia documents	Electronic document collections

ment classes rather than actual hierarchies of components in document instances. Since descriptive metadata in Dushay (2002) is meant for resource discovery, it is the category closest to the focus of the thesis. If the categories of Böhm and Rakow (1994) were metadata of collections, then the category of content-descriptive metadata would the closest to the theme in the thesis. Since the contextual and structural metadata are dependent both of the content and domain, the most fit category of the Boll, Klas and Sheth (1998) would be the domain-specific metadata. From Hill et al. (1999) the reference here can be made to the contextual metadata. The relationships between the thesis and other metadata categorizations are revisited in Section 6.

From different roles for metadata the interesting ones from this thesis point of view are document management in general, but especially the creation of documents and information retrieval. Research concerning document managment will be discussed in the following sub-section. The thesis also includes a suggestion for a specific structure of contextual metadata, which could also be called as a suggestion for a metadata standard.

1.2 Research in Enterprise Document Management

Documents as a means for communication have a long history. Also the connotations related to the document concept vary (see, for example, Schamber, 1996; Buckland, 1997; Päivärinta & Tyrväinen, 1998; Salminen, 2000). The definitions of the term range from a simple statement like "any expression of human thought" (Buckland, 1997) to more descriptive listing of characteristics of a document (Salminen, 2000;2003):

- It is intended for human perceptions, to be understood as information pertaining to a topic.
- It has content and one or more external representations.
- The content consists of parts, parts, consists of symbols, parts are structured to support human understanding.
- It is stored on media.
- It can be identified and handled as a unit.

In this thesis, the definition of Sprague (1995, pp 32) for the term document is considered sufficient. Sprague states that a document is "… a set of information pertaining to topic, structured for human comprehension, represented by a variety of symbols, stored and handled as a unit".

Document management also has been studied for quite a while in different research areas in information systems research, like office automation, computer-supported collaborative work, enterprise resource planning, and naturally, electronic publishing. The history of document management research has been discussed in more detail in Päivärinta (2001).

Enterprise document management can be focused in quite many areas. Examples of these are document standardization (Salminen, 2000), and informa-

tion retrieval (Gordon, 1997; Blair, 2002). From the viewpoint of the focus of this thesis research on *structured documents* is essential. Structured documents are documents, which have a formally defined structure exploitable by computers. Standard Generalized Markup Language, SGML (Goldfarb, 1990), and Extensible Markup Language, XML (Bray et al., 2000) are languages that can be used for structured documents. In both SGML and XML the structure of the document instances is indicated by markup. The rules for marking up the documents are defined in *document type definitions* (DTDs) or other kinds of schemas (e.g., Fallside, 2001). A schema defines names and hierarchic structure of the elements forming the logical structure of documents. Attributes attached to elements increase the possibility to define elements in more detail. XML is a simplified form of SGML having a smaller set of rules and features especially planned for use in the Internet environment (Salminen, 2003). Hypertext Markup Language, HTML (Raggett et al., 1999) is an application of SGML with its own DTD. The use of the structured documents offers various means for automation, and therefore also several research possibilities have risen. The research areas include document transformation (Lindén, 1997), document assembly (Heikkinen, 2000), information retrieval (Salminen & Tompa, 1992; Salminen & Watters, 1992; Kuikka & Salminen, 1997), and document modeling and design (Watson & Shafer, 1995; Maler & El Andaloussi, 1996; Salminen & Tompa, 1999). Document analysis, DTD design, and markup on the domain of legal documents are discussed in the study of Magnusson Sjöberg (1998).

Enterprise document management is not only a technical issue, but concerns also business processes and work of people (Sprague, 1995; Gordon, 1997; Päivärinta, 1999). This organizational context is considered, for example, in the method of Sutton (1996), the genre-based method (Päivärinta, 2001), and the requirements definition method of Barry (1993). In the document analysis methodology expanded in this thesis, organizational context has been a characteristic feature from its beginning (Salminen et al., 1996; Salminen et al., 1997).

Theory of genres offers one approach for studying document management in organizations (Päivärinta et al., 1999; Tyrväinen & Päivärinta, 1999; Karjalainen et al., 2000; Karjalainen & Salminen, 2000; Päivärinta, 2000; Karjalainen & Tyrväinen, 2001; Päivärinta & Peltola, 2001). A *genre* can be defined as a prototypical model for communication (Swales, 1999). Genres are not bound by technology or media used, instead they concentrate on communicational aspects in organizational context. One framework for defining the organizational context with respect to genres is called 5W1H (Orlikowski & Yates, 1998; Yoshioka et al., 2001). The framework considers the following aspects of genres: *why* (purpose of communication), *what* (expected content), *how* (media, or type of language), *who(m)* (who are the communicators), *when* (time schedules, deadlines, duration), and *where* (physical or virtual places).

Research based on genre theory in conjunction with document management leads to considering the communication in the organizations more broadly, noticing also other forms of content than those that can be thought of as documents. Content management is a field that covers different tools and

methods for collecting, processing and delivering content of diverse types (McIntosh, 2000; Boiko, 2002). In the thesis the focus, however, is in managing documents.

In the following sections first the research settings will be discussed in Section 2. Then the creation and utilization of contextual and structural metadata will be described in Sections 3 and 4, respectively. Section 5 includes overviews of the articles included in the thesis. Finally, Section 6 concludes the thesis.

2 RESEARCH GOAL, METHODOLOGY AND PROCESS

In this section the research goals of the thesis as well as the methodology used are discussed. In addition, a description of the research as a process is given.

2.1 Research Goal and Methodology

The overall goal of the research has been to **develop electronic document management in the enterprises by use of contextual and structural metadata**. The research questions can be stated as follows:

1. How to collect and create contextual and structural metadata?
2. How to utilize contextual metadata in document analysis and document management?
3. How to visualize contextual metadata for users?

In order to utilize contextual and structural metadata it should first be either created or collected from existing sources. Metadata creation should not be too time-consuming in relation to the benefits the metadata offers. Otherwise, the effort is not worthwhile.

When the contextual metadata exists, there should be feasible use for it. In the thesis, the focus is in utilizing contextual metadata during document analysis and, later on, during document management tasks.

For the users to be able to utilize the contextual metadata in information retrieval, the metadata should be made visible, i.e. visualized. The visualizations should be simple enough to be understood intuitively, without long explanations. At the same time, visualizations should offer enough information for the users so that they find it useful in their tasks. The research questions are discussed more deeply both in the introductory part and the individual seven articles following the introduction.

The research described in the dissertation can be seen as a multimethodological development, consisting of four research activities. The activities are interrelated (Figure 4) (Nunamaker et al., 1991):

1. *Theory building* includes development of new ideas, concepts, frameworks, methods, or models. In this study, different kinds of methods were built. The methods include general document analysis method (Article 1), method for participatory design of document structures (Article 7), methods and techniques for user requirements elicitation (Articles 2 and 6), usability evaluation method for structured document archives (Article 3), and method for visualizing contextual metadata for information retrieval (Article 5). Also, a model for the structure of the contextual metadata was developed (Article 5). The structure is defined in the format of XML DTD.
2. *Experimentation* normally includes laboratory and field experiments, or computer and experimental simulations. In this study, experimentations were conducted in order to test the methods and techniques built. Experimentations include case studies (Yin, 1994) of user requirements elicitation (Articles 2 and 6), action research (Susman & Evered, 1978; Kock et al., 1997) on defining document structures (Article 7), usability evaluation (Nielsen & Mack, 1994) for a prototype archive (Article 3), and testing the graphical models visualizing the contextual metadata in user interfaces (Article 5).

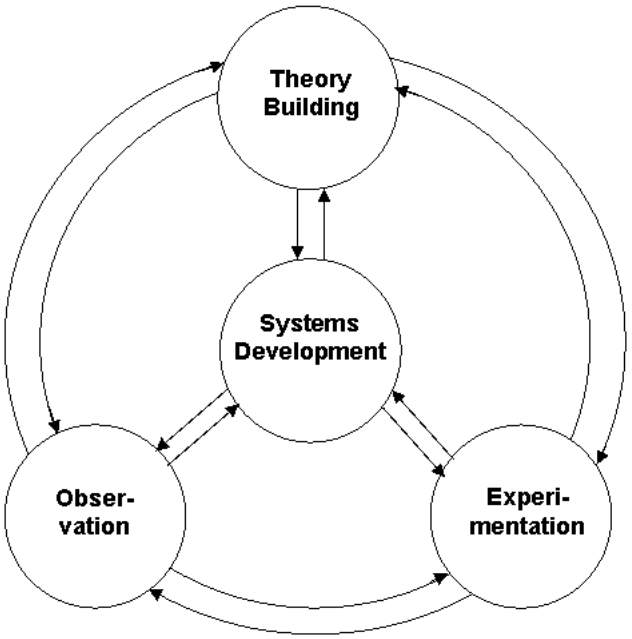

FIGURE 3 Multimethodological Research Approach (Nunamaker et al., 1991, pp 209)

3. *Observation* includes case studies, field studies and sample surveys that are unobtrusive research operations. In this study, observation was conducted in order to find out the effects of the document standardization efforts where document analysis method was tested (Article 4).
4. *Systems development* traditionally includes concept design, constructing the architecture of the system, prototyping, product development, and technology transfer. In this study the developed systems, in addition to the prototypes of structured document archive and visualized graphical user interfaces for document retrieval, were document management environments in case organizations. The prototype systems were implemented by different partners in our research projects.

2.2 Research Process

The constructive, multimethodological research was carried out in the Department of Computer Science and Information Systems at the University of Jyväskylä in different document management projects. The projects were collaborative efforts with several partners outside the university. The work on my part started in 1996 in a project called RASKE (Rakenteisten AsiakirjaStandardien KEhittäminen, Development of standards for structured documents), continued in a European level project called EULEGIS (European User Views to Legislative Information in Structured Form), and final results were developed in the HeTi project (Helluntaiseurakunnan Tiedottamisen kehittäminen, Development of information about church services), which started in 2003.

The major practical purpose of the RASKE project was to improve document management in the Finnish Parliament and ministries, and accessibility of the information created in the Parliament and ministries in the Finnish society. The project had started in 1994 with document analysis concerning Finnish legislative documents (Salminen et al., 1996; Salminen et al., 1997). The practical goal of the analysis was to adopt SGML as document production format for the Finnish parliamentary documents. When I joined the project, in its second period, the domains to be analyzed were the Finnish budgetary process and documents relating to the Finnish participation in EU legislative work.

As a result of the work in the RASKE project, a document analysis methodology was developed (Salminen, 2003). During the second period of the project we complemented process modeling components in the methodology (Article 1), and developed the user requirements analysis methods and techniques (Article 2). In the methodology the contextual metadata is collected during the document analysis process, and it is used in the user requirements elicitation phase as a tool to focus the interviews. The use of the methodology produces structural metadata in the form of document type definitions.

A publishing house participating in the RASKE project implemented a prototype archive of legislative documents in SGML format. In order to evalu-

ate the archive, we developed a usability evaluation method for structured document archives. The method was also tested during the project (Article 3).

After we had designed preliminary document structures for the documents in the analyzed domains, the actual implementation phase was commenced. A commercial software house was selected to be responsible for the effort. When the implementation was completed, we evaluated the results of the SGML standardization case (Article 4). The evaluation was based on the user interviews and reports of the implementation phase.

The EULEGIS project was founded in 1998 by a consortium of nine full members from seven European countries. The main practical intent of the project was to offer a single point Web access to European legal information created in different legal systems and at different levels - European Union, a country, a region, or a municipality – through a unified interface (Lyytikäinen et al., 2000a;b). From scientific research point of view the purpose was to investigate possibilities to use structured documents in European legislative domain.

The EULEGIS project lasted two years. In the beginning a user needs analysis was conducted. People around Europe were interviewed either with semi-structured interviews or by queries. The analysis revealed that people who use databases available in the Internet containing legal documents of European countries need more information concerning legislative systems in Europe. Due to the differences in legal systems, legislative databases, and also in users' expertise, the information about the context where the documents were created was considered as a valuable aid for the users. In order to give the users the information they needed, graphical user interfaces based on the models created during document analysis were developed. Through the visualized models the users were able to access the legislative documents in various databases around Europe without needing to know in advance in which database the documents they wanted to reach reside. The visual user interfaces showed the relationships of the legislative documents, the actors that produce the documents, and the production processes in a certain legal system. In order to accomplish the implementation of the user interfaces we had to define a formal structure for the contextual metadata. For this we chose to use the modeling capabilities of XML DTDs. Opinions of the users were collected in small scale by demonstrating the prototype system in group sessions.

In the fall 2002 HeTi project was started. The purpose of the project was to improve the communication of the information concerning church services in a church of 1700 members residing in Central Finland. The church has 9 employees in Finland, and some work as missionaries in different countries. In addition to the employees major portion of the church members work actively as volunteers in various tasks of the church. The information concerning church services should reach both people involved in planning and implementing the services as well as the audience that come to visit the church. In the beginning of the project the domain of the analysis was not yet very clear. Therefore, a new method was needed to find out the most central documents along with other kinds of information flowing in the domain. The new method consisted of

the user requirements elicitation techniques developed in the RASKE project complemented by a genre-based method, where information flows of the domain are defined in a participatory way with a wall diagram (Article 6). After all the information flows had been defined, the most central of them were selected for more detailed analysis. Again we used a participatory method for defining the preliminary structures for the documents (Article 7). HeTi project is currently in design and implementation phases in its lifecycle. The final implementation is expected to be in testing during 2004.

In the following, the results of the research are described. The Section 3 discussed the life-cycle of contextual metadata from its creation in document analysis to its utilization. The role of structural metadata, in turn, is discussed in Section 4.

3 CONTEXTUAL METADATA LIFE-CYCLE

Metadata concerning the document collections in the enterprises has a lot of potential use during the management of documents' lifecycle. The environment of electronic document management in enterprises consists of activities, actors, systems, and documents (Figure 4). The model has been defined in the RASKE project, and is therefore called the RASKE model for electronic document management environments (Article 1). The documents are produced and used in some activities by actors. Systems assist in managing the documents. In the thesis this model for electronic document management environment is used as a basis for analyzing the contextual metadata needed in enterprises. The model differs from the model of Bearman et al. (1999) in that here the focus is in class level, not in individual instances of documents. Therefore, the information about place or the subject of the documents is not relevant in this context. The role of mandates emphasized by the ISO metadata standard for records management (ISO TC 46/SC 11/WG 1, 2003) could be included in the notion of systems in the RASKE model.

