JYVÄSKYLÄ STUDIES IN COMPUTING
266

Reija Nurmeksela

Implementing Structured Document Production to Support Enterprise Content Management

JYVÄSKYLÄN YLIOPISTO

JYVÄSKYLÄ STUDIES IN COMPUTING 266

Reija Nurmeksela

Implementing Structured Document Production to Support Enterprise Content Management

Esitetään Jyväskylän yliopiston informaatioteknologian tiedekunnan suostumuksella
julkisesti tarkastettavaksi yliopiston Agora-rakennuksen auditoriossa 2
marraskuun 11. päivänä 2017 kello 12.

Academic dissertation to be publicly discussed, by permission of
the Faculty of Information Technology of the University of Jyväskylä,
in building Agora, auditorium 2, on November 11, 2017 at 12 o'clock noon.

UNIVERSITY OF JYVÄSKYLÄ

JYVÄSKYLÄ 2017

Implementing Structured Document Production to Support Enterprise Content Management

JYVÄSKYLÄ STUDIES IN COMPUTING 266

Reija Nurmeksela

Implementing Structured Document Production to Support Enterprise Content Management

UNIVERSITY OF JYVÄSKYLÄ

JYVÄSKYLÄ 2017

Editors
Marja-Leena Rantalainen
Faculty of Information Technology, University of Jyväskylä
Pekka Olsbo, Ville Korkiakangas
Publishing Unit, University Library of Jyväskylä

Permanent link to this publication: http://urn.fi/URN:ISBN:978-951-39-7210-3

URN:ISBN:978-951-39-7210-3
ISBN 978-951-39-7210-3 (PDF)

ISBN 978-951-39-7209-7 (nid.)
ISSN 1456-5390

Jyväskylä University Printing House, Jyväskylä 2017

ABSTRACT

Nurmeksela, Reija
Implementing Structured Document Production to Support Enterprise Content Management
Jyväskylä: University of Jyväskylä, 2017, 83 p.
(Jyväskylä Studies in Computing
ISSN 1456-5390; 266)
ISBN 978-951-39-7209-7 (nid.)
ISBN 978-951-39-7210-3 (PDF)
Finnish summary
Diss.

Within enterprise content management (ECM), the major goal is to develop and deploy systematic solutions for managing documents and other content items. ECM implementation concerns the development and deployment of new content management solutions and practices in an organization. Extensible Markup Language (XML) offers a standardized format for documents supporting the management and preservation of documents as structured documents. However, the deployment of XML may require a demanding standardization process, changes in work practices, and new tools for document management. Consequently, this research explores the implementation of structured document production environments. The focus is on documents that end-users author during ongoing business processes. Moreover, the aim of this study is to increase the understanding of structured document production and to provide a framework for XML standardization. The framework enables the analysis and development of a structured document production environment in an organization. This research follows the design science and case study approaches. The standardization of the Finnish Parliamentary documents in the Government of Finland and in the Finnish Parliament is used as the major case environment for analyzing structured document production. This case is compared to two other cases. By analyzing these cases and the previous literature, content production strategies are introduced, and challenges to XML standardization are presented. In addition, models for XML document management are proposed. This study shows that an XML document management environment is a complex combination of varying content items, processes, actors with diverse backgrounds, and evolving systems. Structured document production is a strategic choice requiring management and end-user commitment during the standardization process. The usability of novel tools requires special focus. This research shows that developing custom-designed editors, hiding document structures from users, and automating document and metadata creation increase user acceptance of novel tools and practices.

Keywords: Enterprise content management, ECM, XML, structured documents

Author
Reija Nurmeksela
Faculty of Information Technology
University of Jyväskylä
Finland

Supervisors
Airi Salminen
Faculty of Information Technology
University of Jyväskylä
Finland

Mikko Siponen
Faculty of Information Technology
University of Jyväskylä
Finland

Reviewers
Professor Jan vom Brocke
University of Liechtenstein
Liechtenstein

Professor Samuli Pekkola
Industrial and Information Management
Tampere University of Technology
Finland

Opponent
Professor Tero Päivärinta
Department of Computer Science, Electrical and Space Engineering
Luleå University of Technology
Sweden

ACKNOWLEDGEMENTS

I would like to thank my supervisor, Professor Airi Salminen, for the opportunity to participate in the RASKE2 research project, and for her patience, guidance, and support throughout the long thesis process. Our discussions, our joint ideas, and her feedback have been invaluable. I also would like to thank my second supervisor, Professor Mikko Siponen, for guiding me in finalizing this thesis.

I would like to express my appreciation to my thesis opponent, Professor Tero Päivärinta, and thesis reviewers, Professor Jan vom Brocke and Professor Samuli Pekkola, for their valuable opinions and suggestions for this thesis. I am also thankful to all co-authors of the articles included in this thesis, as well as colleagues in the RASKE2 and RAKE research projects. Particularly, I would like to thank Anne Honkaranta, Eliisa Jauhiainen, and Maritta Pirhonen for shared study moments at the university.

I appreciate the opportunity I have had to work as a researcher and as a practitioner in the case organizations involved in this thesis work. Particularly, my long-term cooperation with the Finnish Parliament and numerous experts there has been invaluable. I would like to extend my warmest thanks especially to Maija-Liisa Jääskeläinen, Jaana Kaakkola, Nina Katila, Tuula Kulovesi, Antti Linna, Olli Mustajärvi, Mia Nordlund, Stefan Nordman, Eila Nummi, Maari Paasilinna, Päivi Parviainen, Pauliina Peltokorpi, Maarit Peltola, Teuvo Räty, Timo Tuovinen, Päivi Virta-Leskinen, and, particularly, Sanna Turpeinen. I additionally would like to thank Ulla Kaipainen, Terja Ketola, Marika Paavilainen and Tanja Rönn at the Prime Minister's Office, and Johanna Orivuori at the Ministry of Finance for their cooperation regarding the topics of this thesis.

I would like to thank my long-term employer, Tieto, and numerous colleagues and managers there. Special thanks go out to current and former managers Sissi Kolbe, Hanna Mattinen, Jyrki Poteri, Jouni Seppälä, and Marita Tolvanen for your support of my studies during these years. I have been lucky to work with so many professional colleagues, particularly during the most recent years of the Eduksi project. It has been great to work with all of you. It is not fair to list only a few names here, but regarding this thesis, I have appreciated the discussions, shared ideas, and experiences with Matti Autio, Marjut Eklund, Esko Heiskanen, Virpi Kemppainen, Heli Launonen, Tiina Leppänen, Lasse Nurmeksela, Jaana Patrikainen, Tuomo Penttinen, Janne Rossi, Hannu Ryynänen, Sole Salmijärvi, Eija Säkkinen, and Jarmo Vuorinen during the "Eduksi years."

Finally, I would like to thank my beloved family: My son, Sakari, and my husband, Pentti, have encouraged and supported me in this PhD work in the course of the past 13 years. The years would have been much tougher without your support and love. Now, it´s my time to support you in your journey.

Laukaa 15.10.2017
Reija Nurmeksela

FIGURES

TABLE

CONTENTS

1 INTRODUCTION

Documents produced in business processes are important means for human communication and evidence of business transactions in organizations. Thus, a great deal of the information resources in organizations consists of documents (Rockley et al. 2003). A *document* is a recorded information set structured for human comprehension and represented in some media by a variety of symbols that pertain to a topic (Sprague 1995). In contemporary organizations, documents and other content are produced and used in complex business process environments involving many kinds of information systems. The emergence of new technical innovations, such as social Web applications, has expanded both the characteristics (Hausmann & Williams 2015) and management requirements of documents (Grahlmann et al. 2012) in recent years. In the field of information systems research, *document management* is a term referring to the creation, storage, organization, transmission, retrieval, manipulation, update, and eventual disposition of documents in an organization to satisfy its purposes (Sprague 1995). Both researchers (Tyrväinen et al. 2006; Grahlmann et al. 2012) and practitioners (Herbst et al. 2014) regard document management as a subset of enterprise content management (ECM) where the major goal is to develop systematic solutions for managing documents and other content resources in organizations. Grahlmann et al. (2012, p. 5) define the term *enterprise content management* as follows:

> "*Enterprise Content Management* comprises the strategies, processes, methods, systems, and technologies that are necessary for capturing, creating, managing, using, publishing, storing, preserving, and disposing content within and between organizations."

For a long time, documents have been authored in digital format using word processors. With these tools, human authors write new or reuse existing content, for example, by retyping the content or by using copy-paste functions. Authors aggregate the content into the document, focusing on the layout for human consumption, and store the document in digital format for further authoring, editing, or reading purposes. Typically, the author also adds metadata related to the document, such as the document's title, author's name, the date, and the version. In ECM, *metadata* are a means to facilitate content management concerning functions related to the document and other content (Tyrväinen et al. 2006).

Digital documents may be processed by computer software in many different ways. In addition to supporting the authoring and reading needs of human users, a software application may automatically create new content or use and manipulate the existing content of digital documents. The automated management of document content may be based on the *structured document* approach, where authors, domain experts, or software designers have specified and named different parts of the document so software applications can identify, retrieve, and process the parts (Salminen & Tompa 2011). Extensible Markup Language (XML) (Bray et al. 2006) offers a standardized open format for structured documents. Nowadays, XML is widely used in various technical environments and is also a recommended format in many domains, for example, for parliamentary documents (the United Nations and the Inter-Parliamentary Union 2014).

In a *structured document production* environment, documents are created and stored as structured documents where the structure definitions, document instances, and layout specifications can be handled as separate content items. Structured documents may be produced in many different ways by human actors or automatically by a software system. Both the authoring and management of structured documents differ essentially from documents produced with traditional word processors. When creating or editing a structured document, a user or computer software focuses on the logical structure of the document instead of on the layout by marking the document content with the named parts. Computer software creates the document layout automatically by mapping the named parts to a layout specification. The names and order of the parts as well as the layout must be agreed upon beforehand in the environment. The agreement and implementation of rules for the document's structures, layout, management practices, and tools to be used in the work process is regarded as the *standardization* of the content management environment (Salminen 2005).

The document production practices in an organization have a major effect on the extent to which the content is accessible and how well the content supports operational efficiency and open data. The structured document approach has several well-known advantages concerning data interchange, document management, and the management of metadata related to documents (Salminen 2005). According to Salminen (2005), the organization may benefit from:

- more consistent and correct documents regarding structure, content, and layout,
- rich information-retrieval capabilities related to document content,

- the possibility of information reuse,
- the possibility of multichannel publishing,
- independency of particular software providers, and
- the long-term accessibility of information stored in documents.

However, the implementation of a structured document production environment may require a demanding document standardization process in an organization, as well as major changes to the work of people and to the tools used for document management (Salminen et al. 2000). The complexity and new competence needs have also been recognized in practical guidelines, for example, in technological recommendations for parliaments in Europe (OPPD 2010). Thus, managing the complexity and changes in the document production environment are the major issues in the implementation of structured document production (Salminen 2005), as in any ECM implementations (e.g., Päivärinta & Munkvold 2005).

Research on the implementation of structured document production started in the 1990s, when SGML (Standard Generalized Markup Language) (Goldfarb 1990), the mother language of XML, was adopted in organizations (e.g., Maler & El Andaloussi 1996; Braa & Sandahl 1997; Salminen et al. 1997). There was a little research during the initial years after XML's publication (e.g., Rockley et al. 2003; Salminen 2005). Since the publication of the first XML standard version in 1998, XML-related research has been very active, but the focus has been on technical issues. This is evidenced by simple Google Scholar searches (excluding patents). For example, a search for "XML" in June 2017 covering the years 2012–2017 resulted in 497,000 hits. Browsing the hits showed a clear emphasis on technical aspects. A search for "XML document management" covering the same years resulted in 114 hits. Among these, only a few relate to the questions of this thesis. Boer (2016) and Gen et al. (2016) have investigated the application of XML to legislative documents. In addition, Anderson and Eberlein (2015) have studied the implementation of component content management regarding technical documentation, and Jauhiainen (2014) and Jung et al. (2016) have investigated the adoption of open XML document formats in public sector organizations. Some researchers have proposed XML-based solutions to improve content and document management in particular domains, such as emergency medical services (Poulymenopoulou et al. 2014) and legal document and knowledge management (Boella et al. 2012).

XML research has been active in areas like access control (Su et al. 2014), data interchange (e.g., Moskal et al. 2015), data integration (e.g., Fan et al. 2016), the storage and categorization of documents (e.g., Feki et al. 2013; Di Iorio et al. 2014; Rezk et al. 2016), technical change management (e.g., Brahmia et al. 2016), and the visualization of XML documents (e.g., Luo et al. 2017). Additionally, new techniques for XML information retrieval (e.g., Chatvichienchai et al. 2015;

Ikhsan & Hasbi 2016; Thiam 2016) have been developed. In the 2010s, some reviews and analyses of ECM literature have been conducted (e.g., Alalwan & Weistroffer 2012; Grahlmann et al. 2012; Rickenberg et al. 2012a), and the results confirm that the number of academic studies on ECM is still low.

This research explores the implementation of structured document production environments and factors influencing the development of these environments in organizations. The focus is on documents authored by human users during ongoing business processes. Earlier research on the area has been rare, possibly because it requires access to real implementation projects in organizations. The researcher of this thesis participated in several implementation projects during this long-lasting research process as an XML specialist and consultant. The practical work naturally slowed the progress of the dissertation work. Seeing the huge number of research results that evolved during the years concerning new techniques for the effective management of XML data kept the researcher's motivation for this study strong. The successful adoption of XML for documents in the ECM environment of an organization would enable the use of a great number of XML technologies for the management of data in the documents. Consequently, the aim of this study is to increase the understanding of structured document production and to provide a framework for XML standardization.

This research follows the design science (March & Smith 1995) and case study (Yin 1994) approaches. The standardization of the Finnish Parliamentary documents in the Government of Finland and Finnish Parliament is used as the major case environment for analyzing structured document production. This case is compared to two others: the standardization of an invoice document for an international ICT (information and communication technology) provider company and its customers, and the standardization of the administrative documents and statements in the Finnish Centre for Pensions regarding earnings-related pensions.

The main contribution of this study is threefold. First, the research shows that structured document production is a strategic choice for content production in an organization requiring management commitment during the standardization process. The implementation of the environment may be long-lasting, and several problems may need to be solved when implementing the environment. In a complex environment, such as that involving Finnish Parliamentary documents, challenges may be faced regarding all the entities of the ECM environment: the activities, actors, systems, and content items of the domain.

Second, the research shows that an XML document management environment is a complex combination of varying content items, activities, actors with diverse backgrounds, and continuously evolving systems. Because of this complexity, the implementation requires analysis. The models proposed in this study provide tools for analyzing the XML document management environment, the XML document life cycle, and integrated XML document production. The models should help researchers achieve a better understanding of the characteristics of the environment and innovate further research ideas. More importantly, the models should help practitioners to develop, deploy, and maintain structured document production environments within practical ECM solutions. The models also suggest the connection of structured document production first to document and records management, and second to case and business process management. For practitioners, these models should help to develop document production solutions that are integrated with various ECM systems and facilitate automated document and metadata creation. The models proposed in this study would also be used to support the deployment of the developed solutions in organizations.

Third, because structured document production differs significantly from traditional authoring, usability requires special focus. This research shows that hiding document structures from the users, developing custom-designed editors, and automating document creation increases the user acceptance of novel tools.

The research also increases knowledge concerning the rare empirical research of ECM implementations (Alalwan & Weistroffer 2012) and successful ECM implementations (Usman et al. 2009). In addition, it brings more knowledge regarding rare empirical studies concerning the implementation of structured authoring.

The study presented in this thesis is limited to documents that human users author during ongoing business processes and require management and preservation as records. These are typical characteristics of the documents produced in the public sector as evidence of activities in the domain. Another limitation concerns data from the major case environment: the adopters of the structured document approach in the Finnish Parliament and Government can be classified as early adopters. Research has shown that early adopters sometimes face tremendous obstacles because the tools and technology are not yet mature (Chen 2003). Regardless of the limitations, the models proposed in this study could be applied, or at least tested, in further research in other domains.

This thesis is structured as follows. The next chapter introduces the core concepts and theoretical aspects of the thesis. The third chapter describes the research design and methodology. The fourth chapter introduces the results of each article, and the fifth chapter summarizes the contribution of this thesis. Finally, the sixth chapter discusses the theoretical and practical implications of this work, along with its limitations. Future research topics are also presented, with the original articles attached as appendices.

2 THEORETICAL FOUNDATION

In this chapter, the core concepts and theoretical foundation of this thesis are introduced. The focus is on the concept definition that is typical for ECM research (vom Brocke & Simons 2014). First, the chapter discusses research on ECM and presents the content management model designed for supporting the analysis of content production practices in organizations. This is followed by a more specific analysis of XML document management and the introduction of XML document production. Finally, methods for XML document management are presented.

2.1 Enterprise content management (ECM)

This section discusses the first research on ECM and its relationship to research on records management. Then, an ECM environment model is presented. The model is used as a tool for various analyses in this study.

2.1.1 Research on ECM

Due to the continuous increase of digital information assets and the rapid implementation of ECM systems in organizations, ECM research has become an important and complex topic for information systems (IS) research (Tyrväinen et al. 2006). According to the literature review of Grahlmann et al. (2012) and the case analysis by Herbst et al. (2014), ECM implementations may cover various viewpoints and activities related to digital information management, such as document, record, case, workflow, and process management. Thus, the functionalities of ECM solutions in organizations vary regarding access to content, content management processes, services related to the content, and repositories where the content is stored (Grahlmann et al. 2012). An ECM implementation process may be analyzed from development and deployment viewpoints (Tyrväinen et al. 2006).