In this section the life-cycle of the contextual metadata is discussed from its creation or gathering in document analysis to its use in user requirements elicitation and information retrieval. Although user requirements elicitation is

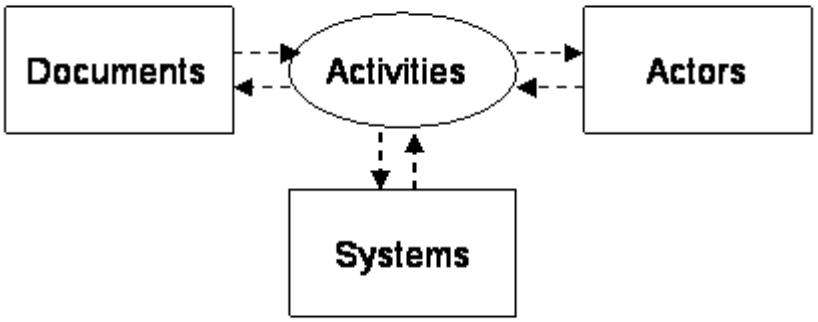

FIGURE 4 The RASKE Model for Electronic Document Management Environments

included in the document analysis it is explained separately, because it both produces and utilizes contextual metadata. The phases of the document analysis are explained in more detail in Salminen (2000) and in the Article 1 included in the thesis. Here the focus in document analysis is placed in describing the aspects related to the contextual metadata that is utilized later.

3.1 Collecting Contextual Metadata in Document Analysis

Document analysis forms a basis of a document standardization process, where the rules for defining the information of the documents are agreed upon (Salminen, 2000). The document analysis process (Figure 5, Article 1) consists of domain definition followed by process modeling, document modeling and role modeling, which can be performed parallel. Process modeling produces descriptions of the activities of a domain, while role modeling defines actors of the domain. During the RASKE project there was no need to model the systems of the domain, since the goal was to develop document standards independent of particular hardware or software. Therefore the modeling of the systems has been left out of the document analysis process as an individual phase. Description of the systems of the domain can be included in the domain specification phase, if needed. The last phase after user needs analysis is collecting the analysis report. In the RASKE project methods for different phases of the analysis

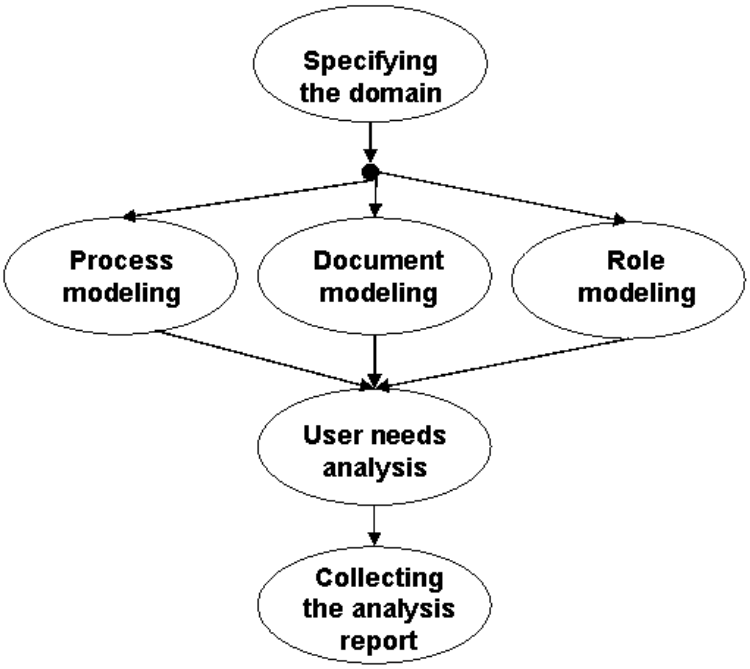

FIGURE 5 Document Analysis Process

have been defined forming the RASKE methodology.

During the modeling phases the contextual metadata concerning the documents of the domain are collected and graphical presentations are created. Document modeling phase produces also structural metadata in a form of preliminary document type definitions defined by a chosen modeling language. The choice of the modeling techniques resulted from the following three principles (Article 1):

1. The models are for human readers for supporting communication and understanding (Aguilar-Savén, 2003). There is no need to show all special cases, instead there has to be space for human judgments in determining the detailed meaning in special cases.
2. The models have to be graphical, as clear as possible, with a few notions and symbols, supporting intuitive understanding.
3. To avoid too many models in one communication situation, process models must allow showing the most important entities in the same model: activities, actors, and documents. Process models are not intended for process automation but to show how work is actually done. Thus they have to be descriptive, not prescriptive. The capability for separating the control and data flow is important to depict the creation and use of documents in business processes. Separate description of data and control flow also allows the use of the same symbols to describe control flow between activities only, or data flow only when needed.

Specifying the domain results an agreement of the activity whose document management should be improved during the analysis process. The agreement can be formulated as a domain definition, which in turn can be used in the role modeling phase.

Role modeling phase produces as a result descriptions of the organizations participating in the activities of the domain (Tiitinen, 2003). The results are first depicted in the organizational model. In Figure 6 there is an example of such a model with the Planning and Communicating Church Services as a domain. The rectangles depict organizations, organizational units, or even individual persons acting in a certain role in the domain. The actors can be grouped, so that their organizational hierarchy can be seen. The domain to be modeled is depicted by a circle in the middle of the graph. The roles of the actors are shown by labeled arrows pointing from the actors to the domain. In order to make the graph clearer to read, the arrows are labeled with identifiers of the roles. Identifiers are then clarified elsewhere. The tasks of the different roles are further defined in textual format during role modeling phase.

Process modeling phase produces as a result a description of the process of creating and using the documents of the domain. The process is defined by a process model with notation borrowed from the Information Control Nets, ICNs (Ellis, 1979). The techniques of the ICNs have been, however, further developed. For example, the role of people, which was not emphasized in original ICNs (Auramäki et al., 1992) has been added to the graphs. An example of a

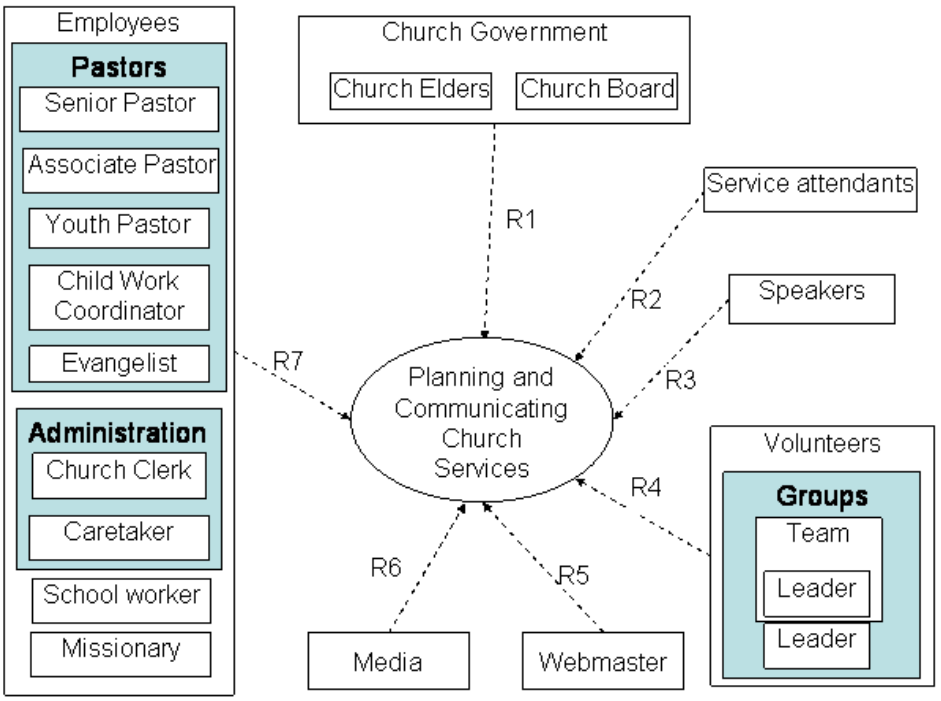

FIGURE 6 Organizational Model

process model depicting output materials to the activities is in Figure 7. Such a process model is called document output model. The example is given from the church domain, and it illustrates the process of planning and informing the church services from strategic planning to announcing weekly services in newspapers. An activity in the process is depicted by an ellipse. The upper part of an ellipse shows actors performing the activity. The solid arrows illustrate the control flow in the process. Because the idea is not to implement an automated workflow system, the starting and finishing times of the activities are not defined accurately. Instead, the control flow only indicates the starting order of the activities. The finishing time of an activity has not been indicated. Hence, there can be many activities performed in parallel even if the graph would seem to suggest successive order for the activities. In addition to the document output model we have used also document input model to illustrate the material needed in the activities. The names of input or output material of an activity are placed on a broken arrow pointing to or from the activity, respectively. The lifecycles of the most important documents can be further defined in document modeling phase by state transition diagrams originating from the OOA methodology (Shlaer & Mellor, 1992).

In document modeling, the document types and their relationships to each other are defined. The results can be illustrated by a document-relationship model. The document-relationship model is derived from the more general en-

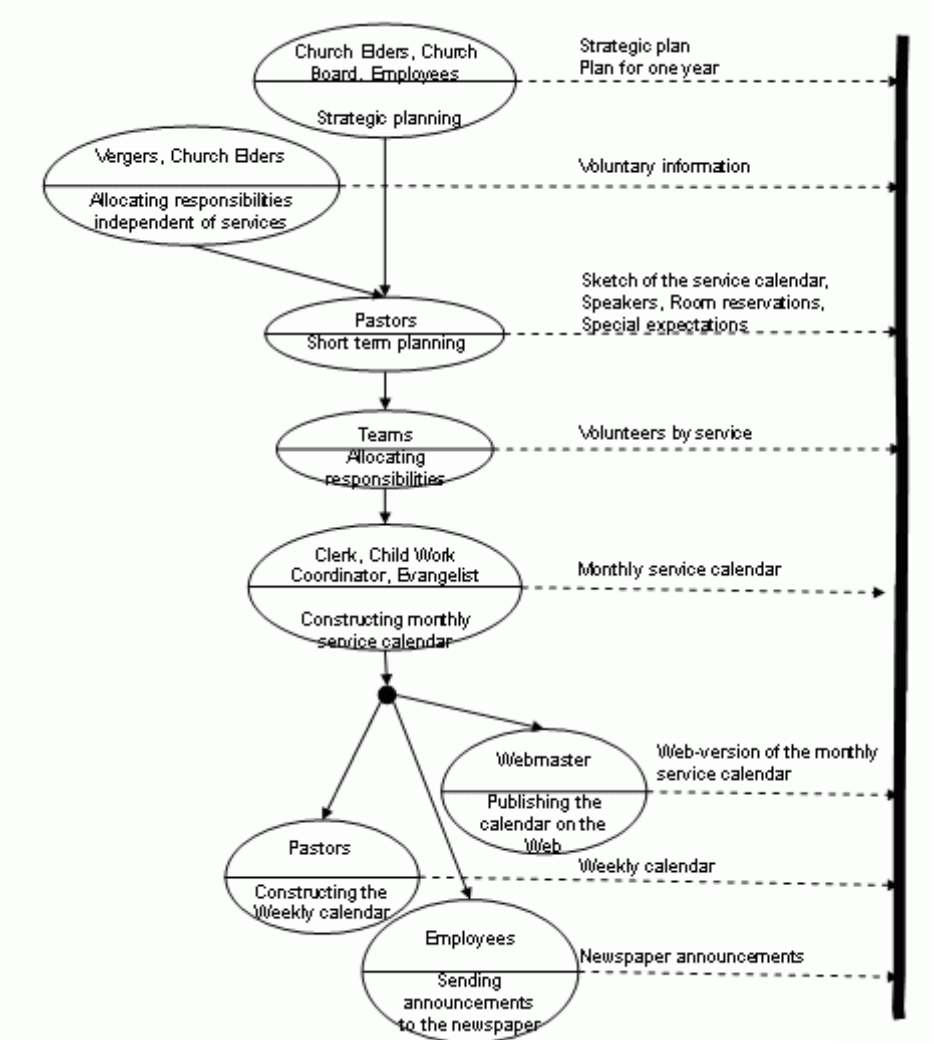

FIGURE 7 Document Output Model

tity-relationship model, also known as ER-model (Chen, 1976). Document modeling is described in more detail in Section 4.1.

3.2 User Needs Analysis and Contextual Metadata

For managing document and document collections in enterprises extensive knowledge about the domain, its processes, documents, and the needs of both organizations and individual people has to be gained. The RASKE methodology includes methods for eliciting the user requirements. The methods have been developed during case studies where the needs concerning electronic document management in public administration were studied (Article 2). The knowledge about the domain can be gathered by many means: discussing informally with the domain experts, studying sample documents, instructions, and documents from previous document management projects, and by semi-structured interviews. The original method of RASKE has been further developed by combining the methodology with a genre-based method (Karjalainen

et al., 2000). The unified method (Figure 8) can be used in cases where the domain definition is quite vague in the beginning, as well as in cases where the domain and its organizational actors are already established (Article 6).