ECM as a concept has evolved during the past several years. The term *enterprise content management* was first introduced in 2001 by AIIM International, a nonprofit global community of information professionals (Blair 2004). Although various definitions for the concept (e.g., Smith & McKeen 2003, p. 648; Grahlmann et al. 2012, p. 5) exist, ECM research is understood to cover the strategies, methods, processes, systems, technologies, and social issues related to the management of information in organizations (Tyrväinen et al. 2006; Grahlmann et al. 2012).

When developing document production in contemporary organizations, it might be difficult to recognize digital content items that should be considered documents (Päivärinta & Tyrväinen 1998, Honkaranta 2003; Hausmann & Williams 2015). This study is limited to the production of documents created during ongoing business processes as evidence of business transactions. Because of their nature as legal evidence, these documents require management and preservation as records. In records management standard 15489-1 (ISO 15489-1, 2001, p. 3) of the International Organization for Standardization (ISO), a *record* is defined as

> "information created, received, and maintained as evidence and information by an organization or person, in pursuance of legal obligations or in the transaction of business."

Several similarities and differences exist between records management and ECM (Svärd 2014), but one common topic in both ECM research (e.g., Grahlmann et al. 2012) and records management research (e.g., Kettunen & Henttonen 2010) is metadata. In short, *metadata* can be defined as "data about data." Regarding document production, metadata provide information about the context in which the document is produced or used (Salminen 2005). In ECM, metadata are a means to facilitate information management (Tyrväinen et al. 2006), whereas in records management, metadata serve as "evidence of and information about business activities and transactions" (ISO 15489-1, 2001, p. 3). In content management, metadata are used to describe resources, manage information and intellectual property rights, and facilitate information retrieval and interoperability between information systems (Haynes 2004). Records management researchers have argued that metadata solutions for ECM can also support records management (e.g., Tough & Moss 2003).

2.1.2 ECM environment model

The development of document production practices in organizations requires an understanding of the business context in which the documents are produced and used. This is essential in the development of any document management ecosystem (vom Brocke et al. 2009), but particularly in the standardization of document production (Salminen et al. 2000; Novakovic & Huemer 2014). *Business context* refers to information that can be used to characterize the situation of a person, place, or object within a business process in a business environment (Novakovic & Huemer 2013). Important business context characteristics are geographical location, industry domain, and activity (Novakovic & Huemer 2014).

Business-context characteristics regarding records management in the public sector of Finland are business function, business case, activity, actor, and document (Arkistolaitos 2009).

Several ECM environment models have been introduced in previous ECM literature. Salminen (2005) has proposed a model that offers a means to analyze content management in the business context of an environment. In the model, activities connect actors, content items, and systems. Tyrväinen et al. (2006) have presented a widely referenced model for characterizing ECM research. The model has been adapted and extended from the model of Salminen (2005) and shows four perspectives to ECM, namely, content, technology, process, and enterprise as a context in which the content is managed. Users, information, and systems are included in the content perspective of the model (Tyrväinen et al. 2006). vom Brocke et al. (2011) and vom Brocke and Simons (2014) have adapted the model presented by Tyrväinen et al. (2006) and utilized the adapted model to connect the management of business processes and content (vom Brocke et al. 2011) and to analyze research on ECM (vom Brocke & Simons 2014). Alalwan et al. (2012) have provided also a model for analyzing the ECM literature. The model includes strategic aspects of ECM, the ECM system life cycle, and four ECM dimensions: tools, strategy, processes, and people. Rickenberg et al. (2012b) have introduced a process-driven approach for analyzing content and technology in organizations. Their model combines models presented by Tyrväinen et al. (2006) and vom Brocke et al. (2008). Grahlmann et al. (2012) have also proposed a model regarding ECM's functions.

The ECM environment model selected for this study is the model that Salminen (2005) proposed. It has been developed from the earlier EDM (electronic document management) model (Salminen et al. 2000), and both the models have been used in several case studies to analyze and describe the case environments. They have proven useful, not only as analysis tools but also as tools in the communication between researchers and people working in the case environments (e.g., Lyytikäinen et al. 2001, Honkaranta et al. 2005, Salminen & Virtanen 2005). The model is presented in Figure 1 (next page) and is part of the RASKE methodology (Salminen 2000; Salminen et al. 2000) developed for supporting the implementation of structured content production. The term *RASKE* comes from the Finnish words "*Rakenteisten Asiakirja Standardien Kehittäminen*," which means the "development of standards for structured documents." The methodology is introduced in more detail in Section 2.2.4.

The content management environment is presented in the model as a construction of two types of entities, activities and information resources, and information flows between the entities. In Figure 1, activities are depicted by the oval, and information resources by rectangles. Information flows between the resources, and activities are depicted by dashed arrows.

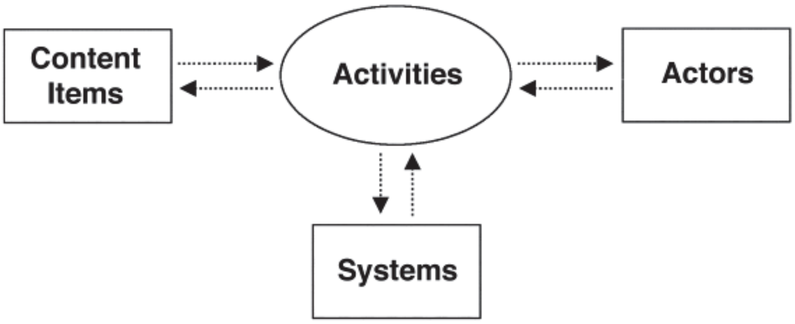

FIGURE 1 Components of the content management model (Salminen 2005).

An *activity* consists of actions that one or several actors perform during an organizational process. In the private sector, the process may be, for example, an insurance claims process or an invoicing process, and in the public sector, the process can be, for example, a legislative or budgetary process. An example of an action taken in a legislative process is the introduction of a motion.

The information resources are divided into three types according their different roles in the activities: actors, systems, and content items. An *actor* is an organization, a person in some organizational role, or a software agent acting on behalf of an organization or a person. In the legislative process, the parliament and the government are examples of organizational actors, and a member of parliament is one of a person's roles.

Systems include technical systems, such as hardware and software, but also agreed-upon and adopted standards and mandates. Software may be, for example, a word processor, a document management system, a case management system, or a records management system. XML language is an example of the standards that may be used in the environment. If document authors are a numerous and heterogeneous group of people, guidelines may be created for standardizing document-authoring practices. For example, in Finland, the Ministry of Justice has published guidelines for authoring the text and structure of a statute (*http://lainkirjoittaja.finlex.fi/*). Mandates are regulations and legislation governing the content management of the domain. For example, the content management of the Finnish legislative process is governed by the Finnish Constitution, the Administrative Procedure Act, the Act on the Openness of Government Activities, and the Archives Act, to name the most important.

Content items are addressable units of stored data, such as documents, Web pages, wiki sites, social media posts, and tweets, including information concerning the activities of the domain. Content items may be clustered in a collection, and metadata may be associated with the collection. A collection may be, for example, document storage supporting ongoing organizational processes, or it may be a document archive for long-time preservation. If metadata are accessible in the activities of the environment as content items, it is possible to divide content items in the environment into primary content items and metadata content items. Metadata content items provide information about the primary content items of a collection and about their production, storage, and use environments. For example, the metadata of primary content items may be stored in a document management system, and the metadata of production, storage, and use environments may be stored in a case management system or in a records management system. Metadata provide a glue for connecting document and process management. Both primary and metadata content items may be produced and managed as structured documents.

In this study, the content management model of Figure 1 is used as an analysis tool for studying content production strategies (Article 1), challenges in the implementation of structured document production (Article 3), content integration (Article 4), and the structured document management environment (Article 5).

2.2 XML document management

This section first introduces the key concepts of XML documents. Then, characteristics of XML document management and XML document production are described. Finally, proposed methods for XML document management are presented.

2.2.1 XML documents

XML (Bray et al. 2006) is a de facto standard for defining and representing information as structured documents. The development of XML was started in 1996 by the World Wide Web Consortium (W3C). The aim was to derive from Standard Generalized Markup Language (SGML) (ISO 8879) a restricted and simpler format for the purposes of internet communication in various application domains. The first W3C recommendation for XML was published in 1998 (Bray et al. 1998). In addition to XML, W3C published in the same year a document object model (DOM) recommendation (Apparao et al. 1998). DOM is a standardized programming interface for XML documents, and it facilitates an update of the structure and content of XML documents by software applications.

An information resource is an *XML document* if it fulfills the syntax rules defined in the XML specification. An XML document has both a logical and a physical structure. Physically, an XML document consists of one or more storage units called *entities*. Logically, an XML document is a collection of character data and markup. Character data include the meaning-carrying content of the document. Markup is constructed of declarations, elements, attributes, comments, character references, and processing instructions. Markup describes the document's storage and interchange format. Because markup is indicated in the document explicitly, it provides the possibility of exchanging information between different software applications in a standardized way.

The content, layout, and structure of an XML document can be separated. The separation of these three different facets facilitates the modular design and management of the document's architecture, content processing, and external presentations for a class of documents (Salminen & Tompa 2011). Figure 2 illustrates the facets and languages related to each facet.

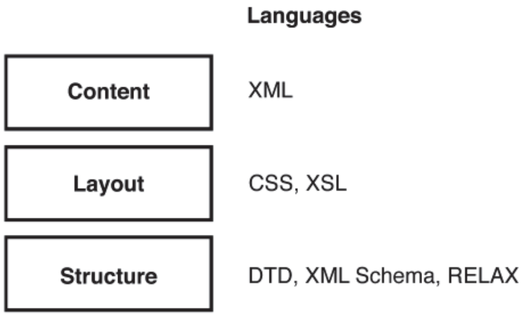

FIGURE 2 Three facets of structured documents (Salminen & Tompa 2011).

As presented in Figure 2, the content of an XML document is described using XML language. The layout and structure may be defined separately using other languages. In the following, these facets and the languages are introduced in more detail.

XML is a metalanguage for describing markup languages. Like databases, the logical structure and other constraints for a class of documents on a domain may be described by a *schema* (Salminen & Tompa 2011). An XML document is *valid* if the document complies with the constraints expressed in the related schema. The schema defines a particular markup language for the domain. The rules, which specify how information is represented in the documents of the domain, are agreed upon in standardization activities, and these rules are expressed by XML schemas. Schemas are implemented for authoring tools and other selected software before creating or manipulating XML documents in the environment. Schemas guide and control the structure of the document during its authoring.

Schemas are based on the modeling of texts by formal grammar and grammar rules. Thus, the schemas may be used to formulate (1) constraints for data input and thus for validity checking, (2) queries and text-editing operations, (3) meaningful views, (4) text transformations, (5) query optimization strategies, and (6) presentations of documents and query results (Salminen & Tompa 1999). Several general-purpose schema languages have been proposed by both standardization organizations and researchers. Examples of the former are XML DTD (Bray et al. 2006), W3C XML Schema (Thompson et al. 2004), and RELAX NG (Clark & Murata 2001). The latter ones include XDuce (Hosoya & Pierce 2003) and DSD (Klarlund et al. 2000), for example. In addition, special-purpose schema languages have been proposed for a particular type of information. An example of these is resource description framework (RDF) schema

(Brickley & Guha 2004, Manola & Miller 2004) for expressing metadata as XML documents.

If an XML document is intended to be used by humans, it has to be *rendered* into humanly perceivable external representation. The layout of XML documents on an output medium is usually specified by means of *style sheets*. W3C has published two different style-sheet languages to define a layout: Cascading Style Sheets (CSS) (Bos et al. 2011) and Extensible Stylesheet Language (XSL) (Berglund 2006). CSS was originally developed for rendering HyperText Markup Language (HTML) documents (Raggett et al. 1999) but can also be used for the simple rendering needs of XML documents on screen, in print, or in an aural medium. XSL is a language for describing the page layout and formatting of large or complex multilingual XML documents to be published in HTML, Portable Document Format (PDF), or other formats. The rendering of an XML document is a special case of XML transformation. Transformation is one of the typical operations in the structured document management environment, and it requires *mapping* between the source and target structures (Amano et al. 2014). XSLT (XSL Transformations) (Kay 2007) is a language that W3C has proposed for transforming XML documents into other XML documents, text documents, or HTML documents.

2.2.2 Characteristics of XML document management

Features of XML document management are characterized in several sources: Arnold-Moore et al. (2000b) and Salminen and Tompa (2001) describe functional needs for XML document management systems, da Graça Pimentel et al. (2009) analyze the subject from a document engineering point of view, and Salminen and Tompa (2011) consider the issue by comparing XML document management to database management. W3C has classified XML-related standards at *https://www.w3.org/standards/xml/*. In the following, the features are introduced.

Design. As early case descriptions highlight (e.g., Poulin et al. 1997a; Sandahl & Jenssen 1997), the implementation of structured document production requires design before any XML document is created in an ECM environment. Design is also one requirement for an XML document management system (Arnold-Moore et al. 2000b).

The minimum design requirement concerns a document structure resulting in an XML schema. Numerous design guides have been published for schema design by software vendors (e.g., Obasanjo 2003; Ogbuji, 2004; Khan & Sum 2006), academics (e.g., Routledge et al. 2002; Lee et al. 2009), and standardization organizations. An example of the lattermost is the public administration recommendation JHS 170 (JUHTA 2012c) in Finland. Methods for XML schema design are reviewed, for example, by Jauhiainen and Honkaranta (2006). Sedlar (2005) presents some schema design problems as well.

If humans consume the documents, the design encompasses a document layout usually resulting in a style sheet as well. Experiences with layout design are described, for example, in the case study that Kerer et al. (2001) report. Several reported case studies show that schema and layout design are interwoven activities (see, e.g., Honkaranta 2003; Salminen et al. 2004).

Depending on the domain where XML is adopted, content design may also be required, particularly if content reuse is considered. For example, in the technical publication domain, content reuse plays a central role in the adoption of XML (e.g., Sapienza 2004; S1000D; Anderson & Eberlein 2015), and content design may result in a special solution for component content management (Andersen & Batova 2015). In complex content reuse environments, content strategy (Batova & Andersen 2016; Rockley et al. 2003) may be required. If XML is adopted for documents related to business processes, master data may be used for content reuse. *Master data* refer to a single source of high-quality data that provide core business-information items to various systems (Fan et al. 2012).

Content production. XML documents may be authored by human users or generated automatically with various types of software applications. The content production alternatives are introduced in more detail in the next section. In the environment, the creation and control of content validity may be fragmented among several software applications (Salminen & Tompa 2011). Software may support only certain schema languages, and thus, the conversions of schemas to another schema language may be required in an environment. Because the structured document approach supports different content production alternatives, the implementation of structured document production may differ among ECM environments (see, e.g., Braa & Sandahl 1997).

Correctness. Compared to databases, XML documents are typically accessible through multiple independent software systems (Salminen & Tompa 2011). Hence, compliance is required for all of the systems used for XML document management. *Compliance* concerns the consistency, security, and availability of content. An access control mechanism (Kudo & Hada 2000) may be required to secure the content, or techniques for detecting the changes of XML documents (Cobena et al. 2002) may be needed.

Operations. Because XML facilitates the automatic creation and manipulation of documents, several kinds of operations must be considered in XML document management. Operations include creation, validity checking, transformation, assembly, rendering, information retrieval and browsing, publishing, and annotations. Schemas may guide the operations (Salminen & Tompa 1999). The operations may be implemented with various software applications, each designed to facilitate a particular operation.

Evolution. XML schemas evolve over time. Sedlar (2005) discusses schema evolution and versioning, along with the implications of versioned schemas for software applications operating with XML documents. Geneves et al. (2011) discuss the impacts of XML schema evolution first on the validity of existing documents, and second on applications operating with documents whose structures are described by the original schema. In the first case, schema changes concern data consistency: Existing documents may be invalid for a new version of the schema, and new documents may be invalid for some previous versions of the schema. Because schemas may guide operations (Salminen & Tompa 1999), in the second case, schema changes may concern several software applications.

Repository management. In XML content management, both document instances and document collections must be considered (Salminen & Tompa 2001). XML documents have some special features: Documents may be large, with complex structures. In addition to text, documents may contain other media types, such as images. In some environments, XML document variants may be required, for example, when there is a need to hide personal data from a published XML document but maintain the data in the document handled internally in an organization. In some cases, it might be reasonable to store one or more renditions of an XML document in a repository to address performance or consistency issues: Besides having a structured form, a document may be stored in plain text and PDF formats in a repository.

2.2.3 XML document production

From a creator-actor point of view, XML document production may be classified in automatic creation by software applications and authoring by human actors. In the literature, XML documents are typically divided into *data-centric* XML documents intended for data integration between the software applications and *document-centric* XML documents meant for human consumption (Bertino and Catania 2001). An order and an invoice are typical examples of document types that may be characterized as data-centric XML documents. These are document types that may be created in XML format automatically by software applications. A book and a record of a plenary session are examples of document-centric XML documents whose production requires human authoring. According to Bertino and Catania (2001), the structure of a document-centric XML document is more irregular than a data-centric XML document. Their content may be heterogeneous, and the meaning of the document depends on the document as a whole. Furthermore, document-centric XML documents typically contain larger sections of text, much mixed content, and less machine-readable data compared to data-centric XML documents (Bertino & Catania 2001). These are features that typically require human authoring.

There are many ways to produce XML documents, either by software applications or human actors (Braa & Sandahl 1998). Production with software applications may involve the following:

- Creating the document from database content with the export capabilities of the database system (Bertino & Catania 2001).

- Creating the content in an XML database (Salminen & Tompa 2001).