After establishing a steering committee to guide and coordinate the work, the domain of the analysis should be defined. After this, literal sources of the domain can be collected and experts interviewed, so that modelling phase can start. The modelling phase includes the same process modeling, document modeling and role modeling aspects as the original RASKE methodology. After the modeling phase has started, the unified method exploits group sessions

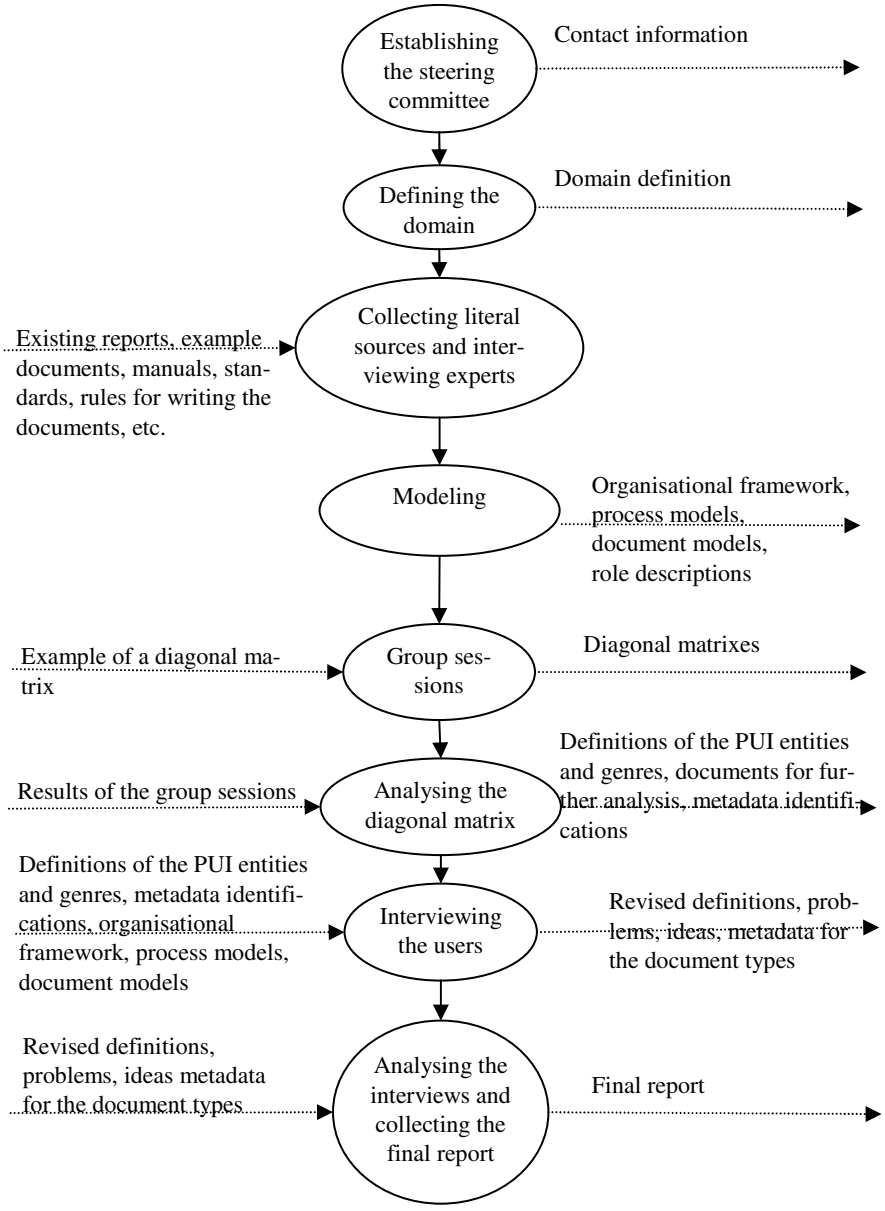

FIGURE 8 Overview of the Requirements Analysis Method

where diagonal matrix technique is used to define the producers and users of the information as well as the information elements flowing between the producers and users. After the group session, representative sample of the producers and users of information are interviewed with semi-structured interviews. The interviewees are oriented to the domain in question by showing them the graphical models concerning the contextual metadata, which were created in preceding phases of the document analysis process. More detailed metadata related to the documents along with problems and development ideas are collected by questions related to the use and processing of the documents. During the interview use cases relating to documents are being identified and defined (Dervin, 1992). In discussing the use cases the interviewee might also reveal some tacit knowledge, because they are inspired to express their tasks in activities in each of the cases (Erdmann & Studer, 1998). The models drawn during the analysis process can also be seen as a means to capture tacit knowledge of the persons operating on a certain domain. In our case organizations the uncomplicated graphs have inspired positive feedback at least because there had not been such a simple models of the domains before.

The genre-based method included in the unified requirements analysis method seems to be useful in cases where the domain is not so well established. The documents to be further analysed can be identified together, and their initial structure can be negotiated in a participatory way. The genre-based method also benefits the participants of the group sessions, because there the users can discuss together the problems and development ideas relating to the doamin. The group sessions, however, increase the time needed for the analysis, because in many cases it is difficult to arrange session times suitable for all participants. It seems that individual interviews with only a few people present at a time are easier to schedule.

3.3 Contextual Metadata in Information Retrieval

The problems of information retrieval were encountered in the EULEGIS project where legal information from different European countries was examined. The user needs study revealed that the heterogeneity of legal systems and information sources is a problem in legal information retrieval. In addition, the users who need the legal information in Europe are quite heterogeneous, too. (Lyytikäinen et al., 2000a)

The solution offered to the information retrieval problem was to collect contextual metadata about the legal systems and to show it to the users in a graphical interface. The users were able to access the information sources without needing to know which repository to connect to. The contextual metadata was visualized by the same graphical models that were earlier used in document analysis and in eliciting user requirements (Article 5). The models enable three different views to the documentation: information source view (a

variant of the document-relationship model), actor view (organizational framework), and process view (document output model).

An example of the graphical user interface is in Figure 9. In the figure, a part of the process of the production of legislative documents in the Finnish legal system is described. The interface can be used in two ways. First, the user can have further information about the elements in the graph, i.e. the activities, actors, or document types, by selecting the desired element. Secondly, the user can search for the documents produced by a selected activity or an actor. In the latter case, the selection leads to a query form helping to formulate a query targeted to the database, which contains desired documents.

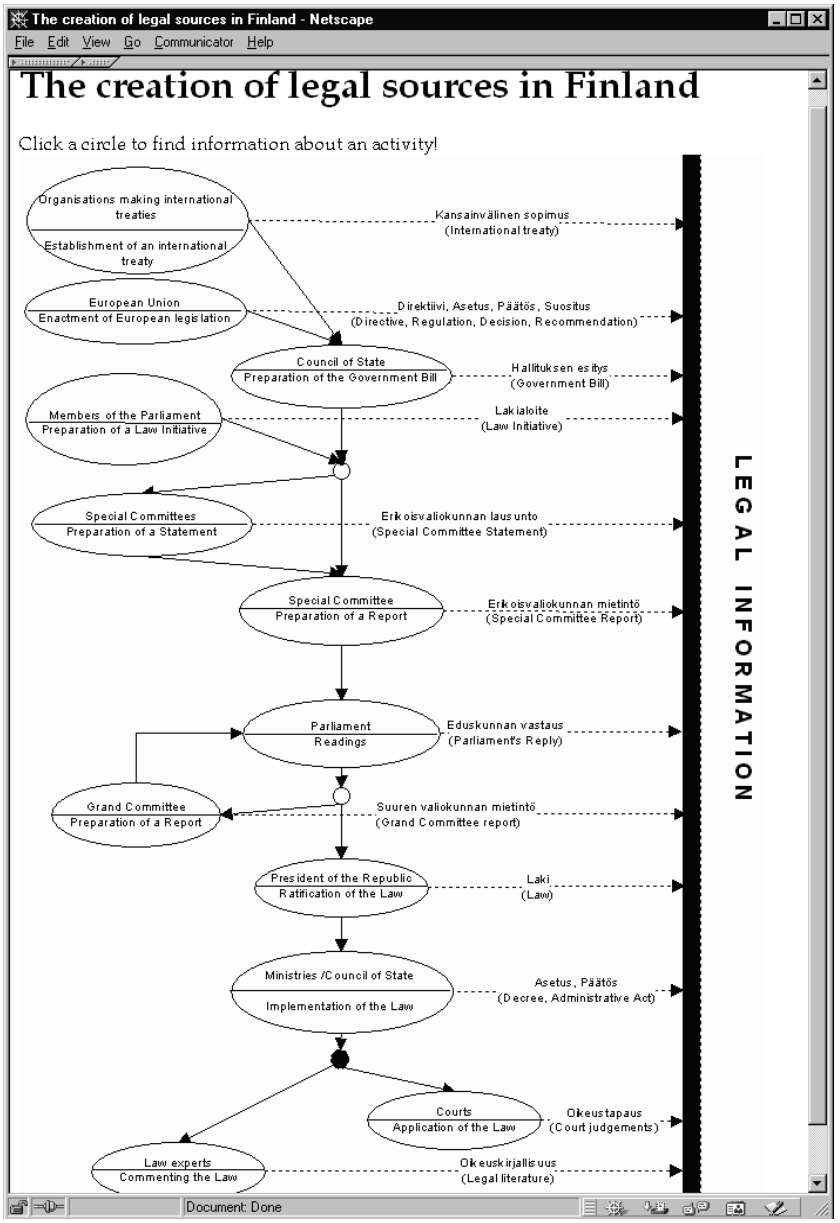

FIGURE 9 A Process View to the Production of Legislative Documents in Finland

3.4 XML Model for the Contextual Metadata

In the EULEGIS project, an XML DTD for describing legal systems was developed (Article 5). The DTD was meant to be used in gathering the metadata about legal systems also from such legal systems that had not been involved in the case analysis. The DTD was also used as a basis for metadata database, which was utilized when users of the EULEGIS prototype wanted to make queries based on contextual metadata. Here, a more generalized version of the DTD is presented. With this DTD, any domain that is of interest from the point of view of contextual metadata, can be described. The Domain DTD in its graphical format, illustrated by Near&Far® Designer 3.0, can be seen in Figure 10. Appendix 1 includes the XML DTD in textual format and also an example of its use.

With the DTD all three views to a domain – actor view, information source view and process view - can be described. In the DTD, the systems interacting with the documents, documents, actors, and processes of certain domain are defined together with their relationships to each other. The DTD is also designed in a way that enables the models to be described in many languages.

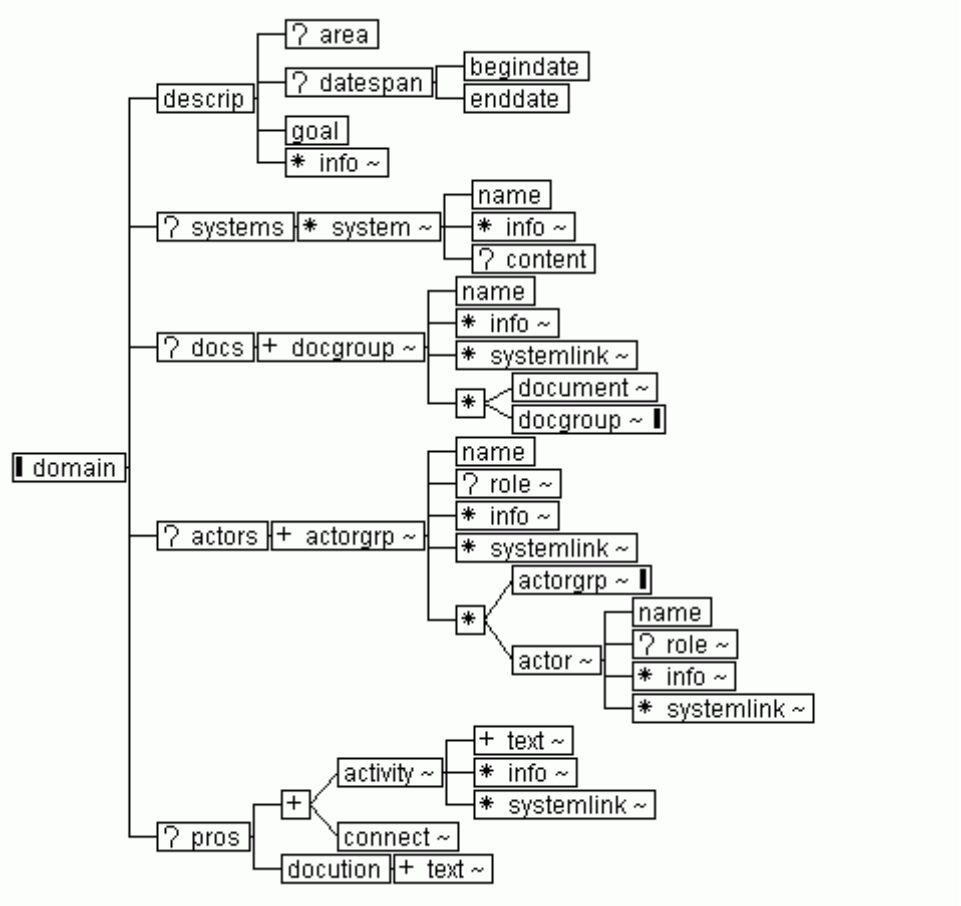

FIGURE 10 Graphical Presentation of an XML DTD for Describing Contextual Metadata

The *descrip* element in the DTD is used for describing the domain in general terms. The domain can cover some geographic *area* (e.g. Finland) and, *datespan*, and its main purpose (e.g. creation of the legal sources in Finland, or planning and communicating church services) is described by a *goal* element. Both area and goal can be described in several languages, defineable by an attribute. With the *info* element more detailed textual description about the domain can be inserted. Inclusion of the datespan element was seen necessary after the EULEGIS implementation, because the user who retrieves documents may be interested in whether the documents of the domain are current ones or originate from the 18th century.

The systems interacting or managing documents in the domain are defined by *systems* element. Each *system* needs to have at least a *name* and some identification, which is defined as an attribute. The more detailed description of the system can reside also elsewhere, and be defined with other DTDs, like for example, SearchDB-ML (Powell & Fox, 1998).

Docs element defines the metadata for document classes in a collection. Documents can form hierarchic groups, which eventually contain the document classes. Links to systems that include the documents belonging to the document class are established via an attribute in *systemlink*.

Actors element defines metadata related to the producers and users of documents in a domain. Similarly to the definition of document classes, actors can form hierarchic actorgroups, which eventually contain the descriptions of actor classes. Systemlink element is used to link the description of an actorgroup or actor to the system including document classes produced by that actor.

Processes, where the documents are produced, are described by element called *pros*. Processes consists of activities and connectors, which can appear in any order. Systemlink element connects the *activity* to the system containing documents produced by that activity. A type of a *connector* can be either 'or' or 'and'. The type is defined by an attribute. *Docution* element denotes the documentation as a whole, which is produced during the process.

The advantages of an XML solution are its capabilities of being source to multiple models in multiple languages. It also enables transformations to be done on the metadata into different presentation formats in case of technological changes in the document management environment. XML files are also easy to transfer in the network environment since XML solutions are system and vendor independent. This also enables long term archiving of the data. Since XML is designed to be used in WWW environment, the metadata described in XML format can be easily used in this global network. With XML DTD the authors of the metadata can be assisted during metadata creation phase. The DTD enables validation of the metadata file, so that it corresponds to the defined structure.