When human authoring is required, several alternatives are available for XML document production:

- Using word processors or Web browsers with XML support. For example, a Microsoft Word word processor stores document in XML format using WordprocessingML language. The document may be further converted from a WordprocessingML structure file into a custom structure. For example, see an XML-based application for ITU Telecommunication Standardization Sector (ITU-T) Recommendations at https://www.itu.int/en/ITU-T/committees/scv/Documents/T42010000020002PDFE.pdf.

- Using word processor document templates with style and transforming the content of a word processor file automatically to an XML document using the style information attached to the document (see, e.g., Braa & Sandahl, 1997). For example, a Microsoft Word .dotx document template may contain styles the author uses during authoring. A software application may be developed to map the styles to the document structure and to produce an output XML document.

- Using a custom-designed interface developed for a certain document type separately (see, e.g., Agnoloni et al. 2007 or Bacci et al. 2009). An example of such a tool is Adobe FrameMaker + XML application or SDL Xopus application.

- Using a generic syntax-directed editor that validates content with respect to the structure's definition (see, e.g., Sandahl & Jenssen, 1997 or Salminen et al., 2001). An example of such a tool is Altova XML Spy.

- Creating the content in an XML database (see, e.g., Meier, 2002).

In the simplest authoring solution, one author works with one document at a time using a tool and method selected in an XML document production environment. A more complex authoring solution is required if several human actors work with the same document, each author preparing a component of the same document. This kind of modular content production is preferred, for example, for technical documentation (e.g., Sapienza, 2004). In this domain, technical writers compose reusable content components, and the created components are constructed into a user manual in a separate content management activity. Another example of modular content production is the creation of a large document simultaneously in several organizations, for example, when each ministry prepares its own budget proposal at the same time for the State Budget Proposal in Finland. In this case, the document is divided beforehand into parts, and each organization works with its own part. Simultaneously, authoring may also take place to prepare Plenary Proceedings in parliaments: Several authors transcribe speeches given by members of parliament (MPs) into verbatim texts authored into speech documents. When all speeches are captured, a final record of the proceedings is composed of the speeches, votes, and decisions of the session. In these kinds of XML document management environments, component content management (Andersen & Batova, 2015) supports XML document production.

From the author's point of view, a shift to XML document production may be soft, guided, or enforced (Braa & Sandahl, 1998). *Soft production* refers to the authoring of text with a word processor, but a technical assistant or document editor marks up the document structure afterward. Special tools have been developed to mark up document structures according to certain XML schema of a domain (see, e.g., Bacci et al. 2009). In some domains, native XML support of a word processor may be utilized.

In *guided production*, the author also uses a word processor but marks up the content of the document with predefined styles by using a style editor included in the word processor. The author must be familiar with the allowed styles and the logical order in which the styles may be used, as the style information is the basis for the conversion of the document into XML format.

In *enforced production*, the document is authored directly in the XML format by using one of the alternatives listed above. Although XML specialists and software developers may produce XML documents with syntax-directed editors, the enforced production requires the development of custom-designed user interfaces for document authors and other end users.

2.2.4 Methods for XML document management

Methods and models supporting the adoption of the structured document approach in inter-organizational business processes have been proposed by professionals (e.g., UN/Core Components *http://www.unece.org/cefact/ codesfortrade/unccl/ ccl_index.html*) and in the academic literature (e.g., Novakovic & Huemer 2013). However, there is only a limited number of related research focusing on methods for XML document management in organizations. These methods are typically meant to support the management of data-centric documents as used in electronic business data exchange. Examples of these document types are purchase and sales orders, invoices, and shipment documents. In addition, in the public sector, interoperability frameworks (see, e.g., CS Transform 2010; Kawtrakul et al. 2011) have been established for supporting data integration to improve government services, transactions, and government interactions with citizens and businesses. The data-centric structured document approach plays a central role in these frameworks. In the academic literature, the document engineering approach (Glushko & McGrath 2005) is one widely referenced method for data-centric documents. It is meant for e-business software applications and data integration.

Regarding research related to document-centric documents, Boer (2016) and Gen et al. (2016) have studied the application of XML to legislative documents. The studies describe lessons learned from two cases and do not propose any methods for XML document management. Flanders and Jannidis (2012) discuss data modeling generally and related to XML specifically. They propose areas for further research where the relationship and role between data models and process models is one of the proposed areas. The relationship between content and processes is included in the RASKE methodology (Salminen 2005; Salminen et al. 2000) and in the process-driven content analysis method proposed by Rickenberg et al. (2012b).

Only a few published studies address XML document management methods intended for document-centric documents:

- The Maler and El Andaloussi method (Maler, E. & El Andaloussi, 1996)
- Unified Content Strategy (Rockley et al. 2003)
- RASKE (Salminen et al. 2000; Salminen 2005)
- The process-driven approach for analyzing content (Rickenberg et al. 2012b)
- Methods for XML schema design (Jauhiainen 2014)
- Component Content Management (Andersen & Batova 2015)
- Document Centric Modeling of Information Systems (Molnár & Benczúr 2015)

The first three methods listed above are introduced in more detail by Jauhainen (2014). The Maler and El Andaloussi method (Maler & El Andaloussi 1996) is meant for SGML language but is applicable also to XML. This method focuses on schema design. Unified Content Strategy (Rockley et al. 2003) provides a tool for creating a strategy for efficient content reuse. RASKE (Salminen 2005; Salminen et al. 2000) is a methodology developed for supporting the implementation of structured content production. RASKE includes the ECM model presented in Chapter 2.1.2 and a model for SGML standardization (Salminen et al. 2001). The process-driven approach for analyzing content (Rickenberg et al. 2012b) is meant for identifying, assessing, and classifying content in organizations. As in the RASKE methodology, the process-driven approach proposes guidelines and visual representations for integrating different ECM perspectives in an organization. Jauhiainen (2014) focuses on schema design and XML document management. Two of the articles included in this thesis are written in co-operation with Jauhiainen and other researchers. Anderson and Eberlein (2015) have studied the current state of component content management regarding technical documentation. Based on the literature review, they discuss content as a business asset, content strategy, structured authoring, and single sourcing. They identify processes and tools required to adopt component content management. The document-centric modeling of business information systems (Molnár & Benczúr 2015) is a theoretical framework and design method for practical applications. The modeling approach focuses on documents and their interrelationships with business processes and is based on the enterprise architecture presented by Zachman (1987).

The RASKE models are used in the study presented in this thesis, because the aim has been to enhance the RASKE methodology. This research area is important to the academic community, because there is only a limited amount of related research focusing on methods for XML document management in organizations. In the next section, the design and methods of the study are presented.

3 RESEARCH DESIGN AND METHODS

This chapter introduces the background, research methods used, and steps of research presented in this thesis. First, the context and the research approach of the study are discussed, then the research process is presented, and finally, the major case of this study is introduced.

3.1 Research approach

This study has its roots in the development of RASKE's methodology (see, e.g., Salminen 2005; Salminen 2000; Salminen et al. 2000; Salminen et al. 1997). The methodology is part of the results of digital media research at the University of Jyväskylä. The methodology, as presented in numerous research articles (e.g., in (Salminen 2005; Salminen 2000; Salminen et al. 2000; Salminen et al. 1997), provides a framework for document standardization whereby the development of document formats is considered part of the holistic development of document management environments related to business processes. The research process leading to this thesis originated in the RASKE2 research project, in which the aim was to enhance RASKE's methodology with methods for the integration of information resources by means of metadata standardization.

Within information systems research, methods are regarded as artifacts that provide guidance on how to solve business problems with computer and communication technologies (Hevner et al. 2004). In addition to methods, artifacts may be constructs, models, and instantiations (March & Smith 1995). *Design science* is an applicable approach for research in which the aim is to create artifacts that serve human purposes (March & Smith 1995). Design science has been used especially in engineering and computer science but has also been used widely in many earlier research disciplines, including the information systems discipline (Peffers et al. 2008; Järvinen 2007).

Peffers et al. (2008) have proposed a methodology for carrying out design science research, one derived from many prior design science methodologies (e.g., Gregor & Jones 2007; Hevner et al. 2004; Nunamaker et al. 1990). The methodology includes a research procedure comprising a nominal sequence of six steps in the design science research process. The steps are: (1) problem identification and motivation, (2) define objectives for a solution, (3) the design and development of the artifact, (4) demonstration, (5) evaluation, and (6) communication. *The problem identification and motivation step* includes the definition of the specific research problem and, for motivating the research, the justification of the value of a solution. *The define the objectives for a solution step* involves inferring the possible and feasible objectives of a solution. *The design and development step* concerns the creation of the artifact. *The demonstration step* provides a demonstration of the use of the artifact to solve the problem identified. *The evaluation step* contains observations and measurements against how well the artifact supports a solution to the problem identified. Finally, *the communication step* provides a presentation of the problem, a description of the artifact, its utility and novelty, the rigor of its design, and its effectiveness for researchers and other relevant audiences. Design science research is not forced to proceed in sequential order, and actually, the researcher may start at any of the first four steps and move outward (Peffers et al. 2008).

The design science research approach used in RASKE's methodology development is both problem-centered and objective-oriented. In the RASKE2 project, the idea for the research resulted from an observation of the Finnish legislative environment, where the practical problem was a lack of software-independent metadata standards and the fragmentation of metadata among 13 ministries, the parliament, and some other organizational actors involved. The identified research problem was a lack of methods supporting metadata standardization within ECM. The practical goal of the project was to identify and standardize the metadata most essential for improving legislative content management and related services, and the research goal was to provide methods for metadata standardization.

In the RASKE2 project, the special focus of the author of this thesis was the use of structured documents in metadata production and use, and the enhancement of RASKE methodology in this area. The aim of the study presented in this thesis is to increase the understanding of structured document production and to provide a framework for XML standardization to support ECM. The framework should enable the analysis and development of a structured document management environment within metadata standardization in organizations. The focus of this thesis is defined by two research questions:

> **RQ1:** How does one implement structured document production to support ECM?
>
> **RQ2:** How does one analyze and describe XML document and metadata management?

In the articles attached to this thesis, these main questions are divided into more detailed sub-questions. The research presented in this thesis is qualitative and exploratory. Moreover, the general aim of the qualitative research in the information systems discipline is to understand and explain certain social or cultural phenomena related to information systems (Myers 2007). To achieve this goal, a case study method has been used. A case study method is also used as a tool for developing artifacts. Generally, a case study is a suitable approach when the forms of research questions are "what" and "how" (Yin 1994). *A case study* is "an empirical inquiry that investigates a contemporary phenomenon within its real-life context, especially when the boundaries between phenomenon and the context are not clearly evident" (Yin 1994). The case study method is useful for a phenomenon that is broad and complex, where the existing body of knowledge is insufficient, when a holistic and in-depth investigation is needed, and when the phenomenon cannot be studied outside of its context (Paré 2004). These characteristics are typical for ECM research (Grahlmann et al. 2012; Tyrväinen et al. 2006). Within the design science approach, the case study method is one alternative for demonstrating how to use the developed artifact to solve problems (Peffers et al. 2008).

A case study is preferred for examining contemporary events, when the researcher has no possibility of manipulating and controlling relevant behaviors (Yin 1994), whereas *action research* is about making discoveries through taking action in practice (Baskerville 2008). Unlike design science and case studies, action research modifies a given reality and produces knowledge to guide changes in practice (Järvinen 2007). The researcher of this study guided the major case organization of the study within the RASKE2 project and after the project as an ECM consultant in practice. Thus, part of this research was conducted as action research, which is typically a cyclic process (Baskerville 2008). In an action research cycle, the following five steps are often regarded: (1) diagnosing, (2) planning the action, (3) taking action, (4) evaluating, and (5) specifying learning (Kock, McQueen & Scott 1997). *The diagnosing step* includes problem identification and the definition of a problem to be solved in the practice organization. *The action-planning step* involves the analysis of alternative courses of action for solving the problem. *The action-taking step* concerns the selection and realization of the selected course of action. *The evaluating step* includes the study of outcomes regarding the selected course of action. Finally, the *specifying-learning step* contains a study of the outcomes of the evaluating step (Kock, McQueaan & Scott 1997).

Action research and design science are regarded as similar research approaches (Järvinen 2007), but when conducting research, there are several differences regarding the involvement of the researcher, the role of the developed artifact, the structure of the research process, and the emphasis on the researcher's learning and the contribution to existing knowledge (Papas et al. 2012; Baskerville 2008). Peffers et al. (2008) prefer action research as a complementary paradigm for the design science approach used to design and demonstrate information systems research artifacts, if the motivation for the research is to solve problems in a specific organizational context.

In the following section, the research process of this study is presented, and an application of the introduced research approaches is described.

3.2 Research process

The study presented in this thesis includes action research cycles in the RASKE2, RAKE, and Eduksi projects, three case studies, and the development of artifacts using the design science approach. The cases are summarized in Table 1. The research is motivated by Case 1, the major case environment of this study: the Finnish Parliamentary documents in the Finnish Government and Parliament of Finland. A more detailed description of the case is provided in the next section, Section 3.3. The second case concerns an invoice of an international ICT service provider and one of its customers. Data from Case 2 are collected by participating in the Invoice Center project and interviewing the product and development managers of the Invoice Center service. Case 3 includes four document types in the Finnish Centre for Pensions, and the data are collected by participating in the RAKE research project.

TABLE 1 Case studies included in this thesis.

Case	Case description	Project
Case 1	The Finnish Parliamentary documents in the Finnish Government and Parliament of Finland	RASKE2, 2004–2006; Eduksi, 2010–2017
Case 2	An invoice at an international ICT service provider and one of its customers	Invoice Center, 2001–2004
Case 3	Four document types in the Finnish Centre for Pensions	RAKE, 2006

The research process of this study is depicted in Figure 3, where the notation follows the RASKE modeling method (Salminen 2003). The author of this thesis actually started the research process before the RASKE2 project with data collection from Case 2 when working as an XML document manager at an international ICT service provider with a participating Invoice Center project. Before this study and thesis, the data collection resulted in a master's thesis and an article regarding the visualization of XML Electronic Data Interchange (EDI) messages (Korhonen & Salminen 2003). The data collected from Case 2 are used in Article 2 of this thesis.

The author joined the RASKE2 project in 2004, when it had already been under way for one year. In its first year, the project analyzed current information management in the legislative environment from a content management point of view. The second year included an analysis of literal sources concerning metadata, particularly JHS 143 (JUHTA 2012a), a new metadata recommendation for Finnish public administration. Moreover, the interviews of 33 actors in the legislative process were conducted by the author and other researchers. The aim was to understand metadata producers, the production phase regarding the legislative process, and the use of the metadata. Research conducted by the author in the RASKE2 project resulted in Article 1 and Article 4. The ideas about how to use structured documents in metadata production and the ECM environment are presented in the final report of the project (Nurmeksela et al. 2006). Article 4 resulted from a comparison of the findings from RASKE2 and RAKE action research projects; the author participated in a RAKE project in 2006.

Wider data collection and analysis of the Finnish Parliamentary documents from 2006–2007 resulted in Article 3. During 2007–2010, the author worked as an XML document management consultant at an international ICT service provider and participated in several customer projects regarding XML document production and data integration. In 2010, the author joined the Eduksi project as an XML document management consultant and continued the data collection from Case 1, work that resulted in Article 5. The next section describes the content management environment implemented in the Eduksi project and the standardization process resulting in the current XML document management environment of Case 1.

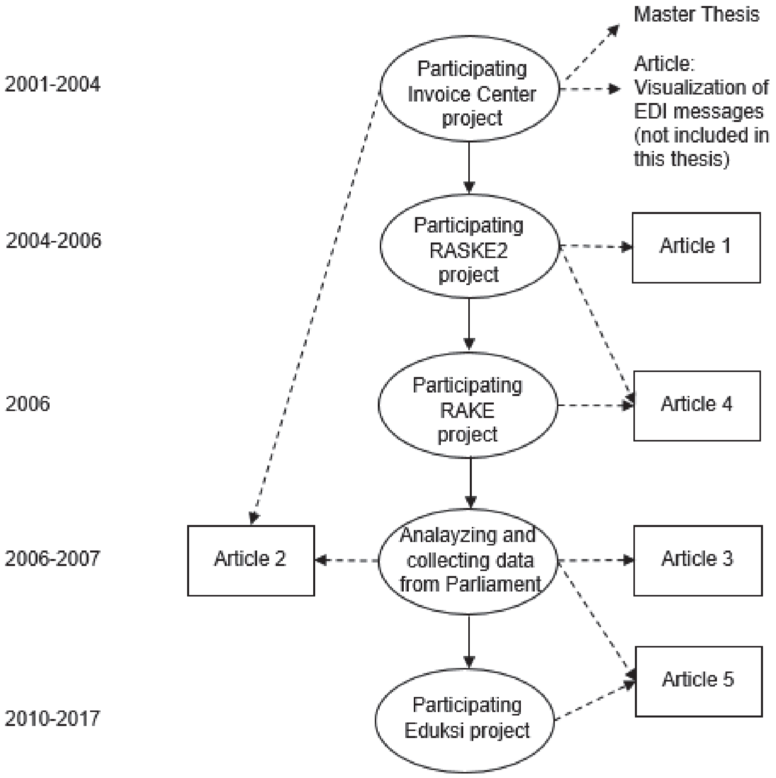
Master Thesis
2001-2004
Participating Invoice Center project
Article: Visualization of EDI messages (not included in this thesis)
2004-2006
Participating RASKE2 project
Article 1
2006
Participating RAKE project
Article 4
2006-2007
Article 2
Analayzing and collecting data from Parliament
Article 3
Article 5
2010-2017
Participating Eduksi project

FIGURE 3 Research process of this thesis.

3.3 Finnish Parliamentary documents

This section describes the complex information and content management environment of Finnish Parliamentary documents that are important at the national level. The environment constitutes the major case environment of the study presented in this thesis. The author of this thesis collected the case's data from the website of the Parliament of Finland (*www.eduskunta.fi*) and by participating in the environment's development activities, particularly the RASKE2 and Eduksi projects. Some civil servants of the Finnish Parliamentary Office have reviewed the content of this section (Appendix 3).