4 STRUCTURAL METADATA IN DOCUMENT MANAGEMENT

This section discusses the role of structured metadata. First, its creation in document modeling phase of the document analysis process is described, and then its utilization in various tasks is discussed.

4.1 Document Modeling

In the RASKE methodology, document modeling includes description of document types, their life-cycles, contents, and relationships to each other. There are three phases in modeling the documents (Salminen, 2000): object modeling, state modeling and content modeling. The execution of the phases can be done in parallel and iteratively.

In the object modeling phase, a set of documents is abstracted as a document object, which will later become a document type due to the formal definition of its structure. Each document object class gets a short description in writing. In addition, the classes and their relationships to each other are modeled in a document-relationship diagram (D-R-diagram), derived from the general entity-relationship model (Chen, 1976). The D-R diagram corresponds also to the information structure diagram of Object-oriented Analysis, OOA (Shlaer & Mellor, 1992). An example of the D-R diagram is shown in Figure 11. The object classes are depicted as rectangles, and their relationships as arrows. The amount of arrowheads symbolized the nature of the relationship: single arrowhead depicts one-to-one relationship; double arrowhead symbolized one-to-many relationship. A letter "C" may be used to indicate a conditional relationship.

The state modeling phase produces descriptions of the dynamic behavior of document objects over time. As a technique for this, the RASKE methodology uses the state transition diagrams of OOA. An example of a state transition diagram for a Monthly service calendar in a church is in Figure 12. The states are

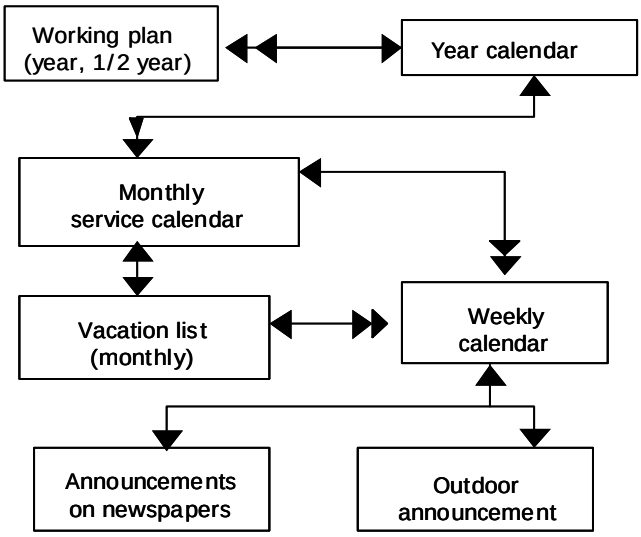

FIGURE 11 A Document-Relationship Diagram

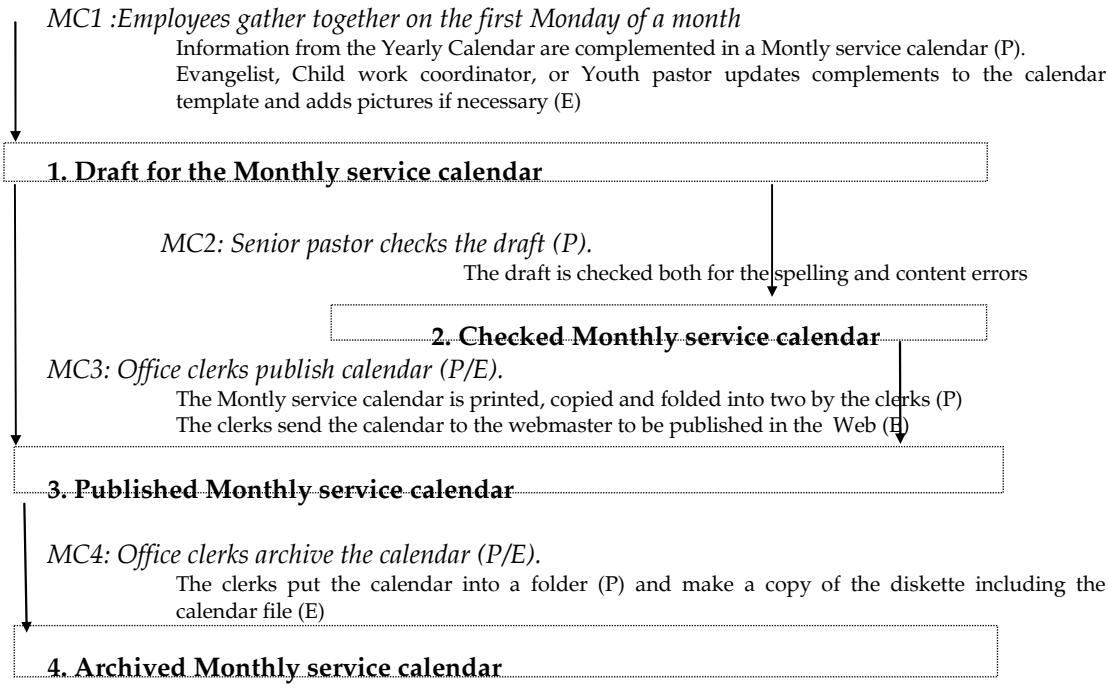

FIGURE 12 A State Transition Diagram for Monthly Service Calendar of a Church

depicted by rectangles, events by texts in italics, actions in normal texts, and transition rules by arrows. The states are numbered and each event has a unique identifier. In such a diagram there are states, events, transition rules, and actions depicted. A state is a situation of a typical instance of the document object in its lifetime. An object can be only in one state at a time. In order to ar-

rive into a state, some action is needed. Each state is associated with one action, but the action can consist of several activities. Events are incidents which cause instances to change state according to transition rules. A transition rule defines which new state follows if a certain event occurs.

In content modeling, the purpose is to specify the hierarchic structure of the documents in the domain (Salminen, 2000). In order to do this, the names for the elements, as well as the order of them in the documents should be defined. Traditionally this can be done by studying existing documents and discussing with domain experts (Maler & El Andaloussi, 1996). A complementary option for this is to define the names of the elements and their order participatory with the domain experts (Article 7). A wall-diagram technique (Saaren-Seppälä, 1997) can be used to place the names of the defined elements in hierarchical order, and enable each participant to see the results of the discussion and comment it. Table 2 illustrates the wall diagram of the topic definition for Monthly service calendar in a church. In the table, the hierarchy of elements is indicated both with their placement on different columns under headings called topics, subtopics, and items.

TABLE 2 A part of the wall-diagram for defining elements in Monthly service calendar

Document type: Monthly service calendar		
Topic	**Subtopic**	**List of items, notes**
Service	Date	Week number Day, time The services may reoccur once a month, on the first Sunday of the month, etc.
	Name	
	Target audience	For example, youths, teenagers, elderly people, women
	Extension	For example Communion, mission service
	Speaker	Title Name There may 0-n speakers per service

After the elements of the document types of the analysis domain have been named, and their hierarchical order defined, the relationships of the elements of different document types can be described in a reuse table (Table 3). In the table elements of the examined document type are placed in the middle column. On the left column there can be information about the source of the element content, for example, names of databases or other document types containing the desired information. On the right column, information about the reuse possibilities of the element content can be seen. The reuse table should not concentrate only to the current practices, but also consider the needs of the future.

SGML or XML offer means for defining more detailed structures for the document types. In order to represent the DTDs or schemas to the users, however, some graphical tools have been used in the RASKE methodology. Such tools are, for example, elm graphs (Maler & El Andaloussi, 1996) or DTD design tools, like Near&Far® Designer.

TABLE 3 A part of a reuse table for Monthly service calendar

SOURCE	ELEMENT	REUSED
Picture from the database	FRONTPAGE	-
Week numbers from general calendar, event information from the yearly calendar	WEEK Weeknumber Event EventName Place Service Date Name Target audience Extension Speaker	Event information is reused in Weekly calendar and Newspaper announcement
Children's event from the calendar of the Child work secretary	EVENTS FOR CHILDREN Event Re-occurring events Remarks	Children's events are reused in Children's calendar

4.2 Utilization of Structural Metadata

The idea of structured documents offers possibilities for various kinds of utilization. Authoring of the documents can be assisted, queries can be targeted to elements in the documents during information retrieval, and documents can be converted and transformed into different presentational formats if their structure is known. In the following, these three advantages are discussed.

4.2.1 Authoring assistance

Controlling the consistency and correctness of documents during their authoring process is enabled due to the utilization of standardized document structures. The authors can have editor applications that are aware of the desired document structure. When using this kind of editor, the software can suggest appropriate elements of the document type to be inserted in the document instance and not allow the existence of elements in places not defined by the structure definition. For example, the software can force the author to type the names of participants immediately after the title of a memo, if that is defined in a structure definition of a memo document type. Examples of such structure-aware editors are FrameMaker+SGML (Adobe, 2004) for documents whose structure is defined by an SGML DTD, or Corel® XMetaL® Author (Corel, 2004) for XML formatted documents.

The authoring environment in the Finnish Parliament is a good example of utilization of structural metadata in document production (Article 4). There, the possible resistance towards new authoring environment and practices was lessened and authoring work was made easier by adding tailored features to the

editor software. Examples of such features were automatic numeration of lists and connections to the databases of the Parliament. The connections enabled automatic inclusion of information concerning the Members of the Parliament.

Another possibility for assisting the authoring of structured documents is to use forms in entering the information. The authors need not to know anything about the structure of the document, or even that the thing they are authoring is stored as document. Instead, they can think of filling a form with predefined fields. The form can then be stored as a structured document with the content that the author entered. An example of such a form for entering information about the events in a church is in Figure 13.

4.2.2 Information retrieval

Structural metadata increases possibilities for information retrieval due to the possibility to use the elements of the document in limiting the target of a query (Kuikka & Salminen, 1997; Salminen & Tompa, 1999). For example, a query could be targeted to concern only the contents of title elements of research papers in a scientific archive. The utilization of this kind of structured query requires at least some level of acquaintance of the structure of the documents (Article 3).

In the Finnish Parliament, the consequences of document standardization effort included also improved information retrieval possibilities. Before the standardization, only limited options were available for users in targeting their queries, but now, there exist individual search forms for every document type

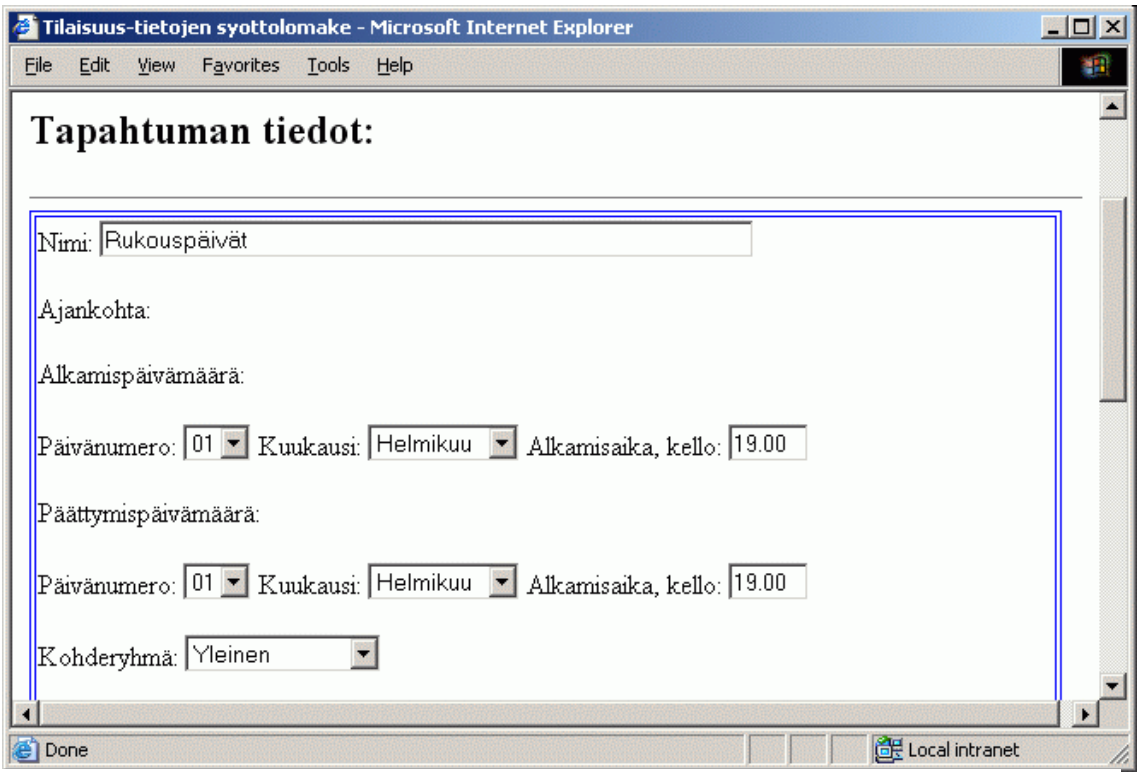

FIGURE 13 A Form for Entering Information About One Event in a Church

enabling the users to use various elements in the documents as a limiting factor in their query. For example, in the case of Committee Statements and Reports, a query can be targeted to such statements and reports only that include protests, are produced during certain year with certain member of the committee participating the preparation of the document.

4.2.3 Document transformations

The existence of structural metadata enables processing of the documents by computers. One document instance can have multiple layouts, parts of the document can be hidden from certain user groups, or multiple documents can be integrated and treated as one new document. Thus, information can be reused reducing the effort needed for multi-channel publishing. Tools and techniques for these kinds of transformations are, for example, Document Style Semantics and Specification Language, DSSSL (ISO-10179, 1996) for SGML formatted documents, XSL Transformations XSLT (Clark, 1999) and Document Object Model, DOM (Apparao et al., 1998) for XML documents. More detailed descriptions of transformations for structured documents can be found, for example, in the studies of Kuikka (1996) and Heikkinen (2000).

Document structures are effectively used for information reuse purposes in the Finnish Parliament. The Parliament's web site (www.eduskunta.fi) includes parliamentary documents in HTML, PDF, and SGML formats, each of which is produced from the same source file. Also, in the Parliament, different kinds of layouts of the parliamentary documents are used in different phases of the document production, for example, in drafting phase the document is printed with bigger line space compared to the finished document. Document transformations enable also the archive of consolidated law, where the legislative documents are retrievable so that the law can be seen as one document regardless of the amount of amendments made during the years.