3.3.1 The documents and actors involved

The Finnish Parliamentary documents were handled or created as part of the work of the Parliament of Finland. These documents represent the formal record of the Parliament's debates and decisions. In Finland, the Parliament is the supreme decision-making authority consisting of 200 members of Parliament (MPs). The MPs enact legislation, approve the state budget, ratify international treaties, consider European Union (EU) matters, and oversee the government. In addition to the plenary session, MPs handle matters in 15 permanent special committees and in the Grand Committee, which focuses mainly on EU affairs. The work in the Parliament involves political debate, interaction, and the exchange of documents among the Parliament, the government, and the other actors participating in parliamentary activities.

Parliamentary documents are produced both inside and outside of Parliament by several civil servants. The decision documents of the government's plenary sessions are the most important of the latter ones. The documents are produced in their respective ministries and are proposed to the Parliament after government plenary sessions. The major document type produced outside the Parliament is the Government Proposal, which in many cases contains a bill. A special document type drafted outside the Parliament is the Citizens' Initiative, which must be considered in the Parliament if at least 50,000 citizens have signed it.

Inside Parliament, the committee secretariat provides documents associated with committee work, and the central office is responsible for documents related to plenary sessions. The central office also technically edits documents that MPs have proposed, translates key Parliamentary documents into Swedish, and is responsible for the publishing, distributing, storing, and archiving of the Parliamentary documents. The handling of 29 different matter types is documented as the Minutes of a Plenary Session and as the Minutes of a Parliamentary Committee Meeting. Decisions are communicated to interest groups within the Parliamentary Reply, Parliamentary Communication, or minutes.

Annually, thousands of documents and tens of thousands of original pages are created and published online on the Parliament´s website (www.eduskunta.fi) and in a printed format.

The series of Parliamentary documents (in Finnish, "*Valtiopäiväasiakirjat*") has been published since 1809, and it currently comprises 35 different document types. During the years, some document types were abandoned, and new ones were embraced. A minority of the document types are produced in the government, whereas most are created inside the Parliament. The most important document types are the Government Proposal, Committee Report, and Parliamentary Reply, including statute and Parliament Communication, which includes the state budget.

Finnish Parliamentary documents have been quite stable for decades. Major changes to document types resulted from EU membership in the 1990s and a new Constitution in 2000. In document production processes, major changes have resulted from the adoption of information and communication technology, which began in the 1980s with the transition from typewriters to word processors (Salminen et al. 2001). Decisions concerning technology during the 1980s and at the beginning of the 1990s were made independently in each organization participating in Parliamentary activities, and this led to inconsistent content management, incompatible tools, and uncertainty about the future usability of archived digital documents (Salminen et al. 2001). The structured document approach was seen as a solution to the problem. The next section describes how the approach has been applied to Parliamentary documents.

3.3.2 XML standardization process

The identification of document management problems and the need for an application-independent standard for digital documents activated the collaboration of the Parliament and some ministries with researchers at the University of Jyväskylä. A project named RASKE commenced in 1994. The structured document approach was the starting point. Analysis reports, articles produced during the project from 1994–1998, and articles reporting experiences from the development activities initiated in the project are accessible at the website *http://www.it.jyu.fi/raske/publications.html*. Many published articles have described those activities and their results (e.g., Salminen et al. 1996; Salminen et al. 1997; Tiitinen et al. 2000; Salminen 2000; Salminen et al. 2000; Salminen et al. 2001; Salminen et al. 2004; Salminen 2005; Nurmeksela 2007).

Researchers in the RASKE project analyzed contemporary document production in parliamentary activities, problems related to document management, and requirements for future solutions (Salminen et al. 2001). They also designed preliminary Document Type Definitions (DTDs). The work in the research-oriented project was followed by practical development projects wherein selected companies designed and implemented SGML solutions for selected document types, and the Parliament and ministries redesigned their work processes. The first implementation was the archive of laws and statutes in SGML format (*www.finlex.fi*) published by the Ministry of Justice in 1997 (Salminen et al. 2001). To test consistent authoring solutions in the Parliament and government, three document types were chosen for the pilot implementation in 1997–1998: the State Budget Proposal, Committee Report, and Committee Statement. The first of these is produced in the government, while the others are produced in the Parliament. A software company's document management consultants designed the DTDs in cooperation with people responsible for authoring the abovementioned types of documents and having extensive experience with their work. A new SGML-based budgetary system was implemented in 1998 in each ministry, and in 1999, the state budget was handled in SGML format in the Parliament. From 1998–2002, SGML was adopted for all Parliamentary document types in the Parliament. At the time, the availability of tools was limited. In the Parliament, Adobe FrameMaker was selected as the syntax-directed editor for structured document production. The government chose to use a style-based word processor solution for most document types. Adobe FrameMaker was selected as a tool for editing the State Budget Proposal as a printed publication. Reused parts of the documents, such as the statute, were transformed into a structured format in the Parliament.

An important milestone in the 2000s was the transfer from SGML to XML as the format of the State Budget Proposal in 2004. In addition, the Parliament piloted XML-based document production with one selected document type: Summary Record of a Plenary Session. Research cooperation among the University of Jyväskylä, Parliament, and some ministries continued during 2003–2006 with a research project called RASKE2, where the goal was to develop methods for the integration of information resources by means of metadata standardization. A special focus in the project was the needs of Finnish legislative work and the adoption of semantic Web technologies. Analysis reports and articles produced during the project from 2004–2006 are accessible from the website *http://www.it.jyu.fi/ raske/publications.html*. The project proposed a new semantic Web-based solution for supporting information management in business processes (Salminen & Virtanen 2005; Järvenpää et al. 2006). Computer-aided support for content management development (Lehtinen & Salminen 2006) was also used in the project.

RASKE2's work was followed in 2008 by the setting up of a practical working group where common metadata for the government and Parliament were defined (Valtiovarainministeriö 2008). In the same year, another working group proposed a unified XML document approach for Parliamentary documents and some other document types for the Government of Finland and Finnish Parliament (Rakenne 2008-työryhmä 2008). Based on the results of the working groups and the experiences of structured document production in the Parliament, selected software companies, the information retrieval experts of ministries, and experienced SGML users of the Parliament designed unified XML schemas and layouts for Parliamentary documents during 2009–2010. The design also covered other document types published in the archive of laws and statutes (*www.finlex.fi*). The author of this thesis participated in the project as an XML consultant. The key motivator for the cooperative design between the government and Parliament was to increase the consistency of the documents exchanged between the organizations and to improve content integration, reuse, and publishing. Major areas of focus were metadatas' structures and reused content, such as the statute and state budget.

In 2009, the renewal of SGML-based document production started in Parliament via the analysis and the definitions for unified XML document production and a new case and document management solution to support work in 29 different types of parliamentary procedures. At the beginning of the 2010s, the government selected a software company to design and implement a new budgetary system. The aim was to integrate into one system the financial calculations of the state budget and the production of the state budget proposal instead of error-prone manual copy and paste. The system was intended to be used in all ministries, the office of the President, and the Parliament. At the same time, the Parliament selected companies for the comprehensive renewing of case and document management systems, structured document production, and websites of the Parliament. The renewal was carried out as an Eduksi project, and the author of this thesis worked as an XML consultant at the software company that was selected to design and implement the new case and document management system and XML document production within the system. The author participated in the design, development, and deployment phases of the new solution for XML document production. Simultaneously, the government started to implement unified XML document production for Parliamentary documents with selected companies.

The production use of the new budgetary system started in 2013. The same year, the Parliament proceeded to produce two essential Parliamentary document types with XML: committee report regarding the state budget proposal and Parliament communication, including the state budget. The comprehensive XML document production of Parliamentary documents started in both the government and Parliament in the year 2015 within the deployment of Parliament's new case and document management system, Eduksi. Since spring 2015, all Parliamentary document types have been produced as XML documents and published at the renewed website *www.eduskunta.fi*. The current XML document management environment is presented in more detail in the next section. The semantic Web solution for the legislative process has also proceeded in recent years. The first attempt at publishing semantically rich Finnish laws from the archive of laws and statutes (*www.finlex.fi*) is presented in Frosterus et al. (2014).

The literature on innovation diffusion often categorizes innovation adopters into categories based on their assimilation levels, for example, innovators, early adopters, early majority, late majority, and laggers (Chen 2003). Among the adopters of the structured document approach in the public sector, the Government of Finland and the Finnish Parliament can be classified as early adopters. Many other early adopters of SGML-based standardization activities in the public sector have been reported, for example, the Norwegian Parliament and ministries (Sundholm 1998), the Supreme Court of Canada (Poulin et al. 1997b), and the Tasmanian government (Arnold-Moore et al. 2000a). These activities were also used to prepare the budget of the European Union (Catteau

1997). Later, early adopters of XML-based document production in public administration were reported, for example, in the Office of Parliamentary Counsel in South Australia (Meyer 2005), the United States Congress (Carmel 2002), and the 15 member states of the EU (Svoboda 2005) including, for example, Estonia (Heero, Puus, & Willemson 2002), Ireland (Doran 2005), and Italy (Marchetti et al. 2002). Research has shown that early adopters sometimes face tremendous obstacles because the tools and technology have not yet matured (Chen 2003).

3.3.3 Current XML document management environment

The XML document management environment in the Parliament of Finland is currently implemented as part of the Eduksi system, which is intended to support case and document management in the organization. The major information resources of the environment regarding XML document management are illustrated in Figure 4. Next, the resources shown in the figure are briefly described.

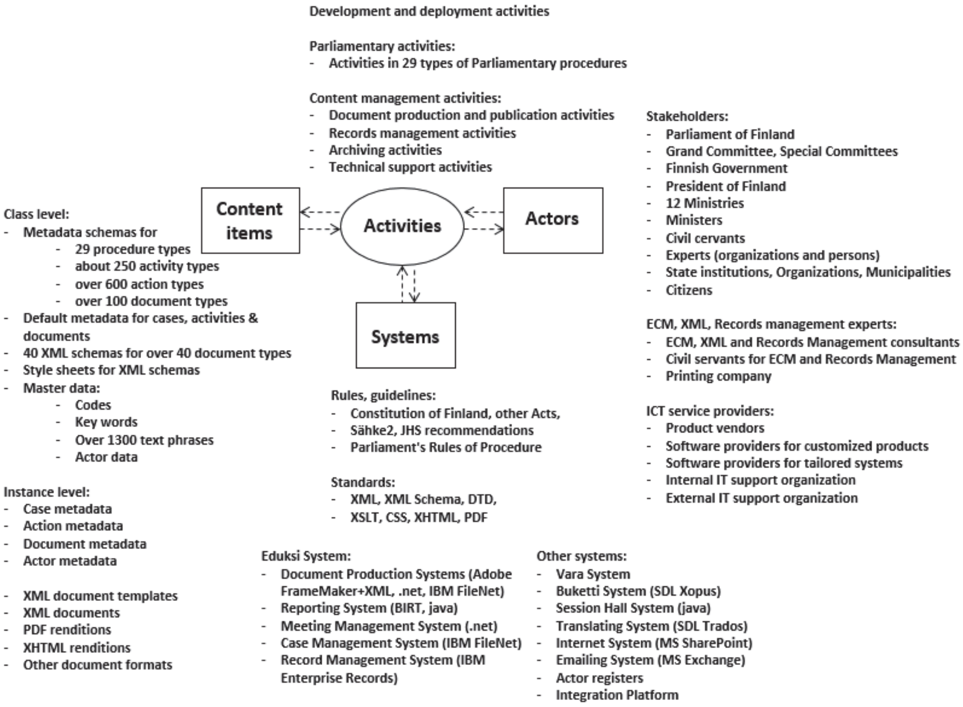
Development and deployment activities
Parliamentary activities:
- Activities in 29 types of Parliamentary procedures
Content management activities:
- Document production and publication activities
- Records management activities
- Archiving activities
- Technical support activities
Content items
Activities
Actors
Systems
Class level:
- Metadata schemas for
- 29 procedure types
- about 250 activity types
- over 600 action types
- over 100 document types
- Default metadata for cases, activities & documents
- 40 XML schemas for over 40 document types
- Style sheets for XML schemas
- Master data:
- Codes
- Key words
- Over 1300 text phrases
- Actor data
Instance level:
- Case metadata
- Action metadata
- Document metadata
- Actor metadata
- XML document templates
- XML documents
- PDF renditions
- XHTML renditions
- Other document formats
Stakeholders:
- Parliament of Finland
- Grand Committee, Special Committees
- Finnish Government
- President of Finland
- 12 Ministries
- Ministers
- Civil cervants
- Experts (organizations and persons)
- State institutions, Organizations, Municipalities
- Citizens
ECM, XML, Records management experts:
- ECM, XML and Records Management consultants
- Civil servants for ECM and Records Management
- Printing company
ICT service providers:
- Product vendors
- Software providers for customized products
- Software providers for tailored systems
- Internal IT support organization
- External IT support organization
Rules, guidelines:
- Constitution of Finland, other Acts,
- Sähke2, JHS recommendations
- Parliament's Rules of Procedure
Standards:
- XML, XML Schema, DTD,
- XSLT, CSS, XHTML, PDF
Eduksi System:
- Document Production Systems (Adobe FrameMaker+XML, .net, IBM FileNet)
- Reporting System (BIRT, java)
- Meeting Management System (.net)
- Case Management System (IBM FileNet)
- Record Management System (IBM Enterprise Records)
Other systems:
- Vara System
- Buketti System (SDL Xopus)
- Session Hall System (java)
- Translating System (SDL Trados)
- Internet System (MS SharePoint)
- Emailing System (MS Exchange)
- Actor registers
- Integration Platform

FIGURE 4 XML document management environment for Parliamentary documents.

Activities. Parliamentary documents are produced and used to handle 29 different types of Parliamentary procedures. In addition to plenary sessions, grand committee and special committee meetings are formal forums for decision-making activities in the Parliament. The documents are produced and managed in content management activities, though technical support may be needed for Parliamentary or content management activities. The environment is implemented in development and deployment activities.

Systems. The handling of matters and the creation of documents are regulated by the Constitution of Finland and numerous other acts. Sähke2 (Arkistolaitos 2009) is a norm for records that are archived only in digital format. Sähke2 and many JHS recommendations (JUHTA 2016) guide information management and the development of ECM environments in Finnish public sector organizations. For example, JHS 156 (JUHTA 2012b) includes recommendations for the registration of documents and related data in digital case management systems, and JHS 170 (JUHTA 2012c) specifies recommendations for XML schemas. Besides XML, important XML-related standards, such as XML Schema and XSLT, have been adopted in the environment. XML documents are rendered for human users utilizing XHTML, CSS, and PDF standards.

Most of the XML documents are produced within the Eduksi system using the document production system and the reporting system. In the document production system, the creation of documents is integrated with the meeting management system, the case management system, and the records management system. A document is created automatically based on data stored in the meeting management system and the case management system, and it is opened for editing in the Adobe FrameMaker+XML application. The meeting management system is a tailored .net application, and .net technology is also used for the tailored parts of the document production system. After editing, the document is stored within the case management system. The technical basis of the case management system and document repository is IBM FileNet. The records management system is used for organizing documents during the handling of Parliamentary matters and setting default metadata for the documents. The records management system is based on IBM Enterprise Records. Some of the document types may be created automatically based on data stored in the meeting management system without editing in the FrameMaker application. Examples of these document types are meeting plan, weekly plan, agenda, and meeting record. A minority of XML documents are currently produced outside the Eduksi system using the Adobe FrameMaker+XML application but are stored in the Eduksi system. The reporting system is implemented using open-source BIRT (*http://www.eclipse.org/birt/*) and Java technologies. The reporting system automatically creates data-centric XML documents using data from the case management system.

In addition to the Eduksi System, XML documents are created in two other systems. First, the Buketti system is intended for production of the State Budget Proposal. The Parliament creates its own part of the document with the system. Second, the Plenary Session System automatically creates voting reports from the data stored during a plenary session. The XML documents are published in PDF and XHTML formats on internal and external websites on the internet based on Microsoft (MS) SharePoint. Parliamentary documents are available for public use on the external website (*www.eduskunta.fi*). Invitations to a hearing session and the announcement of published documents are created automatically from the meeting management system and emailed to the experts using MS Exchange. Actor data are stored in actor registers. An Integration Platform connects the Eduksi system and other systems.

Actors. The most important stakeholders in the environment are the Parliament of Finland, the Finnish Government, and its decision-makers. Civil servants produce the Parliamentary documents in the Parliament and in 12 ministries. Expert organizations and persons are heard while considering matters in the committees of the Parliament. The President of the Republic of Finland ratifies the legislation. The main users of the Parliamentary documents are state institutions, private sector organizations, municipalities, and citizens. Numerous ECM, XML and records management experts, and ICT service providers support the work and development of the environment.

Content items. In the XML document management environment, content may be divided into class-level content items and instance-level content items.

At the class level, content items include metadata schemas, default metadata, XML schemas, style sheets for document layouts, and master data. In the environment, content items regarding cases, documents, and meetings are managed in digital format using case management, document production, records management, and meeting management systems. In Finland, if documents are archived only in digital format, systematic case management is required. *Systematic management* refers to standardized metadata regarding the cases handled, actions taken in the case-handling activities, documents received and produced in the actions, and actors related to the cases, actions, and documents. In the environment, systematic case management has required demanding metadata standardization resulting in metadata schemas for 29 different types of Parliamentary procedures. These consist of about 250 activity types, more than 600 actions, and more than 100 document types. Metadata schemas define metadata structures for a case, an action, a document, and an actor. Default metadata are defined for case, activity, and document automated metadata creation. Currently, XML document production covers more than 40 document types. Forty different XML schemas support the production of the XML documents. The layout for XHTML renditions of the XML documents is created with style sheets. The layout for PDF renditions is created with software applications. Master data include codes, key words, phrases, and actor data in two languages: Finnish and Swedish. Codes are standardized values for metadata, such as publicity class, language codes, and status codes. For key words, Parliament uses the general Finnish thesaurus YSA (*https://finto.fi/ysa/en/*) and the corresponding general Swedish thesaurus ALLÄRS (*https://finto.fi/allars/en/*). Phrases include more than 1,300 alternative phrase texts related to action types. Actor data concern stakeholders, for example, members of Parliament, members of special committees, ministries, ministers, civil servants, such as secretaries of special committees and technical assistants, and experts of hearing sessions. From an XML document management point of view, the standardization of class-level content items of the environment support content reuse and the automated creation of XML documents. In addition, standardized class-level content offers means of automating parliamentary procedures, the integration of information systems of the environment, open data and digital archiving.

At the instance level, content items include metadata content items and primary content items. Metadata content items include cases, actions, documents, and actors. Primary content items are XML document templates, XML documents produced in the environment, and PDF and XHTML renditions of the XML documents intended for human consumption. XML document templates refer to reused XML documents that include typical content of the document type, for example, titles and subtitles in relevant structures. The environment also includes documents in other document formats whose content may be reused when editing XML documents.

Figure 5 illustrates the systems involved in the creation of XML documents within the Eduksi system and the most important external systems related to XML document production. In the figure, the Eduksi system is marked with a gray color, and the information flowing between the systems is depicted with dashed arrows. In the government, Parliamentary documents are produced using the Vara System and the Buketti System and are submitted to the Parliament. Regarding XML document production, the Eduksi system also gets information from other systems of Parliament: the plenary session system and actor registers. The Eduksi system submits XML documents to the plenary session system, internet system, and emailing system. In the following graphic, an example of the automatic creation of an XML document with the Eduksi system is provided.

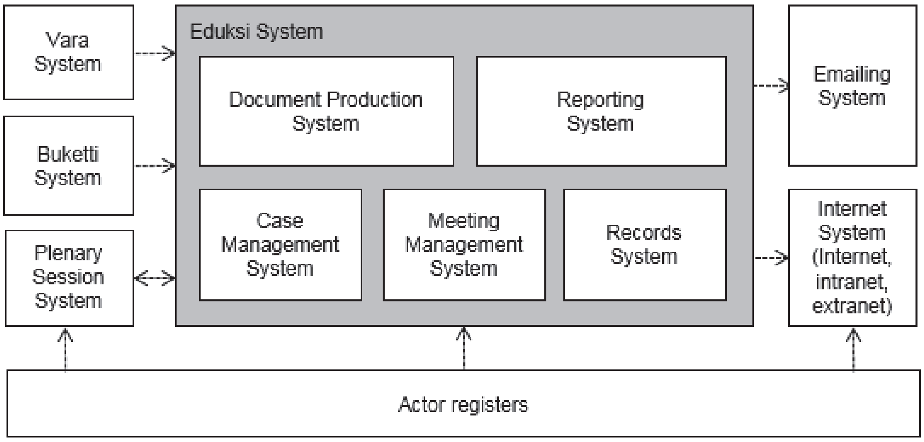

FIGURE 5 The Eduksi system and information flowing between other systems.

A committee report is one of the most important Parliamentary document types. The report covers one or several combined cases. The secretary of a special committee drafts the document for a special committee meeting using the meeting management system and the document production system. The meeting management system includes metadata regarding the case(s) and data experts heard when handling the case(s) in the committee. The secretary starts drafting the document using the meeting management system, and the system activates the document production system.

The creation of an XML document includes several steps: First, the secretary sets a value for the status code of the document and selects whether the document should include the names of the heard experts and members of the committee. Second, the secretary selects a suitable XML document template for the report or, alternatively, a previously created committee report for re-use. Third, the document production system creates an XML document combining the metadata of the case(s) and the actor data of the heard experts and members of the committee with the content of the selected XML document template or reused XML document. Fourth, the created XML document is opened in the Adobe FrameMaker+XML application for editing. The secretary may edit the structure and content of the XML document using both functionalities of the FrameMaker product and functions tailored to the Parliament. A typical example of the previous is the automatic numbering of paragraphs for the reasoning part of the committee report. Another example concerns changes that the committee may propose for the statute: Tailored functions set change codes in the statute part of the committee report. Tailored functions support the discussion and handling of the document's content in the committee meeting. After editing, the secretary selects a tailored saving function. The document production system saves the XML document and its PDF rendition to the document repository and automatically creates metadata of the document in the repository. Some metadata, such as the name of the document, are retrieved from the created XML document. In the repository, the XML document is connected to the case and action metadata and thus is managed as a record from the very beginning of the document's life cycle.

The implementation of current XML document management has required a demanding and iterative standardization process and, in each iteration, changes in work practices and new tools for document management. The next chapter introduces the attached articles of this thesis and provides a framework for XML standardization to support the implementation of XML document production in complex ECM environments, as in the case of Finnish Parliamentary documents.

4 OVERVIEW OF THE INCLUDED ARTICLES

This chapter introduces the attached articles and summarizes their key findings and contributions. At first, viewpoints of the articles are compared with ECM research perspectives (Tyrväinen et al. 2006). According to Tyrväinen et al. (2006), ECM may be studied from content, technology, enterprise, and process perspectives. The content perspective includes three different views, namely, the users, information, and systems view. Technology concerns hardware, software, standards, and other technical issues. Process may be regarded with development and deployment views. Enterprise is the organizational platform of the phenomenon, including its organizational, social, legal, and business aspects.

The implementation of structured document production in an organization is more than a technical issue (Salminen et al. 2000; Salminen 2005). Although the most essential activity in structured document production is the development of document standards (Salminen et al. 1997), the adoption of the standards may require major changes to the document production practices and tools (Salminen et al. 2000). Thus, structured document production should be studied from all the four ECM research perspectives: content, process, technology, and enterprise (Tyrväinen et al. 2006).

At a high level, this research focuses on the implementation of structured document production from the process perspective (articles 1, 2, and 3), where the development and deployment of new content management solutions in organizations are considered. Article 1 concerns the planning of an ECM solution by investigating different content production strategies. Articles 2 and 3 examine the standardization process required for the implementation of structured document production, motivators for the implementation, and changes and challenges faced during the process. In more detail, the issue is investigated from three other perspectives: Articles 3, 4, and 5 focus on the issue from the content perspective. Article 5 presents components of an XML document management environment from the technology perspective. In addition, organizational (articles 3 and 5), social (Article 3), and business (Article 1) issues are considered from the enterprise perspective. It is typical in ECM research that several viewpoints are included (Alalwan & Weistroffer 2012), as in this study.

4.1 Article 1: Content production strategies for e-government

Salminen, A., Nurmeksela, R., Lehtinen, A., Lyytikäinen, V., & Mustajärvi, O. 2008. Content production strategies for E-Government. In A.-V. Anttiroiko (Ed.), Electronic Government: Concepts, Methodologies, Tools, and Applications. Hersley, PA: Information Science Reference.

This article was published earlier in 2007 in the Encyclopedia of Digital Government.

The focus is on RQ1 and the following sub-question:
SRQ 1: What kinds of alternatives do organizations have for producing the content of their information repositories?

4.1.1 Research objectives and methods

Content production practices in an organization have a major effect on the extent to which content is accessible and how well the content supports operational efficiency and open data. Particularly in the e-governmental activities of the public sector, an objective is to get the content available on information networks, including the internet, extranets, and intranets of particular organizations. The selected production strategy affects the ways the content can be used to support e-government's goals. When planning ECM solutions, it is important to understand the alternatives for producing information assets and the consequences of the selected solution. The main objective of the article is to analyze and describe alternative content production strategies.

The analysis is based on the content management model depicted in Salminen (2005) and a case analysis of the Finnish legislative environment. Data are collected from earlier research during the long-term collaboration of researchers at the University of Jyväskylä with the Finnish Parliament and ministries within the RASKE and RASKE2 projects.

4.1.2 Content and results

Based on a literature review and an analysis of the content management environment in the Finnish legislative process, the article introduces three strategies for content production: traditional, structured, and holistic. The content production practices, benefits, and challenges of each of the strategies are evaluated. The strategies and practices are demonstrated by examples from the Finnish legislative environment.

In the *traditional strategy*, word processors, file systems, and database systems create the technological basis of content production. Word processors are used to produce documents and file systems to store them. The strategy is familiar for content producers but often requires the retyping of the same data into various systems. For software application providers, the strategy is easier, as system integration is not considered.

In the *structured strategy*, both database and document content is produced in a structured form by using SGML, HTML, or XML technologies. In the structured strategy, documents are stored as structured documents where the structure definitions, document instances, and layout specifications can be handled as separate content items. The strategy supports the manipulation and use of documents by different software applications, and it facilitates the automatic creation of new documents. Structured content production entails well-known advantages, such as rich information-retrieval capabilities, information reuse, multichannel publishing, and the long-term accessibility of information stored in documents. As a disadvantage, this strategy may require a demanding document standardization process and changes in the traditional work practices.

The *holistic strategy* focuses on systematic metadata solutions to cover the important information resources of the content management environment. The metadata may be described by using XML and Semantic Web technologies. In the holistic strategy, document and metadata schemas offer possibilities for gathering metadata from various sources automatically or semi-automatically. The strategy supports system integration, data integration, information retrieval, and the collaboration of people in work processes. As a disadvantage, this strategy requires extensive metadata standardization. The standardization of the semantic metadata may be particularly difficult. Legal information is an example of a domain where finding agreements about concepts and their relationships is extremely challenging. At the time of this research, the immaturity of Semantic Web technology and the lack of applications using Semantic Web languages, such as RDF and Web Ontology Language (OWL), were seen as technological challenges.

The contribution of this article is the introduction and evaluation of three strategical alternatives that organizations have for producing the content of their information repositories. The findings suggest that structured and holistic strategies are challenging, as they may require demanding document and metadata standardization. This motivates the other articles of this study.

4.2 Article 2: XML document implementation: Experiences from three cases

> Nurmeksela, R., Jauhiainen, E., Salminen, A., & Honkaranta, A. 2007. XML document implementation: Experiences from three cases. In Y. Badr, R. Chbeir, & P. Pichappan (Eds.), Proceedings of the Second International Conference on Digital Information Management, 224–229. Los Alamitos, CA: IEEE.

The focus is on RQ1 and RQ2 and on the following:
SRQ2: What motivates organizations in document standardization?

SRQ3: How is the standardization process realized in different kinds of organizations?
SRQ4: What kinds of changes does implementation cause in document production practices?

4.2.1 Research objectives and methods

This article describes and compares three document standardization cases focusing on the motivation of the standardization and changes in content management caused by the realization of structured document production. The cases include: the Parliamentary documents in the Government of Finland and the Finnish Parliament (Case 1), agendas and memoranda of the Faculty of Information Technology at the University of Jyväskylä (Case 2), and invoice documents in an international ICT service provider and one of its customers (Case 3). Case 1 includes both SGML and XML standardization; cases 2 and 3 are XML standardization cases. The cases fall into different categories in the use of structured documents: In the first two cases, the nature of documents is document-centric, whereas in the third case, the documents have both document- and data-centric characteristics. By the case analysis, the goal of the article is to answer both research questions of this dissertation.

The research is conducted by using a qualitative case study method (Yin 1994). The data are collected by participating in the standardization activities in the case organizations, interviewing the people involved and analyzing documents and schemas. The analysis is based on the document standardization model presented in Salminen et al. (2001). The data are collected during the years 2001–2006, and the research is carried out in the years 2006–2007.

4.2.2 Content and results

The article first introduces the modified document standardization model that Salminen et al. (2001) originally presented, and then, it uses the model as a tool for analyzing the three cases. For each case, the motivation for standardization is presented, and the realization of the phases in the standardization process is described. The phases are: analysis, schema and layout design, work process design, system design, implementation and evaluation and training.

Consistency in content management practices, the automation of business processes, and more effective content reuse were found to be the most important motivators of the adoption of structured documents, but the emphasis of the goals clearly differed in the cases. In all cases, multi-channel publishing was a central focus. As a result, the implementation of structured documents is found to be a domain-specific task related to various kinds of organizational activities, from business processes to document authoring. The implementation requires the cooperation of people and organizations, and thus, as expected, the amount and complexity of the document types as well as the number of people and organizations involved affect the challenges in the implementation process. The layout requirements had a significant impact on the schema design in each case as noted earlier, for example, in Maler and ElAndaloussi (1996) and Honkaranta (2003). Two of the cases show that it is possible to embed XML-based document production into software and to hide the markup from the authors, particularly if the documents are short and have data-centric characteristics. This was seen to lower end-user resistance. Additionally, the training of end users is an important means to reduce user resistance against structured document authoring and novel tools. If the benefits of the structured document production are demonstrated earlier for the end users, the adoption of a novel system may be quite fast and fluent, as in two of the cases.

The contribution of this article is the findings of general motivators of document standardization: multi-channel publishing, more consistent content management, process automation, and content reuse. The other well-known benefits—information retrieval, independency of particular software providers, and long-term accessibility—are not seen as common motivators for all of the cases. The realization of the standardization process is dependent on the amount and complexity of document types as well as the number of people and organizations involved. In addition to the tools used in document authoring, the implementation causes changes in the document production processes. The standardization of the Parliamentary documents is seen as the most complex case. This motivated the researcher to investigate in more detail the case as presented in the next article.

4.3 Article 3: Facing the challenges in implementing XML: The case of the Finnish Parliamentary documents

> Nurmeksela, R. 2007. Facing the challenges in implementing XML: The case of the Finnish Parliamentary documents. In M. Muñoz, A. Freitas, & P. Cravo (Eds.), Proceedings of the IASK International Conference E-Activity and Leading Technologies 2007, Porto 3–5 December, 247–255.

The focus is on RQ1 and the following:
SRQ5: What kinds of challenges may human actors face in the implementation of structured document production in a complex ECM environment?

4.3.1 Research objectives and methods

The shift to structured document production is a challenging change in an environment where various organizations are involved, the number of document types is large, the document content is complicated, and the documents are produced by human authors. These are typical features in the standardization case of the Finnish Parliamentary documents. The objectives of the article are twofold. First, the intention is to describe problems as well as their solutions during the standardization process from a human perspective. Second, the aim is to understand how the standardization impacts the work practices of document authors. Several articles were published earlier about the case, where standardization methods (e.g., Salminen et al. 2000), the impacts of the standardization (e.g., Salminen et al. 2001), and the experiences of digitalization (e.g., Salminen et al. 2004, Salminen 2005) have been described, but in these articles, minor focus has been placed on the experiences of human actors regarding the standardization process, what kind of problems they face, how the problems may be solved, and what the impacts of standardization are on the document authors.

A qualitative case study method (Yin 1994) was used in conducting the research. Different standardization approaches (Braa & Sandahl 1998) and the components of the content management model (Salminen 2005) were used as tools in the analysis. Data about experiences, problems, and solutions to the problems were collected a few years after the standardization process was realized in the case organizations. Data were collected during the RASKE2 project by interviews (Appendix 1) and a questionnaire (Appendix 2), and after the project during the years of 2006–2007 via informal discussions with a specialist who participated in the standardization process of the case.

4.3.2 Content and results

The article presents an analysis of the experiences, problems, and solutions found during the standardization of the Finnish Parliamentary documents. First, literature concerning standardization approaches from the document author's point of view is reviewed. The alternative approaches are soft standardization, guided standardization, and enforced standardization (Braa & Sandahl 1998). Then, authoring methods related to the standardization approaches are presented.

The challenges and the solutions to the problems are analyzed and reported according to the components of the content management model (Salminen 2005). First, organization and person actors are considered. This is followed by an analysis of the development and document production activities. Then, challenges regarding documents and metadata are evaluated. Finally, software systems and standards are analyzed.

As might be expected, the most remarkable problems in the document standardization process concerned documents and metadata: First, the authors are a heterogeneous group of people, and it has been difficult to define document structures that are simple enough and clear for all authors and thus used in the same way. Second, changes to the structures after implementation have caused resistance from the authors. Changes to the structures have caused changes to tools as well as existing documents. Third, the document schemas included structures for metadata to support the use of the documents, for example, internet service. In the case, some authors found the manual creation of metadata for documents to be additional, frustrating work. Challenges were also faced in three other components of the content management model: systems, actors, and activities. However, despite the challenges, the standardization of the Finnish Parliamentary documents has affected positively the interior of the organizations involved as well as the organizations at the national level, for example, as open data. In the case, the XML implementation has taken several years and has iteratively changed the work of the authors significantly. In addition, new work roles have emerged.

The major contribution of this article is the findings regarding the impacts of standardization on the positions of the document authors: The majority of the authors adopted the new work practices, whereas some continued to use the same word processor, and their technical assistants marked up the documents with new tools. A few people did not adopt the new work practices and they changed their jobs. Regarding standardization challenges, the article confirms many findings of the earlier case study (Salminen et al. 2001) and other earlier cases (Sandahl et al. 1997; Weitzman et al. 2002; Nurmilaakso et al. 2002). The same challenges are also considered later in practical guidelines regarding the adoption of XML in a legislative environment (Palmirani & Vitali 2012). One of the challenges found in this study concerned the creation of metadata, which some authors considered to be additional, frustrating work. This was one motivator for the next article, where the automated creation of metadata in document production is considered.

4.4 Article 4: Towards content integration in document production

Honkaranta, A., & Nurmeksela, R. 2007. Towards Content Integration in Document Production. In K. Soliman (Ed.), Information Management in the Networked Economy. 8th IBIMA Conference on 20–22 June in Dublin, Ireland. USA: International Business Information Management Association (IBIMA).