5 OVERVIEW OF THE INCLUDED ARTICLES

This section discusses the included articles with respect to their role and contribution for the dissertation study. For each of the articles the research objectives and methods are described, and then the content and results both generally and in respect to contextual and structural metadata are briefly discussed.

5.1 Article 1: "Putting Documents into their Work Context in Document Analysis"

Salminen, A., Lyytikäinen, V., & Tiitinen, P. 2000. Putting Documents into their Work Context in Document analysis. *Information Processing & Management* 36 (4), 623-641.

5.1.1 Research objectives and methods

The paper discusses document management from the process modeling point of view. The main research question is what kind of techniques should be used to model work context in document analysis. The process models form a part of the contextual metadata collected during the document analysis process. The paper is a result of constructive research conducted in the RASKE project. Along with a constructive part, where selected process and life cycle modeling techniques are implemented as part of the document analysis method, the paper gives theoretical reasoning for evaluating appropriateness of different modeling techniques for describing the work context in document analysis.

5.1.2 Content and results

The paper discusses the requirements of document standardization concerning the modeling of work with standardized documents. For the basis of the standardization work, the RASKE model (Figure 4) for document management environments is presented. As a result of the study a variant of Information Con-

trol Nets (ICNs, Ellis, 1979) is introduced as one technique to be used to model processes in document analysis. Various forms of the process models are introduced. Document output model describes the productions of various document types in a work process. Document input model instead can be used to show the material needed in the activities of the process. The two models can also be combined and shown as a one graph. Another technique described in the paper, state transition diagrams of the Object-Oriented Analysis method (Shlaer & Mellor, 1992), can be used to capture the life cycle of a document type object and actor roles in the life cycle.

5.2 Article 2: "User Needs for Electronic Document Management in Public Administration: A Study of Two Cases"

Tiitinen, P., Lyytikäinen, V., Päivärinta, T., & Salminen, A. 2000. User Needs for Electronic Document Management in Public Administration: A Study of Two Cases. In H.R. Hansen, M. Bichler, & H. Mahrer (Eds.), *Proceedings of ECIS 2000, European Conference on Information Systems, Volume 2,* Wien: Wirtschaftsuniversität Wien, 1144-1151.

5.2.1 Research objectives and methods

The paper describes a study of two cases. In the cases we studied the needs of people concerning electronic document management and how SGML standardization addresses those needs. In addition, the studies addressed the question of how to elicit user needs for electronic document management in e-Government. Data for answering the research questions were gathered by various means: by informal discussions with the project intermediaries and domain experts, by studying documentation of the domain, and by interviewing the users of the domain in semi-structured interviews. The data gathered was analyzed by revising the models of the domain drawn after initial acquaintance to the domain, and using the RASKE model for document management environments (Figure 4) as an analysis framework. The utilization of the model leads to grouping the data by user roles into the needs concerning documents, information technologies, and work activities related to documents.

5.2.2 Content and results

The paper describes techniques used in eliciting requirements in the Finnish e-Government environment. The first case concerns the creation of the state budget in Finland, the second case Finnish participation in EU legislative work. In both of the cases the study revealed many needs related to different aspects of electronic document management: documents, information technology, and work with documents. From contextual metadata point of view the paper shows, how to exploit models describing the domain context in user interviews.

As a result of the case study a method for eliciting user needs concerning document management practices was developed. The method exploits semi-structured interviews with use cases along with contextual metadata in the form of process models and organizational framework.

5.3 Article 3: "Usability Evaluation of a Structured Document Archive"

Salminen, A., Tiitinen, P., & Lyytikäinen, V. 1999. Usability Evaluation of a Structured Document Archive. In R.H. Sprague, Jr. (Ed.), *Proceedings of the Thirty-Second Hawaii International Conference on System Sciences* (file ddhfu06.pdf at CD-ROM). Los Alamitos, CA: IEEE Computer Society.

5.3.1 Research objectives and methods

Suitable criteria for evaluation are needed when usability inspections are conducted. However, for evaluating specific features of document databases containing structured document there had been no predefined criteria earlier. The research objective in this paper was to develop an evaluation framework which takes into account the specificity of structured documents in databases. The framework was tested in a usability evaluation of a prototype archive implemented by a publishing house in the RASKE project. Two researchers, who were familiar with the application domain, carried out the evaluation at first individually. Later, they discussed their findings and produced a common report. The report includes the description of the archive and the usability problems found during the evaluation. Both the description of the archive and the usability problems were reported according to the evaluation framework. The evaluated archive contained three kinds of documents from the Finnish Parliament and ministries in SGML format.

5.3.2 Content and results

The evaluation framework, which we also call as the usability inspection method, was based on an earlier design-oriented evaluation method (Garzotto et al., 1995) used for general hypermedia applications. In addition, we utilized a grammar-based layered hypermedia model, which has been defined by Salminen and Watters (1992), and Salminen et al. (1995). The developed framework includes four dimensions that are used in the evaluation: content and hierarchic structure, hypertext structure, dynamics, and presentation. The dimensions include evaluation objects, such as document structures, linking structures, and functions search capabilities or printing possibilities. The dimensions can be evaluated with six criteria: richness, reuse, consistency, ease, self-evidence, and suitability. The implications of the study give basis for the future design, implementation and evaluation of structured document database applications.

From structured metadata point of view, the study stresses the importance of informing the users of the document structures. Without this knowledge it is not possible to utilize the advanced capabilities of structured search functions.

5.4 Article 4: "Experiences of SGML Standardization: The Case of the Finnish Legislative Documents"

Salminen, A., Lyytikäinen, V., Tiitinen, P., & Mustajärvi, O. 2001. Experiences of SGML Standardization: The Case of the Finnish Legislative Documents. In R.H. Sprague, Jr. (Ed.), *Proceedings of the Thirty-Fourth Hawaii International Conference on System Sciences* (file etegv01.pdf at CD-ROM). Los Alamitos, CA: IEEE Computer Society.

5.4.1 Research objectives and methods

The paper introduces a case study of SGML standardization in the Finnish Parliament. The main idea of the study was to find out the consequences of SGML standardization effort and to learn from the experiences gained during the standardization process. The data for the study was collected from the different literal sources, like earlier published articles and reports of the standardization process, and by interviewing people involved in the process and now utilizing the SGML-based standards. The analysis of the data gathered was based on the RASKE model of a document management environment (Figure 4).

5.4.2 Content and results

The paper first describes the situation in the beginning of the document standardization process in early 1990's and reveals the reasoning why such a process was started. Then, the standardization process is described and lastly the impacts of the process and use of the structured documents are discussed. The impacts concerned documents, technologies, work with documents, organizations, and the whole society. A lesson learned was that the SGML implementation is a tedious task. In an e-Government environment where several organizations work together, inter-organizational co-operation from the early phases of the project is needed. The standardization affects not only the documents themselves but also the work of the people dealing with them. Motivating the needs for changes and demonstrating future benefits is extremely important. The case also revealed the importance of the models as means to describe the context where the documents are produced and used. The models were also a means to gather tacit knowledge into more tangible form.

5.5 Article 5: "Visualizing Legal Systems for Information Retrieval"

Lyytikäinen, V., Tiitinen, P., Salminen, A., Mercier, L., & Vidick, J.-L. 2000. Visualizing Legal Systems for Information Retrieval. In M. Khosrowpour (Ed.) *Challenges of Information Technology Management in the 21st Century*, *Proceedings of 2000 Information Resources Management Association International Conference*, Hersley, PA: Idea Group Publishing, 245-249.

5.5.1 Research objectives and methods

The research presented in the paper concerns the development of e-Government at European level. From this thesis point of view the paper seeks answers to the questions of how to utilize the contextual metadata in legal information retrieval, and how to visualize the contextual metadata related to legal documents for the users. The goal in the work is to help the citizens and people working in enterprises around Europe to find the information they need even if it is produced in a foreign legal system. The problem was approached by constructive research method, and as a result two constructs were built: a method for visualizing contextual metadata, and a prototype that demonstrated the idea. Motivation for this research was gained from the user needs study performed in the EULEGIS project. In the study semi-structured interviews and questionnaires were used in order to find out the needs of various user groups of legal information.

5.5.2 Content and results

The paper discusses the problems related to information retrieval, encountered by the users of European legal information. The need for legal information from foreign countries and various levels of legislation in Europe is greater than ever before. Due to the increase in legal information repositories on the Internet, this information is also widely available in digital form. The information, however, is scattered in numerous databases where documents are structured, organized and classified in different ways. These differences are related to differences in legal systems. The retrieval and utilization of European legal documents entails knowledge of these legal systems. The paper introduces a method for supporting legal information retrieval by graphical data models showing the documents in the context of their legal system in the user interface. The visualization of the contextual metadata is intended to help users of European legal information to cope with the complexity of the legal domain in Europe, to better understand the differences in legal systems, and to better locate information from correct sources. The results of using the method were demonstrated by a prototype system.

5.6 Article 6: "Analysing Requirements for Content Management"

Lyytikäinen, V. 2003. Analysing Requirements for Content Management. In O. Camp, J. Filipe, S. Hammoudi & M. Piattini (Eds.), *Proceedings of the 5th International Conference on Enterprise Information Systems, Angers, France, April 23-26, 2003* (Vol. 3), Portugal: Escola Superior de Tecnologia do Instituto Politécnico de Setúbal, 104-111. Also published in O. Camp, J. Filipe, S. Hammoudi & M. Piattini (Eds.), Enterprise Information Systems V. Kluwer Academic Publishers B.V.

5.6.1 Research objectives and methods

The paper addresses the question of how to elicit user requirements in developing content management environment in an organization. *Content management* in the article refers to management of both textual and multimedia objects. The data for the study was gathered by an action research in a church environment. The church wanted to improve their planning and informing of services, therefore the first thing was to get acquainted with the current situation and the needs of the people involved in the domain. The analysis domain was not very clear at the beginning, so the user requirements analysis method developed earlier in the RASKE project needed refining. In order to get data to be analyzed informal discussions with domain experts were arranged, exemplar documentation were studied, and group sessions were organized. The participants in the group sessions were people actively involved in producing and using the documentation of the domain. They defined co-operatively the producers and users of information and the information flow between them. More detailed information about documents and tasks related to them were gathered via semi-structured interviews. The gathered data was analyzed and reported according to the RASKE methodology.

5.6.2 Content and results

The paper first discusses the requirements the content management environment places to the user requirements analysis method. Then the suggested method is introduced, and its use in the case organization is described. The paper closes with the lessons learned and implications. The user requirements analysis method developed in the RASKE project was complemented by the use of group sessions, where a diagonal matrix was constructed in a participatory way. The diagonal matrix employed a genre-based method, so that the communicators of genres of the domain were first identified and placed on the diagonal of the matrix on the wall. Then the information flowing between the communicators was identified. The content of the most important genres was handled in the case as documents. The documents were later discussed in detail in semi-structured interviews, where contextual metadata in the form of graphical

models were shown to the users as a tool to focus the interviews. The use of the genre-based method in the middle of the original user requirements analysis method resulted iteration in requirements elicitation, which was considered valuable by both the participants of the group sessions and the analysts. It also created common understanding of the purpose of the new content management system.

5.7 Article 7: "Operationalizing a Genre-Based Method for Content Analysis: A Case of a Church"

Honkaranta, A., & Lyytikäinen, V. 2003. Operationalizing a Genre-Based Method for Content Analysis: A Case of a Church. In W. Abramowicz & G. Klein (Eds.), *Proceedings of the 6th International Conference on Business Information Systems. Software Engineering Track. Colorado Springs, 4-6 June, 2003*. Poland: Department of Management Information Systems at the Poznan University of Economics, 108-116.

5.7.1 Research objectives and methods

The paper considers the adoption of the genre-based, participatory method for content analysis (Karjalainen & Salminen, 2000) for content analysis in a case of a church. *Content analysis* in the case refers to defining schemas and rules for assemblies for documents consisting of many kinds of content units, which may origin from different sources. The sources may be, for example, databases or other documents. The adopted method aids in defining names for content elements of documents, and thus produces structural metadata for document collections. The paper discusses how the method was elaborated in the study, and summarizes the findings with respect to the method, techniques used, and to the theory of genres. The study was a continuum for the requirements analysis described in the Article 6. The study was an action research of its nature, where two researchers, including the author, elaborated the genre-based content analysis method. The method differed from the original one in the sense that in this case it was possible to focus to the selected documents in the beginning since the domain definition was already accomplished in the previous analysis process. We also arranged only one workshop where we collected the contextual metadata in the form of 5W1H framework (Orlikowski & Yates, 1998; Yoshioka et al., 2001) instead of collecting a portion of metadata values in many workshops.

5.7.2 Content and results

The paper begins with reasoning for a participatory, genre-based content analysis and design method. Such a method is then described and its use in a case of

a church is discussed. The implications of the case suggest that such a participatory method is useful in a sense that it offers simple wall-diagram models and tables, with which users and analysts can together define the names for the content units of documents. Simplicity is important for the users, since they may not necessarily be acquainted with the ideas of structured documents in advance. In addition, the participatory method helped in bridging the vocabulary gap between analysts and users in the group sessions. The group sessions were a valuable means for the users to negotiate and decide over the content unit names and hierarchy, since the structures of the documents were not totally stabilized before the sessions. Contradictory proposals for a future system could be negotiated, and agreed upon in workshops. Furthermore the joint work of the users and analysts seemed to reduce the fear against possible changes in work practices. From structured metadata point of view, the method introduced in the paper offers a tool for defining preliminary structures for documents together with their users. The expertise of the analysts is still very important in defining the final DTDs.

5.8 About the Joint Articles

The joint articles of the thesis were produced as a result of close co-operation in our document management group at the University of Jyväskylä. Articles from 1 to 4 were produced in the RASKE project research group, Article 5 describes results of the EULEGIS project group, and Article 7 is a joint work of researchers in the HeTi project.

Airi Salminen was the main author of the first article. Pasi Tiitinen and I had a less role in the writing process, but the process modeling aspects of the modeling methodology were included in my responsibilities in the RASKE project.

For the Article 2, where Pasi Tiitinen was the main author, I contributed almost equally in the writing process. Airi Salminen and Tero Päivärinta gave also their valuable contributions as authors. My responsibility in conducting the research described in the paper concerned the case of creation of the Finnish state budget. I was also very much involved in the other case of the paper, the case describing document management needed in the Finnish participation in the EU legislative work. The case was mainly in Pasi Tiitinen's responsibility. The requirements elicitation methods reported in the paper were developed together with other authors.

The Article 3 was done cooperatively with Airi Salminen and Pasi Tiitinen, Airi Salminen being the main author. The usability framework described in the paper was produced cooperatively by all authors. I also contributed in testing the framework with Pasi Tiitinen by evaluating a prototype application containing legislative documents. Describing the document type definitions of the documents in the evaluated archive was mainly my task.

In the Article 4, where Airi Salminen was the main author, I, Pasi Tiitinen and Olli Mustajärvi from the Finnish Parliament contributed roughly equally as co-authors. In the actual study, I and Pasi Tiitinen shared the responsibility of planning the data gathering. We also interviewed the users in the Finnish Parliament together in order to find out the effects of the standardization work.

In the Article 5, the main responsibility of the authorship was shared between me and Pasi Tiitinen, while Airi Salminen had a slightly smaller contribution. Laurent Mercier and Jean-Luc Vidick contributed mainly in describing legal systems in general. The first three authors contributed equally in developing the visualizations concerning legislative information sources, actors and processes, although in the project this area was in my and Pasi Tiitinen's responsibility. Experts of the European legislation both inside and outside the EULEGIS project group gave important contribution to the development of the visualizations by giving their opinions on our suggestions. I was responsible for the development of the XML DTD for the views.

In writing the Article 7, both of the two co-authors were responsible with an equal share. Method 2, which was originally developed by Anne Honkaranta was adjusted in the study together by the authors. The DTDs produced as a result of the study were developed by me.

6 CONCLUSION AND FURTHER RESEARCH

This section gives conclusion of the research by first revisiting the research questions stated earlier. Then the contributions of the thesis are summarized briefly. The last subsection introduces some avenues for further research based on the studies described in the thesis.

6.1 Research Questions Revisited

In section 2 there were three research questions defined. The questions were:

1. How to collect and create contextual and structural metadata?
2. How to utilize contextual metadata in document analysis and document management?
3. How to visualize contextual metadata for users?

In the following, the research of the thesis is summarized according to the research questions stated.

How to collect and create contextual and structural metadata?

The attributes characterizing metadata according to Gilliland-Swetland (1998) describe the conditions where collection or creation of metadata takes place. The attributes are the following: method of metadata creation, source, nature, status, structure, semantics, and level. The attributes are used here to summarize the creation of contextual and structural metadata.

The contextual and structural metadata described in the thesis is created manually during the document analysis process. After the analysts have got acquainted with the literal sources and have discussed with the domain experts in order to form a comprehension of the domain, the modeling phase may begin. The modeling phase includes drafting of graphs describing the relationships of

the document types of the domain, actors involved in the domain, as well as processes producing the document types.

The source of the contextual and structural metadata is external. Metadata described here is typically created separately from the creation of the documents in the document collection, by experts of the domain. In Murphy's (1998) terms, the metadata is thus both physically and logically separable. The status of the metadata is static and long-term, since the metadata is not supposed to change very often. The metadata is also structured, since the predefined XML DTD can be used to define the content of metadata elements. Since there is no controlled vocabulary for the content of metadata elements, it can be said that the metadata here is semantically uncontrolled. The level of metadata is metadata for document collections.

How to utilize contextual metadata in document analysis and document management?

The contextual metadata is both created and utilized during document analysis. Document analysis includes user needs analysis, where the models containing contextual metadata are shown to the users during interview sessions. The models help in focusing the interviewees to the domain in question. Besides, the interviews offer possibilities to check and correct the models in case of misunderstandings.

In document management, the utilization of the contextual metadata offers new possibilities for information retrieval. If the graphical models containing contextual metadata are transformed into user interfaces, the users can query the databases containing the documents they need without having the knowledge of the structure and actual content of the database. This approach is demonstrated in the environment of European legal documents, where the amount of databases available for the users has increased and the heterogeneity of legal systems and databases causes problems in information retrieval.

How to visualize contextual metadata for users?

The principles for the choice of visualizations for contextual metadata are the same that were used for selecting modeling techniques in document analysis. The purpose for visualizing contextual metadata is to offer information for human users for supporting the understanding of the domain. In addition, the retrieval of documents can be supported as well. In our cases there has been no need to show all special cases in the models, but instead we have left space for human judgments in details. The models should be as clear and intuitive as possible, with few symbols and notions.

The visualizations of contextual metadata were tested by a prototype in the EULEGIS project. The feedback from the users implied that the chosen models were simple enough, yet contained the information needed to cope with the complexity of the European legal domain.

6.2 Contributions of the Thesis

The research described in the thesis contributes in introducing ways of collecting and using structural and contextual metadata. The methods developed during the research include

- document analysis method, especially techniques for the process modeling phase,
- method for participatory design of document structures,
- method for user requirements elicitation,
- usability evaluation method for structured document archives, and
- method for visualizing contextual metadata.

In addition to the above methods, a model for the structure of the contextual metadata was also developed during the research. The XML DTD was used in the EULEGIS project in defining contextual metadata for different legal systems.

The methods were developed and tested in different case environments. The cases differ in many ways, for example, in respect to the business areas of the organizations involved, their size, and maturity in the use of information technology. The organization of the first two cases, the case of the Finnish Parliament and ministries, and the EULEGIS, operated on a public domain. Their purpose is to serve citizens and the employees by offering them the information they need in the most suitable format. The use of the information technology has long traditions both in Finnish public organizations and in other European countries involved in the EULEGIS. The third case organization, the church, was a small third sector enterprise, where the goal was to improve communication and planning of services. The church has only a short history of utilizing computers in the work of the employees.

The experiences of the cases seem to suggest that the developed methods are suitable in many kinds of enterprises. The size of the enterprise may have effects on the amount of data to be gathered and interviews to be made during the document analysis. However, the same methods can be used. Also, there did not seem to be any difference, whether the enterprise was a public one, or belonged to the third sector. Since none of the enterprises in the cases represented private sector, implications of the applicability of the methods can be drawn from the case studies. The methods themselves, however, do not have anything particular that would tie them to the public or third sector enterprises and prevent their use in private organizations.

The results of the thesis could be utilized in organizations willing to develop their document management practices. The benefits of the document analysis and use of structured documents, however, increases if the amount of data to be managed through documents grows. The contextual metadata, in turn, could be useful in such cases where there is need to know the process and

actors producing documents. An example of such could be a manufacturing company with many subcontractors. The process of producing a complicated artifact with a user manual includes a lot of information stored in documents. The visualization of contextual metadata could serve as an interface to locate the documents in an intranet between the manufacturer and its subcontractors.

6.3 Avenues for Further Research

The studies described in the thesis open possibilities for further research in the future. The methods developed should be tested in different kinds of environments, for example, in industrial settings, where complex artifacts are manufactured in inter-organizational network. Also, situations, where the content to be analyzed and retrieved is not in the form of documents but, for instance, reside in data warehouses, offer an interesting avenue for future research.

The definition of contextual metadata used in the thesis is not covering all kinds of possible needs people might have for the context of documents. The scope of the context could be widened into many directions. In security sensitive environments, for example, the needs to know about the security issues might be important. Also, for the preservation and ensuring long-term accessibility to the document collections, more technical issues could be included in the contextual metadata definition. For example, according to Jokela (2001) information about the format of the documents and possible dependencies of some hardware or software could be mentioned.

One avenue for broadening the scope of the contextual metadata DTD could be investigating the upcoming XML DTD called Encoded Archival Context (EAC). The DTD describes record creators, like persons, organizations, and families, with their names, essential functions, activities, and characteristics, as well as dates and places they were active (Pitti, 2003). Particularly interesting research area would be to include to the contextual metadata DTD the possibility to indicate the time when a specific actor had a particular role or relationships with the document type the metadata is related to.

In the research of the thesis the focus has been in collections of documents. If the context is known, then the user is able to query for the individual documents of the collection. One avenue for further research could be to examine the issue from the individual documents point of view, how the information about the context of each document instance could be created, preserved and used. In this case the contextual metadata should cover instances of processes and actors, not just classes.

Another avenue for further research would be automatic transformation of the XML formatted contextual metadata into graphical user interfaces. Currently, the contextual metadata is only stored in XML format according to the DTD defined. This enables querying the possible metadata database. However, the visual interfaces are produced manually.

The emergence of different kinds of schemas in addition to the DTDs for defining document structures creates also pressures for developing more sophisticated methods for document analysis. The capabilities to define different data types, for instance, adds possibilities to create more complex schemas to suit better for new kinds of uses, like e-business.

YHTEENVETO (FINNISH SUMMARY)

Useissa organisaatioissa dokumenteilla on varsin keskeinen rooli. Tiedon määrän lisääntyessä tarvitaan uudenlaisia menetelmiä dokumenttien hallintaan. Organisaation dokumenttien hallinta käsittää dokumenttien koko elinkaaren niiden luomisesta tai vastaanottamisesta tuhoamiseen saakka. Lisäksi organisaation dokumenttien hallintaan kuuluu dokumenttien hallinnan ratkaisujen kehittäminen. Dokumenttien hallinnassa metatiedolla on tärkeä merkitys. Metatiedolla tarkoitetaan dokumentteja kuvaavaa tietoa. Metatietoa käytetään dokumentteja haettaessa tärkeiden dokumenttien erottamiseen vähemmän tärkeistä ja jopa turhista. Aiempi dokumenttien metatiedon tutkimus on pääosin keskittynyt yksittäisten dokumentti-ilmentymien metatiedon kuvaamiseen ja määrittelyyn. Tässä väitöskirjassa tutkimuksen kohteena on dokumenttikokoelmiin liittyvä metatieto. Erityisesti väitöskirjassa tarkastellaan dokumenttikokoelman kontekstia. Kontekstilla tässä yhteydessä tarkoitetaan dokumenttien tuottamisprosesseja, tuottaja- ja käyttäjärooleja sekä dokumenttien keskinäisiä suhteita. Lisäksi huomiota kiinnitetään dokumenttien loogisen rakenteen kuvaamiseen, joka tuottaa metatietoa dokumenttien rakenteista. Väitöskirjassa kuvataan menetelmiä ja tekniikkoja, joilla kontekstiin ja rakenteeseen liittyvää metatietoa voidaan kerätä ja hyödyntää.

Menetelmiä ja tekniikkoja testattiin kolmessa projektissa erilaisissa tapaustutkimuksissa. Tapaustutkimukset toteutettiin eduskunnassa ja ministeriöissä, eurooppalaisessa lakitietokantojen yhtenäistämishankkeessa sekä keskisuomalaisessa helluntaiseurakunnassa. Tutkimuksen tuloksena osoitetaan, kuinka konteksti- ja rakennemetatietoa voidaan kerätä ja hyödyntää. Metatietoa kerätään dokumenttianalyysissä, ja sitä hyödynnetään analyysiin kuuluvassa käyttäjätarpeiden määrittelyssä sekä myöhemmin dokumentteja haettaessa. Väitöskirjassa esitetään myös esimerkinomainen rakennemäärittely kontekstimetatiedolle. Rakennemäärittely on laadittu XML-kielen dokumenttityyppimäärittelyn avulla. Tulevaisuudessa tutkimusta voidaan laajentaa esimerkiksi lisäämällä kontekstia kuvaavaan metatietoon tietoturvaan liittyviä määritteitä, kuten tietosuojaan ja tiedon saatavuuteen liittyviä rajoitteita.

REFERENCES

Adobe. 2004. FrameMaker+SGML. Adobe. http://www.adobe.com/products/framemaker/prodinfosgml.html. [January 30, 2004]

Aguilar-Savén, R.S. 2003. Business process modelling: Review and framework. International Journal of Production Economics In press.

Ahonen, H. 1996. Automatic generation of SGML content models. In Electronic Publishing '96. Palo Alto, California, USA.

Apparao, V., Byrne, S., Champion, M., Isaacs, S., Jacobs, I., Hors, A.L., Nicol, G., Robie, J., Sutor, R., Wilson, C. & Wood, L. 1998. Document Object Model (DOM) Level 1 Specification. W3C Recommendation 1 October, 1998. http://www.w3.org/TR/1998/REC-DOM-Level-1-19981001/. [February 2, 2004]

Auramäki, E., Hirscheim, R. & Lyytinen, K. 1992. Modelling offices through discourse analysis: a comparison and evaluation of SAMPO with OSSAD and ICN. The Computer Journal 35 (5), 342-352.

Baker, T. 1998. Languages for Dublin Core. D-Lib Magazine. http://www.dlib.org/dlib/december98/12baker.html. [January 13, 2004]

Barry, R.E. 1993. Electronic Document and Records Management Systems: Towards a Methodology for Requirements Definition. In Document & Management 93, Proceedings of The Document Management Conference. London.

Bearman, D. & Sochats, K. 1996. Metadata requirements for evidence. http://www.archimuse.com/papers/nhprc/BACartic.html. [January 14, 2004]

Bearman, D. 1999. Reality and chimeras in the preservation of electronic records. D-Lib Magazine. http://www.dlib.org/dlib/april99/bearman/04bearman.html. [January 14, 2004]

Bearman, D., Miller, E., Rust, G., Trant, J. & Weibel, S. 1999. A common model to support interoperable metadata. D-Lib Magazine. http://www.dlib.org/dlib/january99/bearman/01bearman.html. [January 12, 2004]

Berners-Lee, T., Fielding, R., Irvine, U.C. & Masinter, L. 1998. Uniform resource Identifiers (URI): Generic Syntax. W3C. http://www.ietf.org/rfc/rfc2396.txt. [January 15, 2004]

Blair, D.C. 2002. The challenge of commercial document retrieval, Part I: Major issues, and a framework based on search exhaustivity, determinacy of representation and document collection size. Information Processing & Management 38 (3), 273-291.

Boiko, B. 2002. Content Management Bible. New York, U.S.A: Hungry Minds, Inc.

Boll, S., Klas, W. & Sheth, A. 1998. Overview on using metadata to manage multimedia data. In A. Sheth & W. Klas (Eds.). Multimedia Data Management.

Using Metadata to Integrate and Apply Digital Media. New York, USA: McGraw-Hill, 1-24.