The focus is on RQ2 and the following sub-question:
SRQ6: How can metadata production and content reuse be automated in document production?

4.4.1 Research Objectives and Methods

In document-oriented business processes, document production requires the integration of metadata and other existing content into the document to be produced. Particularly in e-government, document types used and produced in various processes have a multitude of content sources that are needed in document production, such as law texts, names of contact persons and addresses, repeating phrases and references to legislation, and normative guidelines. In contemporary ECM environments, metadata needed in document production and content items to be reused may be fragmented into many systems. The aim of this article is to better understand metadata and content reuse needs in document production and the connection between document and metadata production.

The action research method (Kock et al. 1997; Susman & Evered 1978) is used in conducting the research in two separate cases: the Finnish legislative environment and expert environment in the Finnish Centre for Pensions regarding earnings-related pensions. The content management model depicted in Salminen (2005) is used as a tool for analyzing metadata requirements. The data were collected by participating in the standardization activities in the case organizations, interviewing the people involved, and analyzing documents.

4.4.2 Content and Results

Based on a literature review and the analysis of the document metadata recommendation for the Finnish public sector (JHS 143; JUHTA 2012a), the article first introduces a metadata classification for content production. In the classification, metadata are divided into document, process, actor, and system metadata according to the components of the content management model (Salminen 2005). Then, the article presents action research cycles on the RASKE2 and RAKE projects regarding integrated document and metadata production. Based on findings from the literature and these two research projects, patterns of document and metadata reuse are identified, and requirements for content integration in document production are proposed. Requirements are provided by an example of a business process in e-government: the process of making a statement.

Based on the analysis, *a model for integrated document production* is presented. The model consists of two separate models: the document architecture model and integrated document production process model. The *document architecture model* is composed of metadata and primary content. Metadata content is divided in the model into document, process, and actor metadata. Primary content includes reused content and new content to be authored in document production. In the *integrated document production process model*, document production is integrated with business process management and document management. In the first phase of the document production process, the document, process, and actor metadata required in document production are collected and combined from systems, and a pre-filled document is generated and provided to the document's author. During authoring, reused, primary content is collected from various data sources and provided to the author parallel with creating new content. When the document is completed, metadata needed for the systems of the domain are extracted in the metadata extraction phase. In the study, XML is used as an enabling technology for integrated document production.

The contribution of this article is the model for integrated document production. The model enhances the understanding regarding the connection between metadata and reused primary content in document production, particularly in the public sector. The model supports the implementation of an automated content production environment and content management activities, and it further offers a means for ensuring content consistency across documents and systems. XML may be used as an enabling technology.

4.5 Article 5: A life cycle model of XML documents

Salminen, A., Nurmeksela, R., & Jauhiainen, E. 2014. A life cycle model of XML documents. Journal of the Association for Information Science and Technology 65 (12), 2564-2580, DOI: 10.1002/asi.23148

The focus is on RQ2 and on the sub-questions:
SRQ7: What are the components of the XML document management environment?
SRQ8: How does one analyze and describe the XML document life cycle?

4.5.1 Research objectives and methods

The goal of this paper is to increase the understanding of XML document management in organizations and to study the XML document life cycle. The aim is to provide a model to enable the analysis and description of XML document management over the whole life of the documents. The content management model that Salminen (2005) presented is used for analysis. The proposed model utilizes the concepts of the RASKE methods (Salminen et al. 1997; Salminen et al. 2000; Salminen 2005; Salminen 2010).

The study followed the design science method and employed a nominal sequence of six steps in the design science research process, as presented by Peffers et al. (2008). Data were collected from the previous literature, observations in development projects, using domain knowledge and expert interviews. In the design and development phase of the research, the case study method (Yin 1994) was used to collect data from two cases: the State Budget Proposal of the Finnish Government and the other concerning a faculty council meeting agenda at a university. Based on the analysis of the data, earlier RASKE methods were adapted. In the demonstration phase, example documents of the cases were analyzed using the developed artifact. The artifact was evaluated by comparing the cases and collecting feedback from the case organizations.

4.5.2 Content and results

The article is structured according to the steps of the design science research process (Peffers et al. 2008). After the introduction and motivation of the study, the article describes key concepts regarding XML document management and provides *an XML document management model*. The content management model (Salminen 2005) is adapted for the model presented in the article. Next, the developed artifact, *an XML document life cycle model* with five phases, is introduced. The phases are design, content production, capture and dissemination, use, and retention. Typical activities related to the management of XML documents in each phase are described. In addition, typical actors, systems, and types of content items concerned in the activities of the phase are identified. After the introduction of the models, the use of the models is demonstrated in two case studies: one concerning the state budget proposal of the Finnish Government and the other concerning a faculty council meeting agenda at the University of Jyväskylä. The first case describes part of a complex information and content management environment having importance at the national level. The other case concerns content management in a faculty office. In both cases, one central document type of the environment has been chosen for the life cycle description. The case descriptions are divided into four parts: data gathering methods, XML document management environment, life cycle description, and impact analysis.

The key contributions of the article are two models developed for (1) the XML document management environment and (2) the XML document life cycle. The result also shows that the XML document management environment is a complex combination of various content items, processes, actors with different backgrounds, and continuously evolving systems.

4.6 About the joined articles

The articles included in the thesis are a result of the in-depth cooperation of the content management research group at the University of Jyväskylä. Article 1 is the joint work of researchers in the RASKE2 and ASG projects. Article 2 combines research from the RASKE, RASKE2, and Tag2IT projects, and a practical implementation case regarding invoice management in an international ICT service provider and one of its customers. Article 3 is based on findings from the RASKE and RASKE2 project case environments. Article 4 combines research from two separate research projects: RASKE2 and RAKE. Article 5 synthesizes the results from long-term RASKE methodology development efforts and experiences from two cases: the state budget proposal of the Finnish Government and the other concerning the meeting agenda of the faculty council at the University of Jyväskylä.

University of Jyväskylä Researcher Virpi Lyytikäinen presented the idea of Article 1, whereas the main author of the article was University of Jyväskylä Professor Airi Salminen. The author of this thesis had a minor role in the writing process, but she reviewed literature regarding methods of XML document production as well as the advantages and possibilities of the structured content in the structured and holistic strategies. In addition, she gathered information with University of Jyväskylä Researcher Antti Lehtinen about the use of the structured documents in the Finnish legislative process.

The inspiration for Article 2 is based on the professional work and research findings of the author of the thesis, as well as discussions with University of Jyväskylä Doctoral Student Eliisa Jauhiainen and University of Jyväskylä Senior Lecturer Anne Honkaranta. The article is written in cooperation with Eliisa Jauhiainen, Airi Salminen, and Anne Honkaranta. The author of this thesis was the main author and was responsible for the description of Case 3 of the article as well as the complementary data of Case 1 of the article. The comparisons of the cases and conclusions were written in cooperation with all authors.

Article 3 was written solely by the author of this thesis. She also conducted the study discussed in the article. The author obtained feedback and help with revising the language of the article from her supervisor, Airi Salminen.

Article 4 was motivated by discussions between Anne Honkaranta and the author of this thesis. They identified similar patterns of document use and requirements for content integration in document production on two separate research projects: RASKE2 and RAKE. The article was coauthored with Anne Honkaranta, who was the main author. The author of this thesis collected data about metadata requirements and was the main contributor to the literature review, metadata classification for content production, and the document architecture model proposed in the article. She was also responsible for the RASKE2 research description. The integrated document production process was coauthored with Anne Honkaranta.

The research of Article 5 was motivated by the professional work of the author of this thesis as an XML document management consultant in two customer assignments. The implementation in the assignments concerned a solution where XML document production was integrated with case and document management systems. One of the customers was a large municipality in Finland, and the other was the Finnish Parliament. The author of this thesis realized the need to analyze the life cycle of XML documents systematically to better understand the metadata needs of different activities. Airi Salminen was the main author of the article. The author of this thesis contributed to the XML document management environment model and the XML document life cycle model together with Airi Salminen. The idea to include design and content production, as well as capture activities in a document life cycle came up in discussions with Airi Salminen. The author of this thesis provided the data collection, analysis, and description of Case 1 presented in the article.

5 CONTRIBUTIONS

This chapter presents the results and contributions of the research by introducing the framework for XML standardization. The researcher took part in two research studies and two practical projects regarding three different XML standardization cases in which the data were collected. The framework consists of strategic and managerial aspects, author aspects, and models for XML standardization, and it is presented under the research questions. The models are described in more detail under the second research question.

How does one implement structured document production to support ECM?

Strategic and managerial aspects. The research shows that structured document production is a strategic choice for content production in an organization. Traditional, structured, and holistic content production strategies are presented in Article 1. In the structured strategy, documents are produced by using XML or XHTML (former SGML or HTML) technologies. The structured strategy supports the manipulation and use of documents by different software applications, facilitates the automated creation of new documents, and offers possibilities for automating metadata gathering from documents. Thus, structured document production offers possibilities for a holistic strategy in which systematic metadata solutions are considered in various activities regarding a document's life cycle. The strategy aspect is noticed also in related research on ECM (e.g., Alalwan et al. 2012) and the management of XML documents (Molnár & Benczúr 2013).

As reported in Article 2, consistency in content management practices, the automation of business processes, more effective content reuse, and multichannel publishing were the common and most important motivators for the implementation of structured document production in the investigated cases. Other motivators varied between the cases. The most important motivators are the same as the goals in the component content management in the technical documentation domain, but in that domain, single sourcing is of particular interest (Andersen & Batova 2015).

The research shows that the implementation of a structured document production environment is a domain-specific task related to various kinds of organizational activities, from business processes to document authoring. As Brocke et al. (2008) and Rickenberg et al. (2012b) suggest, the business process should be the starting point for content management. Particularly if the documents are evidence of activities in the domain, Molnár and Benczúr (2015) suggest paying attention to the design of the activities by which the documents are created, modified, and used. Thus, the implementation is more than a technical issue, particularly when the documents are authored by human users. The development and deployment requires management commitment and the cooperation of people and organizations during the standardization of content items, the design of new work practices, and the design of novel tools for document management. The implementation of the environment may be quite fast and straightforward, as in the case of the invoice center, or iterative and long-lasting, as in the case of Finnish Parliamentary documents. In the latter case, several organizations have been involved, and standardization has proceeded and iterated in different ways with the various organizations involved in the business process. The cooperation, motivation, and training of document authors in the deployment of new tools, roles, and content management practices is essential.

As expected, the case comparison in Article 2 revealed that the amount and complexity of document types as well as the number of people and organizations involved affect the challenges in the implementation process. In a complex environment, such as the Finnish Parliamentary documents, challenges may be faced regarding all of the entities of the ECM environment (Salminen 2005): activities, actors, systems, and content items of the domain, as discovered in Article 3. The same kinds of challenges, as found and reported in Article 3, are also noticed in the previous literature, where structured document production environments are analyzed (e.g., Salminen et al. 2001, Sandahl et al. 1997; Weitzman et al. 2002; Nurmilaakso et al. 2002). The results of this study indicate the need for tough leaders to improve change management regarding new roles, work processes, and novel tools.

Author aspects. If structured documents are produced by human authors, the usability of document structures and authoring tools needs special focus because structured document production differs significantly from traditional authoring. The research presented in Article 2 shows that developing custom-designed editors, where document structures are hidden from the users, increases users' acceptance of novel tools. In addition, automating document and metadata creation motivates authors, too. For example, the document-centric modeling of information systems (Molnár & Benczúr 2015) may be used when designing automated solutions. The research also shows that different authoring tools and work practices may be needed in the same business process to support variable user needs in divergent roles. For example, in the legal domain legislative drafters and legal publishers may have different tools and work practices (Boer 2014). If the benefits of the structured document production were demonstrated earlier for the authors, the adoption of a novel system may be quite fast and fluent, as reported in Article 2 regarding two of the cases. In addition, Article 3 shows that authors need motivation to commit to changes and to learn new tasks and novel tools. An earlier case report concerning Finnish Parliamentary documents (Salminen et al. 2001) also pointed out the importance of committed and reformist document authors who set good examples for other authors. However, Article 3 reveals that a few authors had difficulties with understanding the idea of structured documents and adopting new work practices, and consequently, they changed their jobs.

Models for XML standardization. As reported in Article 5, the implementation of a structured document production environment may result in a complex document management environment including varying content items, activities, actors with diverse backgrounds, and continuously evolving systems. Because of complexity, the implementation requires analysis. In this study, a previous model of a standardization process (Salminen et al. 2001) was adapted and tested as an analysis tool. In addition, three new models are presented as tools for analysis and design. The models and their contributions are described under the second research question.

How does one analyze and describe XML document and metadata management?

To answer the second research question, this study presents models aimed at enabling the analysis and development of a structured document management environment within metadata standardization in organizations. A case study (Yin 1994) and the design science approach (March & Smith 1995) are used in the development of the models, and RASKE process modeling techniques (Salminen 2000; Salminen et al. 2000) are utilized in the modeling.

The implementation of structured document production results in an XML document management environment. The environment presented in Figure 6 is introduced and demonstrated in Article 5, and it is summarized in the following. Moreover, Section 3.3.3 of this thesis demonstrates how the model may be used to analyze and describe an XML document management environment in an organization.

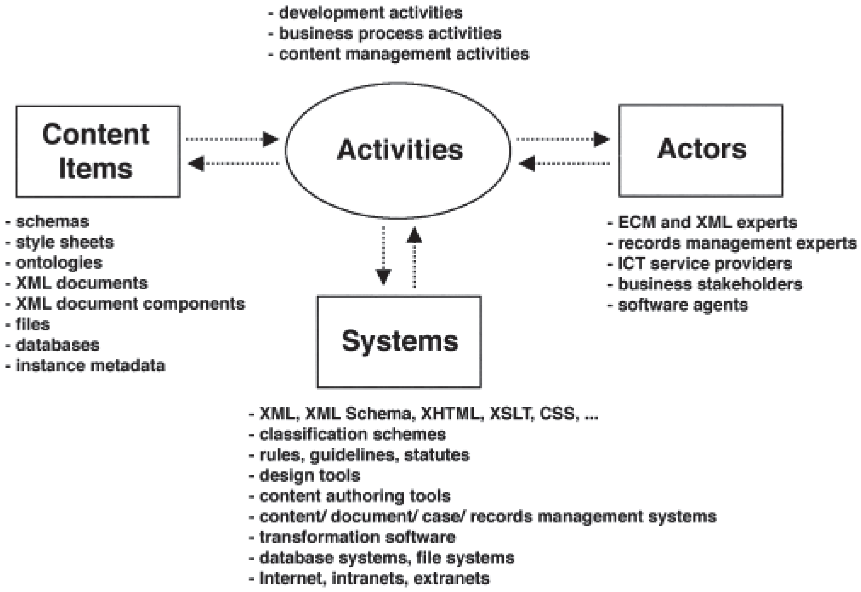
- development activities
- business process activities
- content management activities
Content Items
Activities
Actors
Systems
- schemas
- style sheets
- ontologies
- XML documents
- XML document components
- files
- databases
- instance metadata
- ECM and XML experts
- records management experts
- ICT service providers
- business stakeholders
- software agents
- XML, XML Schema, XHTML, XSLT, CSS, ...
- classification schemes
- rules, guidelines, statutes
- design tools
- content authoring tools
- content/ document/ case/ records management systems
- transformation software
- database systems, file systems
- Internet, intranets, extranets

FIGURE 6 An XML document management environment (Article 5).

In the environment, *activities* may be divided into development activities, business process activities, and content management activities. Implementation starts with development activities and results in an environment that is deployed to support business process and content management activities. The content management activities include the creation, capture, and update of documents, the creation and update of related metadata, publishing and use activities, records management activities, and archival activities. *Actors* involve organizations and experts needed for the different kinds of activities. Schemas and style sheets are class-level *content items* required for the production of XML document instances. The documents are stored as files or in a database with instance metadata. Ontologies refer to concepts used in schemas, term dictionaries, or a more complex collection of terms and their relationships. The *systems* needed in an XML document management environment consist of numerous software but also classification schemes for organizing the content units as well as rules, guidelines, and statutes regulating the domain.

The XML document management environment in an organization is a result of a standardization process. A model for the standardization process is depicted in Figure 7. The circles represent phases of the standardization process, and the arrows represent the order for starting the activities. The small black circle denotes that all of the following three activities may be started either in parallel or in any order. The model is adapted from the SGML standardization model presented by Salminen et al. (2001). The adapted model is demonstrated as an analysis tool in Article 2.

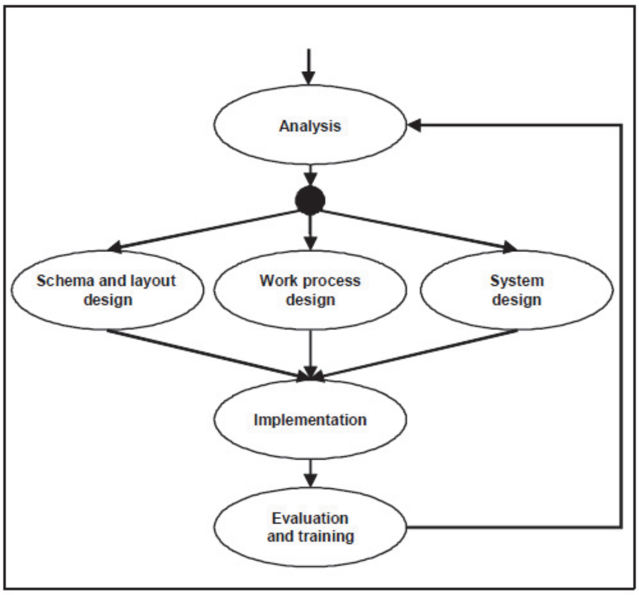

FIGURE 7 A model for the standardization process (Article 2).

The standardization process starts with an *analysis* phase resulting in descriptions and development plans of the XML document management environment of the domain. The design is composed of a *schema and layout design* and *systems design* parallel with a possible *work process design*. If the holistic strategy is selected, the schema design involves both metadata and document schemas. A layout design is required if the documents are intended for human users. Different layouts may be needed for document authors and users, as well as divergent use environments, such as websites and printed formats. Collaboration between the separate design activities and with the future document authors during the design reduces problems in the design activities, system customization, and implementation. Both the technical and organizational *implementation* of the new XML document management environment is required, possibly consisting of major changes in content management activities. *Evaluation and training* is an important phase for the successful adoption of the new environment and content management practices. In addition, an evaluation may reveal new design needs. After some operational use, the standardization process may proceed with the next iteration. In the following, three models for the analysis activities are summarized.

Many documents produced during ongoing business processes require their management and preservation as records. A life cycle of these kinds of documents is much longer than the business process where the documents are created. If the documents are preserved several years or even permanently, the documents' metadata are essential. A life cycle of XML documents and related metadata may be analyzed and described with the model presented and demonstrated in Article 5. A life cycle model of XML documents consists of five activities, as illustrated in Figure 8. The circles depict phases of the life cycle. The solid arrows represent the order for starting the activities, and the arrows indicate the output of the activity.

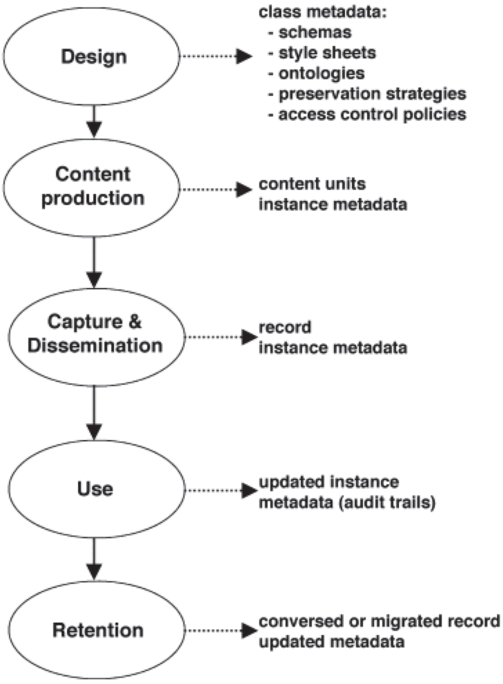

FIGURE 8 A life-cycle model of an XML document (Article 5).

The *design* includes the development and deployment of the ECM environment and solutions for XML document management. The phases of the design activity are covered in the XML standardization model (see Figure 7), and it results in the class-level metadata of XML documents, such as schemas, style sheets, and ontologies as well as preservation strategies and access control policies regarding XML documents. The operational use of the deployed XML environment comprises *content production* and *capture and dissemination* activities of the XML documents during the business process. The activities produce content unit and record instances, as well as related metadata. The *use* activity results in updated metadata during or after the business process. *Retention* refers to activities for maintaining the usability, integrity, and authenticity of the documents created originally as XML documents. It may also include activities for converting documents into XML format, if the documents are originally created and handled in other formats during business processes. Previous research has proposed several document life cycle models, particularly to support the development of content management systems (see, e.g., Molnar et al. 2015; Rickenberg et al.

2012b). Compared to other models, the life-cycle model of an XML document includes design as an important phase.

An architecture model of an XML document is proposed in Article 4 and presented in Figure 9. In the model, the logical component of an XML document architecture is depicted with a rectangle, and solid lines represent the relationship between components. The model divides the content of an XML document into metadata and primary content. Metadata describe the document, the business process, and the case where the document is created, along with actors related to the document content. The primary content is divided into reused content, existing content, and new content the author creates with an XML editor. The model may be used when designing automation for the content production and capture and dissemination activities of the XML document life cycle.

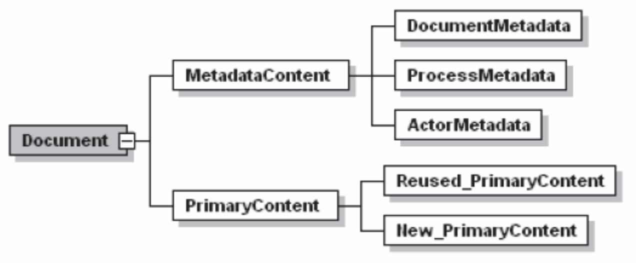

FIGURE 9 An architecture model of an XML document (Article 4).

Article 4 also presents and demonstrates the use of a model for the integrated XML document production process as presented in Figure 10. The model is based on the idea of the architecture model of the XML document (see the previous Figure 9), and it describes document production activities as part of case management activities during ongoing business processes. The circles represent the activities of XML document production, and the arrows denote the order for starting the activities. The small black circle means that both of the following two activities may be started either in parallel or in any order.

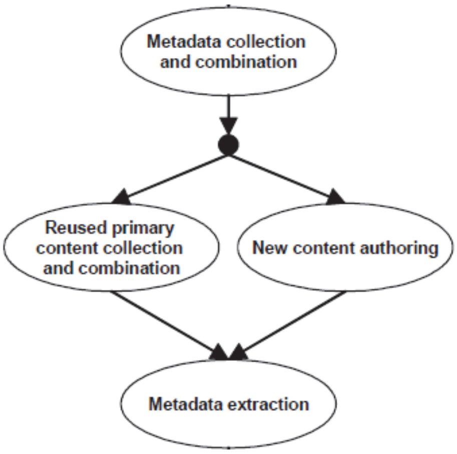
Metadata collection and combination
Reused primary content collection and combination
New content authoring
Metadata extraction

FIGURE 10 A model for integrated XML document production (Article 4).

In integrated XML document production, the document production activity is initiated by a business process activity: a need to create a document as a business action. For example, a committee statement is required for expressing the committee´s opinion of a case. Within systematic case management, most of the metadata included in the document to be produced is typically known already in the beginning of the document production process. Thus, the process begins with a *metadata collection and combination activity* where document, process and case, and actor metadata are retrieved from case and document management systems and transformed into the XML document structure. If the metadata are standardized and stored in the systems, the activity may be automated by an XML document production system, as in the case of the Finnish Parliamentary documents. The activity results in an XML document filled with metadata gathered from systems. After the first activity, the content of the XML document is completed by a document author in two kinds of activities: *reused primary content collection and combination* and *new content authoring* activities. Reuse may consider, for example, text phrases or statute text proposed by the government. When the authoring is done, metadata may be created automatically and gathered from the document in a *metadata extraction* activity. An example of gathered metadata is a document title. The model for integrated XML document production may be used particularly in the analysis and design of document production processes where the documents are authored by human users, but it also is applicable for data-centric documents created automatically by software applications.

Models for XML documents are also proposed in the previous literature. Molnár and Benczúr (2015) have introduced a general model of an XML document and a multi-dimensional model for the interaction of information systems and documents. Molnar et al. (2015) have proposed a conceptualization of the document management domain that is based on the ISO 82045 family of standards. Additionally, in these models, primary and metadata content items are divided from each other, but in the model presented in this thesis, both primary and metadata content are categorized into subcategories, such as document, process and actor metadata, and primary and new content.

This chapter presented the framework for XML standardization, including developed models to support the implementation of a structured document production environment in an organization. Earlier research on the area has been rare, and this research has contributed to enhance knowledge on the XML document management area. More importantly, the research produced several artifacts that may help practitioners to develop, deploy, and maintain structured document production environments within practical ECM solutions. In the next section, the implications of this research and the limitations of the study are discussed.

6 DISCUSSION

This chapter contains the theoretical and practical implications, as well as the limitations of the research presented in this thesis. This research was positioned in the ECM area, in which documents and other content produced and used in organizations are considered. This study was focused on the implementation of structured document production to support document and metadata management. XML and its predecessor, SGML, were considered enabling technologies for structured documents. The study followed case study (Yin 1994) and design science (Peffers et al. 2008) methods. For the study, models and modelling techniques of the RASKE methodology (Salminen 2005; Salminen 2000; Salminen et al. 2000; Salminen et al. 1997), particularly the content management model (Salminen 2005), were used as research tools.

The study consisted of three case studies in different organizations. The case studies were carried out in two research projects and two practical development and deployment projects where the author of this thesis worked as an XML document management consultant. The case organizations were important electronic government (e-government) organizations in the national legislative process in Finland, an international ICT company and one of its customers, an expert organization in the e-government field.

The study proposed a framework for XML standardization consisting of strategic and managerial aspects, author aspects, and models for XML standardization. A model for the standardization process was adapted from the previous SGML standardization model (Salminen et al. 2001). The models proposed in this thesis included an XML document management environment model, an XML document life cycle model, an XML document architecture model, and a model for an integrated XML document production process. The results of the studies were reported in five articles, which are attached as appendices to this thesis. The use of the models were demonstrated in the included articles and in Chapter 3.3.3 of this thesis.

Theoretical implications. The study presented in this thesis increases knowledge concerning the rare empirical research of ECM implementations (Alalwan & Weistroffer 2012) by describing the implementation of XML document production environments in four case organizations. The study also enhances comparatively slight knowledge of successful ECM implementations (Usman et al. 2009). According to Grahlman et al. (2012), the major themes in ECM research regarding the functionalities of ECM solutions focus on content workflow, repositories, and services for content management and use. This research provides more knowledge concerning the content production, capture, and component management functionalities of ECM solutions.

A literature review revealed that there is a lack of knowledge of a holistic viewpoint for the implementation of structured document production in an organization. The previous research has considered various aspects, such as authoring methods (e.g., Braa & Sandahl 1998), tools (e.g., Georg et al. 2007), and XML as a communication technology (e.g., Salminen & Tompa 2011), but research from organizational and implementation viewpoints are limited, as Salminen and Tompa (2011) have noticed. The existing knowledge comprises methods for XML document management, such as the process-driven approach for analyzing content (Rickenberg et al. (2012b), the RASKE methodology (see, e.g., Salminen 2005), and unified content strategy (Rockley et al. 2003), but the methods are limited to the analysis phase of the document standardization process. Moreover, design issues (e.g., Poulin et al. 1997a; Sandahl & Jenssen 1997) and particularly schema design methods (e.g., Routledge et al. 2002; Lee et al. 2009, Jauhiainen 2014) have been proposed. In addition, a theoretical framework for modeling documents within information systems has been proposed (Molnar & Benczur 2015). Even though structured content and structured authoring have evolved in practical ECM solutions, knowledge in this area is limited (Andersen & Batova 2015).

The limitation of this research is that it does not test existing theory or create new theory. However, this study provides a point of departure for future research in the implementation of a structured document production environment in an organization. The models presented in this thesis provide tools for the research community in analyzing and comparing XML document management in an organizational context, the implementation of the environment, and the iterative development of the environment. The models should help researchers in achieving a better understanding of the characteristics of the XML document management environment in organizational settings and iteratively the development of the environment. The models should also help to innovate further research ideas.

Practical implications. This study revealed that document standardization may be a long-lasting process, particularly if the number of document types is large and the environment includes various activities and stakeholders. However, in more streamlined processes where the number of document types is limited, the implementation of structured document production may be fast and fluent. For practitioners, this result should support the management of development projects in different kinds of organizational settings. The adopted XML standardization model should help practitioners in the planning and management of development and deployment projects.

This study shows that in the environments where the documents are authored by human beings, structured document production must be regarded as a strategic choice that requires management and end-user commitment during the standardization process. The standardization has a major impact on the work practices of document authors, and various challenges may be faced regarding all of the entities of the ECM environment (Salminen, 2005): activities, actors, systems, and content items of the domain. The models presented in this thesis should help practitioners to develop, deploy, and maintain structured document production environments within practical ECM solutions. The models should support the planning, analysis, and design of new solutions and new work practices.

This study suggests the connection of XML document production to document and records management, as well as to case and business process management. Katuu (2011) reported a similar finding regarding ECM implementations in current organizations covering integrated document, records, and business process management. For practitioners, the models for integrated content production should help to develop document production solutions that are integrated with various ECM systems facilitating automated document and metadata creation.

Research on document production is rare, and it is also a gray area in practical ECM solutions. A life cycle of content to be managed in an ECM solution might start from the capture phase of existing content, and the creation activity of the content is not a central focus. This study indicates that, if documents are produced as structured documents, document production is part of an ECM solution. Furthermore, according to the findings of this study, the use of the structured document approach may automate content capture and thus streamline content management and improve organizational performance in content-intensive business processes.

Evaluation. The aim of this study was to increase the understanding of structured document production and to provide a framework for XML standardization by analyzing XML implementation cases. An objective evaluation of the study is hard or even impossible, as the implementations are the results of unique processes in unique organizational settings. In these kinds of knowledge-intensive and relatively rare development processes, competence and other attributes of participating actors impact the results and perceptions of the results. The framework developed in this thesis enables the analysis and development of a structured document management environment in an organization. The use of the framework is demonstrated in the included articles, and thus, the first verification of the results is done. However, more testing of the developed models is needed.

Limitations and avenues for further research. The study presented in this thesis focuses on document-centric documents that are authored by human users during ongoing business processes but that require their management and preservation as records. These are typical characteristics of the documents produced in the public sector as evidence of activities in the domain. The results of this study could be compared to those in private sector domains, for example, document-centric document production in the insurance sector or in the private sector's procurement process.

The structure of the documents considered in this study is controlled by a custom schema. The schema development has been a demanding activity in the standardization processes of the analyzed cases. It would be interesting to study structured document production environments where international document standards of the domain are adopted in an organization for document-centric documents. For example, in the legislative domain, the adoption of the Akoma Ntoso standard (*http://www.akomantoso.org/*) for parliamentary, legislative, and judiciary documents (Palmirani and Vitali 2011) and metadata included in the vocabulary (Barabucci et al. 2009, Barabucci et al. 2010) could be studied.

One limitation is the major case environment of this study: Among the adopters of the structured document approach in the public sector, the Government of Finland and the Finnish Parliament may be categorized as early adopters. The research has shown that early adopters sometimes face tremendous obstacles because their tools and technology have not matured yet (Chen 2003).

In the future, the proposed framework should be tested in other research contexts and implementation projects where the utility of the framework could also be verified in other empirical studies. In addition, improvements for the framework could be proposed. For example, the framework could include more detailed models for content creation processes and how structured authors and document editors work in these processes. As Semantic Web solutions evolve, such as in the legal domain (see, e.g., Casanovas et al. 2016), the framework could also be enhanced to support metadata standardization.

YHTEENVETO (FINNISH SUMMARY)

Merkittävä osa organisaation informaatioresursseista koostuu dokumenteista. Organisaatioiden sisällönhallinnan (Enterprise Content Managemet, ECM) avulla pyritään toteuttamaan systemaattisia ratkaisuja dokumenttien ja muun sisällön yhtenäiseksi hallintakäytännöksi organisaatioissa. Sisällönhallintaratkaisun toteutus edellyttää usein sekä uuden teknisen ratkaisun että uusien toimintatapojen kehittämistä ja käyttöönottoa. Yksi mahdollinen tekniikka sisällön tuottamiseen on XML (Extensible Markup Language), joka mahdollistaa dokumenttien ja muun sisällön hallinnan avoimessa, rakenteisessa muodossa. XML:n käyttöönotto voi kuitenkin edellyttää mittavaa standardisointiprosessia, merkittäviä muutoksia työkäytäntöihin sekä uusien työkalujen käyttöönottoa.

Tutkimuksen tavoitteena oli tuottaa lisää tietoa rakenteisesta asiakirjatuotannosta ja kehittää viitekehys tukemaan XML:n käyttöönottoa ja standardisointityötä organisaatioissa. Painopiste oli dokumenteissa, jotka tuotetaan organisaation toimintaprosessin aikana todentamaan tapahtunutta toimintaa, esimerkiksi lakien säätämistä. Tutkimusmenetelminä olivat tapaustutkimus (case study) ja suunnittelututkimus (design science) Tutkimuksessa tarkasteltiin kolmea eri tapausorganisaatiota kahdessa sisällönhallinnan tutkimusprojektissa ja kahdessa käytännön sisällönhallintaratkaisun kehittämis- ja käyttöönottoprojektissa. Tutkimuksen tekijä osallistui tapausorganisaatioiden projekteihin, joista kahdessa hän toimi XML-dokumenttien hallinnan konsulttina. Tapausorganisaatiot olivat merkittäviä suomalaisia kansalliseen lainsäädäntöprosessiin osallistuvia toimijoita, kansainvälinen IT-palvelujen tarjoaja ja yksi sen asiakasyrityksistä, sekä julkishallinnon asiantuntijaorganisaatio.

Tutkimuksessa kehitetty viitekehys XML:n käyttöönottoon muodostuu strategisista ja johtamisen aspektista, sisällöntuottajan aspektista ja XML:n käyttöönottoa tukevista standardisointiprosessin, XML-dokumenttien hallintaympäristön, XML-dokumentin elinkaaren, XML-dokumentin arkkitehtuurin, ja integroidun XML-dokumenttituotannon malleista. Tulokset on raportoitu viidessä tieteellisessä artikkelissa. Mallien käyttöä on havainnollistettu artikkeleissa tutkimuksen tärkeimmän tapausympäristön, valtiopäiväasiakirjojen, avulla ja vertailemalla sitä muihin tapausympäristöihin.