Bray, T., Paoli, J., Sperberg-McQueen, C.M. & Maler, E. 2000. Extensible Markup Language (XML) 1.0. (2nd Edition). W3C Recommendation. 1st version 10 Feb. 1998, revised 6 Oct. 2000. W3C Consortium. http://www.w3.org/TR/2000/REC-xml-20001006. [February 2, 2004]

Buckland, M.K. 1997. What is a "Document"? Journal of the American Society for Information Science 48 (9), 804-809.

Böhm, K. & Rakow, T.C. 1994. Metadata for multimedia documents. ACM SIGMOD Records 23 (4), 21-26.

Chen, P.P.S. 1976. The entity-relationship model - towards a unified view of data. ACM Transactions on Database Systems 1 (1), 9-36.

Clark, J., ed. 1999. XSL Transformations (XSLT) Version 1.0. W3C Recommendation 16 Nov. 1999. W3C Consortium. http://www.w3.org/TR/xslt. [February 2, 2004]

Consultative Committee for Space Data Systems. 2002. Reference Model for an Open Archival Information System (OAIS). http://ssdoo.gsfc.nasa.gov/nost/wwwclassic/documents/pdf/CCSDS-650.0-B-1.pdf. [April, 9, 2004]

Corel. 2004. Corel® XMetaL® Author. Corel. http://www.corel.com/servlet/Satellite?pagename=Corel/Products/productInfo&id=1042152756365&did=1042152754863. [January 30, 2004]

Dempsey, L. 1996. ROADS to Desire: some UK and other European metadata and resource discovery projects. D-Lib Magazine. http://www.dlib.org/dlib/july96/07dempsey.html. [January 14, 2004]

Deniman, D., Sumner, T., Davis, L., Bhusman, S. & Fox, J. 2003. Merging metadata and content-based retrieval. Journal of Digital Information 4 (3), Article No. 231.

Dervin, B. 1992. From the mind's eye of the user: The sense-making qualitative-quantitative methodology. In J. D. Glazier & R. R. Powell (Eds.). Qualitative Research in Information Management. Englewood (CO): Libraries Unlimited, Inc., 61-84.

DublinCore. 2003. Dublin Core Metadata Element Set, Version 1,1: Reference Description. Dublin Core Metadata Initiative. http://dublincore.org/documents/2003/06/02/dces/. [January 12, 2004]

Duff, W.M. 2001. Evaluating metadata on metalevel. Archival Science 1 (3), 285-294.

Dushay, N. 2002. Localizing experience of digital content via structural metadata. In Proceedings of the second ACM/IEEE-CS joint conference on Digital libraries. New York, USA: ACM Press, 244-252.

Ellis, C.A. 1979. Information control nets: a mathematical model of office information flow. In Proceedings of the Conference on Simulation, Measurement and Modeling of Computer Systems. 225-239.

Erdmann, M. & Studer, R. 1998. Use-cases and Scenarios for Developing Knowledge-Based Systems. In Proceedings of Information Technology

and KNOWledge Systems (IT&KNOWS), a part of the 15th IFIP World Computer Congress. Vienna/Budabest.

Fahrenholz-Mann, S. 1999. SGML for electronic publishing at a technical society - Expectations meets reality. Markup Languages: Theory and Practice 1 (2), 1-30.

Fallside, D.C.e. 2001. XML Schema Part 0: Primer. W3C Recommendation, 2 May 2001. W3C Consortium. http://www.w3.org/TR/xmlschema-0/. [February 2, 2004]

Garzotto, F., Mainetti, L. & Paolini, P. 1995. Hypermedia design, analysis, and evaluation issues. Communications of the ACM 38 (8), 74-86.

Gilliland-Swetland, A.J. 1998. Defining metadata. In M. Baca (Ed.) Introduction to Metadata. Los Angeles: Getty Information Institute, 1-8.

Goldfarb, C.F. 1990. The SGML Handbook. Oxford, UK: Oxford University Press.

Gordon, M.D. 1997. It's 10 a.m. Do you know where your documents are? The nature and scope of information retrieval problems in business. Information Processing & Management 33 (1), 107-212.

Han, H., Giles, C.L., Manavoglu, E., Zha, H., Zhang, Z. & Fox, E.A. 2003. Automatic document metadata extraction using support vector machines. In Proceedings of the Third ACM/IEEE-CS Joint Conference on Digital Libraries. Houston, Texas: IEEE Computer Society, 37-48.

Heikkinen, B. 2000. Generalization of Document Structures and Document Assembly. Helsinki: University of Helsinki.

Hill, L.L., Janée, G., Dolin, R., Frew, J. & Larsgaard, M. 1999. Collection metadata solutions for digital library applications. Journal of the American Society for Information Science 50 (13), 1169-1181.

Hodgins, W. 2002. IEEE Standard for Learning Object Metadata. IEEE Learning Technology Standards Committee. http://ltsc.ieee.org/wg12/. [January 12, 2004]

Iannella, R. 2002. Open Digital Rights Language (ODRL) Version 1.1. W3C. http://www.w3.org/TR/odrl/. [January 12, 2004]

ISO-10179. 1996. International Organization of Standardization, Information technology - Text and office systems - Document Style Semantics and Specification Language (DSSSL). ftp://ftp.ornl.gov/pub/sgml/wg8/dsssl/dsssl96f.pdf. [February 2, 2004]

ISO 15489-1. 2001. Information and documentation - Records management.

ISO TC 46/SC 11/WG 1.2003. Information and documentation - Records Management Processes - Metadata for Records. ISO.

Jokela, S., Turpeinen, M. & Sulonen, R. 2000. Ontology development for flexible content. In R. H. Sprague (Ed.) Proceedings of the 33rd Annual Hawaii International Conference on System Sciences (HICSS). Los Alamitos CA: IEEE Computer Society.

Jokela, S. 2001. Metadata enhanced content management in media companies. Acta Polytechnica Scandinavia. Mathematics and Computing Series No. 114. Helsinki: Helsinki University of Technology. Doctoral thesis.

Karjalainen, A., Päivärinta, T., Tyrväinen, P. & Rajala, J. 2000. Genre-based metadata for enterprise document management. In R. H. Sprague (Ed.) Proceedings of the 33rd Annual Hawaii International Conference on System Sciences (HICSS). Los Alamitos CA: IEEE Computer Society.

Karjalainen, A. & Salminen, A. 2000. Bridging the gap between hard and soft information genres. In M. Khosrowpour (Ed.) Challenges of Information Technology Management in the 21st Century. Proceedings of 2000 Information Resources Management Association International Conference. Hershey, U.S.A.: Idea Group Publishing, 92-95.

Karjalainen, A. & Tyrväinen, P. 2001. Defining genres and their features for studying information reuse: Preliminary findings of a case study of training materials. In M. Khosrowpour (Ed.) Managing Information Technology in a Global Economy. Proceedings of Information Resources Management Association 2001 conference (21-23.5.2001 Toronto, Canada). Hershey, U.S.A.: Idea Group Publishing, 346-348.

Kashyap, V. & Sheth, A. 1997. Semantic heterogeneity in global information systems: The role of metadata, context and ontologies. In M. Papzoglou & G. Schlageter (Eds.). Cooperative Information Systems: Current Trends and Directions. Springer-Verlag, 139-178.

Kavadias, G. & Tambouris, E. 2003. A markup language for describing public services and life events. In M. A. Wimmer (Ed.) Knowledge Management in Electronic Government, 4th IFIP International Working Conference, KMGov 2003, Rhodes, Greece, May 26-28, 2003, Proceedings. Springer, 106-115.

Kobayashi, M. & Takeda, K. 2000. Information retrieval on the web. ACM Computing Surveys 32 (2), 144-173.

Kock, N.F., Jr., McQueen, R.J. & Scott, J.L. 1997. Can action research be made more rigorous in a positivist sense? The contribution of an iterative approach. Journal of Systems & Information Technology 1 (1), 1-24.

Kokkelink, S. & Schwänzl, R. 2002. Expressing Qualified Dublin Core in RDF/XML. Dublin Core Metadata Initiative. http://dublincore.org/documents/dcq-rdf-xml/. [April 13, 2004]

Kuikka, E. 1996. Processing of Structured Documents Using a Syntax-Directed Approach. Department of Computer Science and Applied Mathematics. Kuopio: University of Kuopio. PhD Thesis, Kuopio University Publications C. Natural and Environmental Sciences 53.

Kuikka, E. & Salminen, A. 1997. Two-dimensional filters for structured text. Information Processing & Management 33 (1), 37-54.

Lassila, O. & Swick, R.R. 1999. Resource Description Framework (RDF) Model and Syntax Specification. W3C Recommendation 22 February 1999. http://www.w3.org/TR/REC-rdf-syntax/. [January 15, 2004]

Leinonen, P. & Penttonen, M. 1998. Producing SGML-documents with Public Domain Tools. Joensuu: University of Joensuu, Department of Computer Science.

Lindén, G. 1997. Structured Document Transformations. Department of Computer Science. Helsinki: University of Helsinki. Doctoral thesis, Series of Publications A, Report A-1997-2.

Luoma, E., Tiainen, S. & Tyrväinen, P. 2003. Integrated domain model for digital rights management. In M. Khosrowpour (Ed.) The Proceedings of the Information Resources Management Conference, IRMA 2003, Philadelphia Pennsylvania, USA, May 18-21, 2003. Hershey, U.S.A.: Idea Group Publishing.

Lyytikäinen, V., Tiitinen, P. & Salminen, A. 2000a. Challenges for European legal information retrieval. In F. Galindo & G. Quirchmayer (Eds.). Proceedings of the IFIP 8.5 Working Conference on Advances in Electronic Government. Zaragoza: Seminario de Informatica y Derecho, Universidad de Zaragoza, 121-132.

Lyytikäinen, V., Tiitinen, P. & Salminen, A. 2000b. Graphical information models as interfaces for Web document repositories. In V. Di Gesú, S. Levialdi & L. Tarantino (Eds.). Proceedings of the Working Conference on Advanced Visual Interfaces, AVI 2000. New York: ACM Press, 261-265.

Magnusson Sjöberg, C. 1998. Critical Factors in Legal Document Management. Stockholm: Jure.

Maler, E. & El Andaloussi, J. 1996. Developing SGML DTDs. From text to model to markup. Upper Saddle River (NJ): Prentice Hall.

McIntosh, M. 2000. Content Management Using the Rational Unified Process®. Rational Software White Paper. Rational Software Corporation. http://www.rational.com/media/products/rup/TP164.pdf. [November 15, 2002]

McKemmish, S. & Acland, G. 1999. Accessing essential evidence on the Web: Towards an Australian recordkeeping metadata standard. In The Web after a Decade: Proceedings of AusWeb99, the Fifth Australian World Wide Web Conference, April 1999.

Murphy, L.D. 1998. Digital document metadata in organizations: Roles, analytical approaches, and future research directions. In R. H. J. Sprague (Ed.) Proceedings of the 31st Annual Hawaii International Conference on System Sciences (HICSS). Los Alamitos CA: IEEE Computer Society, 267-276.

Nielsen, J. & Mack, R. 1994. Usability Inspection Methods. John Wiley & Sons, Inc.

Nunamaker, J., F., Jr., Chen, M. & Purdin, T., D. M. 1991. Systems development in information systems research. Journal of Management Information Systems 7 (3), 89-106.

Orlikowski, W.J. & Yates, J. 1998. Genre Systems: Structuring Interaction through Communicative Norms. MIT Sloan School of Management. http://ccs.mit.edu/papers/CCSWP205/. [February 2, 2004]

Pitti, D.V. 2003. Creator description: Encoded Archival Context. In Proceedings International Conference Authority Control: Definition and International Experiences, Florence.

Powell, J. & Fox, E.A. 1998. Multilingual federated searching across heterogeneous collections. D-Lib Magazine. http://www.dlib.org/dlib/ september98/powell/09powell.html. [January 13, 2004]

Päivärinta, T. & Tyrväinen, P. 1998. Documents in information management: Diverging connotations of "a Document" in digital era. In M. Khosrowpour (Ed.) Proceedings of the 9th Information Resource Management Association International Conference. Hershey, PA, U.S.A.: Idea Group Publishing,, 163-173.

Päivärinta, T. 1999. A genre approach to applying critical social theory to information systems development. In C. H. J. Gilson, I. Grugulis & H. Willmot (Eds.). Proceedings of the 1st Critical Management Studies Conference; Information Technology and Critical Theory - Stream. Manchester.

Päivärinta, T., Salminen, A. & Peltola, T. 1999. Improving enterprise document management by a quality system: a case study. In J. Pries-Heje, C. Ciborra, K. Kautz, J. Valor, E. Christiaanse, D. Avison & C. Heje (Eds.). Proceedings of the European Conference on Information Systems. Copenhagen: Department of Informatics, Copenhagen Business School, 922-933.

Päivärinta, T. 2000. Towards a Genre-Based, Critical Approach to Enterprise Document Management. Department of Computer Science and Information Systems. Jyväskylä: University of Jyväskylä. Lic. Thesis.

Päivärinta, T. 2001. A Genre-Based Approach to Developing Electronic Document Management in the Organization. Department of Computer Science and Information Systems. Jyväskylä: University of Jyväskylä. PhD Thesis, Jyväskylä Studies in Computing 11.

Päivärinta, T. & Peltola, T. 2001. Engineering of a genre-based method for developing electronic document management: The consultant's viewpoint. In J. Krogstie, K. Siau & T. Halpin (Eds.). Proceedings of the Sixth CAiSE/IFIP8.1 International Workshop on Evaluation of Modeling Methods in Systems Analysis and Design (EMMSAD´01). XIII 1-14.

Pöyry, P., Pelto-Aho, K. & Puustjärvi, J. 2002. The role of metadata in the CUBER system. In Proceedings of the 2002 Annual Research Conference of the South African Institute of Computer Scientists and Information Technologists on Enablement Through Technology. Port Elizabeth, South Africa: South African Institute for Computer Scientists and Information Technologists, 172-178.

Raggett, D., Le Hors, A. & Jacobs, I. 1999. HTML 4.01 Specification. W3C Recommendation 24 Dec 1999. World Wide Web Consortium. http://www.w3.org/TR/html4/. [February 2, 2004]

Recordkeeping Metadata Standard for Commonwealth Agencies. 1999. National Archives of Australia. http://www.naa.gov.au/recordkeeping/ control/rkms/summary.htm. [January 12, 2004]

Rothenberg, J. 1995. Ensuring the longevity of digital documents. Scientific American 272 (1), 42-27.

Rödig, P., Borghoff, U.M., Scheffczyk, J. & Schmitz, L. 2003. Preservation of digital publications: An OAIS extension and implementation. In Proceed-

ings of the 2003 ACM Symposium on Document Engineering. New York, USA: ACM Press, 131-139.

Saarela, J. 1999. The role of metadata in electronic publishing. Acta Polytechnica Scandinavia. Mathematics and Computing Series. Espoo: Finnish Academy of Technology. Doctoral thesis.

Saaren-Seppälä, K. 1997. Seinätekniikka prosessien kehittämisessä. (Using the wall-chart technique in process development, in Finnish), Kari Saaren-Seppälä Ltd., Finland.

Salminen, A. & Tompa, F.W. 1992. PAT expressions: an algebra for text search. Acta Linguistica Hungarica 41 (1-4), 277-306.

Salminen, A. & Watters, C. 1992. A two-level structure for textual databases to support hypertext access. Journal of the American Society for Information Science 43 (6), 432-447.

Salminen, A., Tague-Sutcliffe, J. & McClellan, C. 1995. From text to hypertext by indexing. ACM Transactions on Information Systems 13 (1), 69-99.

Salminen, A., Lehtovaara, M. & Kauppinen, K. 1996. Standardization of digital legislative documents, a case study. In Proceedings of the 29th Annual Hawaii International Conference on System Sciences. Maui (HA): IEEE Computer Society Press, 72-81.

Salminen, A., Kauppinen, K. & Lehtovaara, M. 1997. Towards a methodology for document analysis. Journal of the American Society for Information Science 48 (7), 644-655.

Salminen, A. & Tompa, F.W. 1999. Grammars++ for Modelling Information in Text. Information Systems 24 (1), 1-24.

Salminen, A. 2000. Methodology for document analysis. In A. Kent (Ed.) Encyclopedia of Library and Information Science. New York: Marcel Dekker, Inc., 299-320.

Salminen, A. 2003. Document Analysis Methods. In C. L. Bernie (Ed.) Encyclopedia of Library and Information Science, Second Edition, Revised and Expanded. NY, USA: Marcel Dekker, Inc., 916-927.

Schamber, L. 1996. What is a document? Rethinking the concept of uneasy times. Journal of the American Society for Information Science 47 (9), 669-671.

Shlaer, S. & Mellor, S., J. 1992. Object Lifecycles: Modeling the World in States. Englewood Clisffs (NJ): Yourdon Press.

Soo, V.-W., Lee, C.-Y., Li, C.-C., Chen, S.L. & Chen, C.-c. 2003. Automated semantic annotation and retrieval based on sharable ontology and case-based learning techniques. In Proceedings of the third ACM/IEEE-CS joint conference on Digital libraries. Houston, Texas: IEEE Computer Society, 61-72.

Sprague, R.H. 1995. Electronic document management: challenges and opportunities for information systems manager. MIS Quarterly 19 (1), 29-49.

Susman, G.I. & Evered, R.D. 1978. An assessment of the scientific merits of action research. Administrative Science Quarterly 23 (4), 582-603.

Sutton, M.J.D. 1996. Document Management for the Enterprise: Principles, Techniques and Applications. New York, U.S.A.: Wiley and Sons, Inc.

Swales, J.M. 1999. Genre Analysis. English in Academic and Research Settings. Cambridge: Cambridge University Press.

Tiitinen, P. 2003. User roles in document analysis. In J. Eder & T. Welzer (Eds.). CAiSE '03 Forum, Forum for short contributions at The 15th Conference On Advanced Information Systems Engineering. Maribor, Slovenia: Maribor University Press, 205-208.

Tyrväinen, P. & Päivärinta, T. 1999. On rethinking organizational document genres for electronic document management. In R. H. Sprague (Ed.) Proceedings of the 33th Annual Hawaii International Conference on System Sciences. Maui (HA): IEEE Computer Society Press, HICSS Digital Library. 10 pp.

Watson, B.C. & Shafer, K. 1995. Creating custom SGML DTDs for documentation products. In Proceedings of the 13th Annual International Conference on Systems Documentation. Emerging from Chaos: Solutions for the Growing Complexity of Our Jobs. 189-196.

Wimmer, M.A. & Tambouris, E. 2002. Online One-Stop Government: A working framwork and requirements. In R. Traunmüller (Ed.) Information Systems: The e-Business Challenge. Proceedings of the 17th World Computer Congress of IFIP in Montreal. Boston: Kluwer Academic Publishers, 117-130.

Yin, R.K. 1994. Case Study Research: Design and Methods. Newbury Park: Sage Publications.

Yoshioka, T., Yates, J. & Orlikowski, W.J. 2001. Genre taxonomy: A knowledge repository of communicative actions. ACM Transactions on Information Systems 19 (4), 431-456.

APPENDIX 1
Domain.dtd for Describing Contextual Metadata

```
<!-- ********************************************************
Domain.dtd: describes the contextual metadata of a domain
Author: Virpi Lyytikäinen 30.1.2004
******************************************************** -->
<!-- DOMAIN -->
<!-- Root element for describing the whole domain     -->
<!ELEMENT domain   (descrip, systems?, docs?, actors?, pros?)  >

<!-- Description about the domain including area (geographical cov-
erigdge) as well as the goal of the domain -->
<!ELEMENT descrip       (area?, datespan?, goal, info*)           >
<!ELEMENT datespan      (begindate, enddate)          >
<!ELEMENT begindate     (#PCDATA)                     >
<!ELEMENT enddate       (#PCDATA)                     >
<!ELEMENT goal          (text+)                       >
<!ELEMENT area          (text+)                       >

<!--Systems -->
<!ELEMENT systems (system*)                       >
<!ELEMENT system  (name, info*, content?)         >
<!ATTLIST system
            id                ID      #IMPLIED
            inputsystems      IDREFS  #IMPLIED
            outputsystems     IDREFS  #IMPLIED
            responsibleorgs   IDREFS  #IMPLIED  >
<!ELEMENT content (text+)                         >

<!-- DOCUMENTS -->
<!-- Documents can be grouped into document groups.
The documents may have relationships to each other. In addition,
documents are stored in one or more systems, which is indicated
by attribute reference -->
<!ELEMENT docs      (docgroup+)                   >
<!ELEMENT docgroup  (name, info*, systemlink*, (document| docgroup)*)
>
<!ATTLIST docgroup
            id   ID           #IMPLIED           >
<!ELEMENT document      (name, info*, systemlink*)          >
<!ATTLIST document
            id   ID           #IMPLIED
            children IDREFS   #IMPLIED           >

<!-- ACTORS -->
<!-- Actormodel consists of actor groups, which have names and
roles towards the goal of the environment. Actor groups can include
actors, which also have names and possible roles. Some systems may in-
clude information produced by an actor -->
<!ELEMENT actors   (actorgrp+)                    >
<!ELEMENT actorgrp  (name, role?, info*, systemlink*, (actorgrp |
actor)*)>
<!ATTLIST actorgrp
                  id  ID             #IMPLIED   >
<!ELEMENT role          (text+)                   >
<!ATTLIST role
                  id  ID             #IMPLIED   >
```

```
<!ELEMENT actor   (name, role?, info*, systemlink*)   >
<!ATTLIST actor
                  id  ID              #IMPLIED  >

<!-- PROCESSES -->
<!-- Processes are composed of activities and connectors, which can
occur in any order, followed by the name of documentation (virtual re-
pository for all the documents created in the process.  An activity
has a name followed by optional output documents.
The system including the documents is indicated by an attribute refer-
ence. Connectors are either type and or or. -->
<!ELEMENT pros          ((activity | connect)+, docution)   >
<!ELEMENT activity      (text+, info*, systemlink*)         >
<!ATTLIST activity
            id  ID                  #IMPLIED
                  next IDREF        #IMPLIED
            docsin IDREFS           #IMPLIED
            docsout IDREFS          #IMPLIED
                  orgs IDREFS             #IMPLIED    >
<!ELEMENT connect       EMPTY                           >
<!ATTLIST connect
                  id  ID            #IMPLIED
                  nexts  IDREFS     #IMPLIED
                  type  (and | or)  "or"              >
<!ELEMENT docution      (text+)                         >


<!-- COMMON ELEMENTS-->

<!-- Link to system including desired documents -->
<!ELEMENT systemlink    EMPTY                   >
<!ATTLIST systemlink
            systemref   IDREFS       #REQUIRED  >


<!-- Normal text paragraph in different languages -->
<!ELEMENT text          (#PCDATA)               >
<!ATTLIST text language (fi | en | fr |es | da | de | el | it | nl |
pt | sv) "en" >
<!-- Technical, unique name -->
<!ELEMENT name          (text+)                         >

<!-- Longer description -->
<!ELEMENT info    (title | subtitle | para | list | anchor)+ >
<!ATTLIST info language (fi | en | fr |es | da | de | el | it | nl |
pt | sv) "en" >
<!ELEMENT title         (#PCDATA)               >
<!ELEMENT subtitle      (#PCDATA)               >
<!ELEMENT para          (#PCDATA)               >
<!ELEMENT list          (item+)                 >
<!ATTLIST list  type (bullet | number | dash ) "bullet"     >

<!ELEMENT item          (#PCDATA)               >
<!ELEMENT anchor        (#PCDATA)               >
<!ATTLIST anchor link CDATA         #IMPLIED    >
```

An Example of the Use of the domain.dtd

```
<?xml version="1.0" encoding="iso-8859-1"?>
<!DOCTYPE domain SYSTEM "domain.dtd">
<domain>
  <descrip>
    <area>
      <text>Central Finland</text>
    </area>
    <goal>
      <text>Planning and communicating church services</text>
      <text language="fi">Seurakunnan toiminnan suunnittelu</text>
    </goal>
  </descrip>
  <systems>
    <system id="DM1">
      <name>
        <text>Document management system</text>
      </name>
    </system>
  </systems>
  <docs>
    <docgroup>
      <name>
        <text>Short term plans</text>
        <text language="fi">Lyhyemmän aikavälin
          suunnitelmat</text>
      </name>
      <document id="MSC">
        <name>
          <text>Monthly service calendar</text>
          <text language="fi">Kuukausikalenteri</text>
        </name><systemlink systemref="DM1"/>
      </document>
      <document>
        <name>
          <text>Vacation list</text>
          <text language="fi">Vapaapäivälista</text>
        </name><systemlink systemref="DM1"/>
      </document>
      <document id="WC">
        <name>
          <text>Weekly calendar</text>
          <text language="fi">Viikkokalenteri</text>
        </name><systemlink systemref="DM1"/>
      </document>
      <document id="AN">
        <name>
          <text>Announcements on newspaper</text>
          <text language="fi">Lehti-ilmoitukset</text>
        </name><systemlink systemref="DM1"/>
      </document>
      <document>
        <name>
          <text>Outdoor announcement</text>
          <text language="fi">Seinätauluilmoitus</text>
        </name><systemlink systemref="DM1"/>
      </document>
    </docgroup>
  </docs>
```

```
<actors>
 <actorgrp id="EMP">
  <name>
    <text>Employees</text>
    <text>Työntekijät</text>
  </name>
  <role>
    <text>Operational responsibility in domain</text>
    <text language="fi">Operationaalinen vastuu</text>
  </role><systemlink systemref="DM1"/>
  <actorgrp id="PAS">
    <name>
     <text>Pastors</text>
     <text language="fi">Pastorit</text>
    </name>
    <actor>
     <name>
      <text>Senior Pastor</text>
      <text language="fi">Seurakunnan johtaja</text>
     </name>
    </actor>
    <actor>
     <name>
      <text>Associate Pastor</text>
      <text language="fi">Seurakuntapastori</text>
     </name>
    </actor>
    <actor>
     <name>
      <text>Youth Pastor</text>
      <text language="fi">Nuorisopastori</text>
     </name>
    </actor>
    <actor id="CWC">
     <name>
      <text>Child Work Coordinator</text>
      <text language="fi">Lapsityösihteeri</text>
     </name>
    </actor>
    <actor id="EV">
     <name>
      <text>Evangelist</text>
      <text language="fi">Evankelista</text>
     </name>
    </actor>
  </actorgrp>
  <actorgrp>
    <name>
     <text>Administration</text>
     <text>Hallintohenkilökunta</text>
    </name>
    <actor id="CLERK">
     <name>
      <text>Church Clerk</text>
      <text language="fi">Toimistotyöntekijä</text>
     </name>
    </actor>
    <actor>
     <name>
      <text>Caretaker</text>
```

```
          <text language="fi">Talonmies</text>
         </name>
        </actor>
      </actorgrp>
      <actor>
        <name>
         <text>School worker</text>
         <text language="fi">Koulutyöntekijä</text>
        </name>
      </actor>
      <actor>
        <name>
         <text>Missionary</text>
         <text language="fi">Lähetystyöntekijä</text>
        </name>
      </actor>
    </actorgrp>
    <actorgrp id="WEB">
      <name>
        <text>Webmaster</text>
        <text language="fi">Webmaster</text>
      </name>
      <role>
        <text>Electronic publishing</text>
        <text language="fi">Elektroninen julkaiseminen</text>
      </role>
    </actorgrp>
  </actors>
  <pros>
    <activity id="AC1" docsout="MSC" orgs="CLERK CWC EV" next="C1">
      <text>Constructing Monthly service calendar</text>
      <text language="fi">Kuukausikalenterin tuottaminen
</text><systemlink
      systemref="DM1"/>
    </activity> <connect type="and" id="C1" nexts="AC2 AC3 AC4"/>
    <activity id="AC2" docsout="WC" orgs="PAS">
      <text>Constructing Weekly calendar</text>
      <text language="fi">Viikkokalenterin tuottaminen</text><systemlink
      systemref="DM1"/>
    </activity>
    <activity id="AC3" docsout="AN" orgs="EMP">
      <text>Sending announcements to the newspaper</text>
      <text language="fi">Sanomalehti-ilmoituksen laadinta
</text><systemlink
      systemref="DM1"/>
    </activity>
    <activity id="AC4" docsout="MSC" orgs="WEB">
      <text>Publishing the calendar on the Web</text>
      <text>Kalenterin julkaiseminen Webissä</text><systemlink
systemref="DM1"/>
    </activity>
    <docution>
      <text>Planning documentation</text>
      <text language="fi">Suunnitteludokumentaatio</text>
    </docution>
  </pros>
</domain>
```