Tutkimuksessa havaittiin, että XML-dokumenttien hallintaympäristö on monimutkainen rakennelma erilaisia sisältöjä, prosesseja, eri taustaisia toimijoita ja muuttuvia järjestelmiä. Rakenteinen asiakirjatuotanto on strateginen valinta, jonka kehittäminen ja käyttöönotto edellyttää johdon ja sisällöntuottajien sitoutumista. Rakenteinen asiakirjatuotanto eroaa merkittävästi tavanomaisesta tekstinkäsittelyohjelmalla tuotettavasta sisällöstä, joten työkalujen käytettävyyteen on kiinnitettävä erityistä huomiota. Tutkimus osoittaa, että XML-asiakirjojen rakenteisuuden kätkeminen sisällöntuottajilta, räätälöityjen sisällöntuottamisratkaisujen kehittäminen ja dokumenttien ja niitä kuvailevien metatietojen tuottamisen automatisointi lisäsivät uusien työkalujen ja työkäytäntöjen omaksumista.

REFERENCES

Agnoloni, T., Francesconi, E. & Spinosa, P., 2007. xmLegesEditor: an opensource visual XML editor for supporting legal national standards. In Proceedings of the V legislative XML workshop (pp. 239-251).

Alalwan, J. A. & Weistroffer, H. R. 2012. Enterprise content management research: a comprehensive review. *Journal of Enterprise Information Management, 25*(5), 441 – 461. doi: 10.1108/17410391211265133

Amano, S., David, C., Libkin, L. & Murlak. F. 2014. XML schema mappings: Data exchange and metadata management. *Journal of the ACM, 61*(2)

Andersen, R. & Batova, T. 2015. The current state of component content management: An integrative literature review. *IEEE Transactions on Professional Communication, 58*(3), pp.247-270.

Anderson, R. D. & Eberlein, K. J. 2015. Darwin Information Typing Architecture (DITA) Version 1.3. Part 3: All-Inclusive Edition. http://docs.oasis-open.org/dita/dita/v1.3/dita-v1.3-part3-all-inclusive.html

Apparao, V., Byrne, S., Champion, M., Isaacs, S., Jacobs, Le Hors, A., Nicol, G., Robie, J., Sutor, R., Wilson, C. & Wood, L. 1998. Document Object Model Level 1 https://www.w3.org/TR/1998/REC-DOM-Level-1-19981001/

Arkistolaitos. 2009. SÄHKE2-määräys. Sähköisten asiakirjallisten tietojen käsittely, hallinta ja säilyttäminen. http://www.arkisto.fi/fi/palvelut/normitmaeaeraeykset/saehke2-maeaeraeys

Arnold-Moore, T., Clemes, J. & Tadd, M. 2000a. Connected to the law: Tasmanian legislation using EnAct. *Journal of Information, Law and Technology, 1*, pp.00-1

Arnold-Moore, T., Fuller, M. & Sacks-Davis, R. 2000b. System Architectures for Structured Document Data. *Markup Languages, 2*(1), 11-39.

Bacci, L., Spinosa, P., Marchetti, C., Battistoni, R., Florence, I., Senate, I. & Rome, I. 2009. Automatic mark-up of legislative documents and its application to parallel text generation. In Proc. of LOAIT Workshop (pp. 45-54).

Barabucci, G., Cervone, L., Palmirani, M., Peroni, S., Vitali, F. 2009. Multi-layer markup and ontological structures in Akoma Ntoso. In Proceeding of the International Workshop on AI approaches to the complexity of legal systems II (AICOL-II)., Rotterdam, The Netherlands.

Barabucci, G., Cervone, L., Di Iorio, A., Palmirani, M., Peroni, S. and Vitali, F., 2010. Managing semantics in XML vocabularies: an experience in the legal and legislative domain. In Proceedings of Balisage: The markup conference (Vol. 5).

Baskerville, R. (2008). What Design Science Is Not. *European Journal of Information Systems, 17*, 441-443.

Batova, T. & Andersen, R. 2016. Introduction to the Special Issue: Content Strategy—A Unifying Vision. *IEEE Transactions on Professional Communication, 59*(1), pp.2-6.

Berglund, A. 2006. Extensible Stylesheet Language (XSL) Version 1.1. Available at http://www.w3.org/TR/xsl11/

Bertino, E. & Catania, B. 2001. Integrating XML and databases. *IEEE Internet Computing, 5*(4), pp.84-88.

Blair, B. 2004. An enterprise content management primer. *Information Management Journal, 38*(5), pp. 64-6.

Boella, G., Humphreys, L., Martin, M., Rossi, P. & van der Torre, L. 2011. Eunomos, a legal document and knowledge management system to build legal services. In International Workshop on AI Approaches to the Complexity of Legal Systems (pp. 131-146). Springer Berlin Heidelberg.

Boer, A. 2014. Legislation as Linked Open Data: Lessons from MetaLex XML. Amsterdam: Leibniz Center for Law, University of Amsterdam.

Bos, B., Celik, T., Hickson, I. & Wium Lie, H. 2011. Cascading Style Sheets Level 2 Revision 1 (CSS 2.1) Specification. Available at http://www.w3.org/TR/CSS2/

Braa, K. & Sandahl, T.I. 1998. Approaches to standardization of documents, in T. Wakayama, S. Kannapan, C. M. Khoong, S. Navanthe, and J. Yates, (Eds.). Information and Process Integration in Enterprises: Rethinking Documents. Norwell (MA): Kluwer Academic Publishers, 125-142.

Brahmia, Z., Grandi, F. & Bouaziz, R. 2016. Changes to XML namespaces in XML schemas and their effects on associated XML documents under schema versioning. In 2016 Eleventh International Conference on Digital Information Management (ICDIM).

Bray, T., Paoli, J. & Sperberg-McQueen, C. M. 1998. Extensible Markup Language (XML) 1.0. W3C Recommendation. W3C Consortium. http://www.w3.org/TR/1998/REC-xml-19980210

Bray, T., Paoli, J., Sperberg-McQueen, C. M., Maler, E., Yergeau, F., & Cowan, J. 2006. Extensible Markup Language (XML) 1.1. (2nd Edition) W3C Recommendation. 1st version Feb 2004. W3C Consortium. http://www.w3.org/TR/xml11/

Brickley, D., & Guha, R., V. 2004. RDF Vocabulary Description Language 1.0: RDF schema. W3C Recommendation. Available at http://www.w3.org/TR/rdf-schema

Carmel, J. 2002. Drafting Legislation Using XML at the U.S. House of Representatives. http://xml.house.gov/drafting.htm.

Casanovas, P., Palmirani, M., Peroni, S., van Engers, T. and Vitali, F., 2016. Semantic Web for the Legal Domain: The next step. Semantic Web, (Preprint), pp.1-15.

Catteau, T. 1997. The European Union's budget: SGML used to its full potential. Conference Proceedings of SGML'97 US (pp. 645-653).

Chatvichienchai, S., Lin, J., & Tanaka, K. 2015. Generating an XML-Based Search Index for an Effective Search of Office Documents. *International Journal Of Business And Information, 2*(2).

Chen, M. 2003. Factors affecting the adoption and diffusion of XML and Web services standards for e-business system. *International Journal of Human-Computer Studies, (58)*, 259-279.

Clark, J. & Murata, M. 2001. RELAX NG specification. Available at http://www.oasis-open.org/committees/relax-ng/spec-20011203.html

Cobena, G., Abiteboul, S. & Marian, A., 2002. Detecting changes in XML documents. In Data Engineering, 2002. Proceedings. 18th International Conference on (pp. 41-52). IEEE.

CS Transform. 2010. e-Government Interoperapility. A comparative analysis of 30 countries.

Doran, P. 2005. Legislative XML in the Irish Parliament: our reasons for, and experience of, migration to XML. In Proceedings of the third Workshop on Legislative XML, 22-31.

Fan, W., Li, J., Ma, S., Tang, N. & Yu, W. 2012. Towards certain fixes with editing rules and master data. *The VLDB Journal – The International Journal on Very Large Data Bases, 21*(2), pp.213-238.

Fan, H., Liu, J. & Deng, K. 2016. Towards a Composite XML Schema Matching Approach Using Reference Ontology. In 3rd International Conference on Information Science and Control Engineering (ICISCE), 2016 (pp. 724-728).

Feki, J., Ben Messaoud, I. & Zurfluh, G. 2013. Building an XML document warehouse. *Journal of Decision Systems, 22*(2)

Flanders, J. & Jannidis, F. 2012. Knowledge Organization and Data Modeling in the Humanities.

Frosterus, M., Tuominen, J. & Hyvönen, E. 2014. Facilitating Re-use of Legal Data in Applications-Finnish Law as a Linked Open Data Service. In JURIX (pp. 115-124).

da Graça Pimentel, M., Bulterman, D.C. & Soares, L.F.G. 2009. Document engineering approaches toward scalable and structured multimedia, web and printable documents. *Multimedia Tools and Applications 43*(3), pp.195-202.

Gen, K., Akira, N., Makoto, N., Yasuhiro, O., Tomohiro, O. Katsuhiko, T. 2016. Applying the Akoma Ntoso XML schema to Japanese legislation. *Journal of Law, Information and Science, 24*(2), (pp. 49-70).

Geneves, P., Layaïda, N. and Quint, V., 2011. Impact of XML schema evolution. *ACM Transactions on Internet Technology (TOIT) 11*(1), 4.

Georg, G. & Jaulent, M.C. 2007. A document engineering environment for clinical guidelines. In Proceedings of the 2007 ACM symposium on Document engineering (pp. 69-78).

Glushko, R.J. & McGrath, T. 2005. Document engineering. Cambridge: Mit Press.

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

Grahlmann, K. R., Helms, R. W., Hilhorst, C., Brinkkemper, S. and van Amerongen, S. 2012. Reviewing enterprise content management: A functional framework. *European Journal of Information Systems, 21*(3), 268-286.

Gregor, S., & Jones, D. 2007. The anatomy of a design theory. *Journal of the Association for Information Systems, 8*(5), 312–335.

Hausmann, V. & Williams, S.P. 2015. Social Business Documents. Procedia Computer Science, 64, pp. 360-368.

Haynes, D. 2004. Metadata for information management and retrieval. London: Facet.

Henttonen, P. 2009. A comparison of MoReq and SÄHKE metadata and functional requirements. *Records Management Journal, 19*(1), 26-36.

Heero, K., Puus, U. & Willemson, J. 2002. XML based document management in Estonian legislative system. Tallin: Institute of Cybernetics at Tallin Technical University.

Herbst, A., Simons, A., vom Brocke, J., Müller, O., Debortoli, S., & Vakulenko, S. 2014. Identifying and characterizing topics in enterprise content management: a latent semantic analysis of vendor case studies. In *Proceedings of the 22th European Conference on Information Systems*, Tel Aviv 2014.

Hevner, A., March, S., Park, J., & Ram, S. 2004. Design Science in Information Systems Research. *MIS Quarterly 28*(1), 75-105.

Honkaranta, A. 2003. Evaluating the ʻgenre lens' for analyzing requirements for content assembly. In K. Siau, J. Krogstie & T. Halpin (Eds.). The 8th Caise/IFIP8.1 international workshop on evaluation of modeling methods in systems analysis and design (EMMSAD ʻ03), Velden/Klagenfurt, Austria. June 16–17

Honkaranta, A., Salminen, A., & Peltola, T. 2005. Challenges in the redesign of content management: A case of FCP. *International Journal of Cases on Electronic Commerce 1* (1), 53-69.

Hosoya, H. & Pierce, B. C. 2003. XDuce: A typed XML processing language. ACM Transactions on Internet Technology 3(2), 117–148.

Ikhsan, R. F. & Hasbi, M. 2016. Implementation of relational algebra operations for web. In 2016 International Conference on Computational Intelligence and Cybernetics.

Di Iorio, A., Peroni, S. Poggi, F. & Vitali, F. 2014. Dealing with structural patterns of XML documents. *Journal of the Association for Information Science and Technology 65*(9), 1884–1900.

ISO 8879. 1986. Information processing -- Text and office systems -- Standard Generalized Markup Language (SGML)

ISO 15489-1. 2001. Information and documentation—Records management. Part 1: General.

Jauhiainen, E. 2014. Deployment of XML for office documents in organizations. Jyväskylä licentiate thesis in computing 16. Jyväskylä: Jyväskylä University.

Jauhiainen, E. & Honkaranta, A. 2006. A Review on XML Document Schemas and Methods for Schema Design. In BIS (pp. 201-214).

JUHTA. 2016. JHS- Public Administration Recommendations. Advisory Committee on Information Management in Public Administration. http://www.jhs-suositukset.fi/web/guest/jhs

JUHTA. 2012a. JHS 143 Asiakirjojen kuvailun ja hallinnan metatiedot http://www.jhs-suositukset.fi/suomi/jhs143

JUHTA. 2012b. JHS 156 Asiakirjojen ja tietojen rekisteröinti sähköisen asioinnin ja asiankäsittelyn tiedonhallinnassa. http://www.jhs-suositukset.fi/suomi/jhs156

JUHTA. 2012c. JHS 170 XML- schemas for the public administration. Version: 1.2. http://docs.jhs-suositukset.fi/jhs-suositukset/JHS170_en/JHS170_en.html

Jung, M. R., Oh, S-L.; Yim, J.H. 2016. Effects of Adopting the Open Document Format in Public Records Management. In *Journal of Records Management & Archives Society of Korea 16*(2). Records Management & Archives Society of Korea.

Järvenpää, M., Virtanen, M., & Salminen, A. 2006. Semantic portal for legislative information. In Proceedings of the Fifth International EGOV Conference, Krakow, Poland, September 4-8, 2006. Wien, New York: Springer Verlag.

Järvinen, P., 2007. Action research is similar to design science. *Quality & Quantity, 41*(1), 37-54.

Katuu, S., 2012. Enterprise content management (ECM) implementation in South Africa. *Records Management Journal, 22*(1), pp.37-56

Kawtrakul, A., Mulasastra, I., Khampachua, T. & Ruengittinun, S. 2011. The Challenges of Accelerating Connected Government and Beyond: Thailand Perspectives. *Electronic Journal of e-Government 9*(2), pp183 – 202.

Kay, M. 2007. XSL Transformations (XSLT) Version 2.0. W3C Recommendation (23 January 2007) https://www.w3.org/TR/xslt20/

Kettunen, K. & Henttonen, P. 2010. Missing in action? Content of records management metadata in real life. *Library & Information Science Research, 32*(1), 43-52

Kerer, C., Kirda, E., Jazayeri, M. & Kurmanowytsch, R. 2001. Building and managing XML/XSL-powered Web sites: an experience report. In Computer Software and Applications Conference, 2001. COMPSAC 2001. 25th Annual International (pp. 547-554). IEEE.

Khan, A. & Sum, M. 2006. Introducing Design Patterns in XML Schemas. Oracle technology Network, Java. http://www.oracle.com/technetwork/java/design-patterns-142138.html

Klarlund, N., Moller, A., & Schwatzbach, M. I. 2000. DSD: A schema language for XML. In ACM SIGSOFT Workshop on Formal Methods in Software Practice. Portland, OR.

Kock Jr, N.F., McQueen, R.J. and Scott, J.L. 1997. Can action research be made more rigorous in a positivist sense? The contribution of an iterative approach. *Journal of Systems and Information Technology, 1*(1), 1-23.

Korhonen, R. & Salminen, A. 2003. Visualization of EDI messages: Facing the problems in the use of XML. In N. Sadeh (Ed.), Proceedings of the Fifth

International Conference on Electronic Commerce. New York: ACM Press, 466-473.

Kudo, M. and Hada, S., 2000, November. XML document security based on provisional authorization. In Proceedings of the 7th ACM conference on Computer and communications security (pp. 87-96).

Lee, T., Hon, C.T. & Cheung, D. 2009. XML schema design and management for e-Government data interoperability. *Electronic Journal of E-government 7*(4), 381-390.

Lehtinen, A. & Salminen, A. 2006. Computer aided support for content management development. In Y. Kiyoki, H. Kangassalo, & M. Duži (Eds.), Proceedings of the 16th European-Japanese Conference on Information Modelling and Knowledge Bases, EJC 2006 (pp.275-279). Ostrava: VŠB - Technical University of Ostrava.

Luo R., Wang M., & Zhou S. 2017. Generating Customized PDF Document Based XML Source Data. In: Zhao P., Ouyang Y., Xu M., Yang L., Ouyang Y. (eds) Advanced Graphic Communications and Media Technologies. PPMT 2016. Lecture Notes in Electrical Engineering, vol 417. Springer: Singapore.

Lyytikäinen, V., Tiitinen, P., & Salminen, A. (2001). Supporting access to information created in inter-organizational processes. In A.G. Chin (Ed.), Text Databases and Document Management: Theory and Practice (pp. 223-241). Hersley, PA: Idea Group Publishing.

Maler, E. & El Andaloussi, J. 1996. Developing Sgml Dtds. From Text to Model to Markup. Upper Saddle River (NJ): Prentice Hall.

Manola, F. & Miller E. (Eds.) 2004. RDF Primer. W3C Recommendation. Available at http://www.w3.org/TR/2004/REC-rdf-primer-20040210/

March, S.T., & Smith, G.F. 1995. Design and natural science research on information technology. *Decision Support Systems, 15*(4), 251–266.

Marchetti, A. Megale, F., Seta, E. & Vitali, F. 2002. Using XML as a means to access legislative documents: Italian and foreign experiences. In ACM SIGAPP Applied Computing Review archive (Vol 10 , Issue 1) SPECIAL ISSUE: First European workshop on XML and knowledge management best papers. New York: ACM Press, 54-62.

Meier, W., 2002, October. eXist: An open source native XML database. In Net. ObjectDays: International Conference on Object-Oriented and Internet-Based Technologies, Concepts, and Applications for a Networked World (pp. 169-183). Springer Berlin Heidelberg

Meyer, P. 2005. Case study of Office of Parliamentary Counsel, South Australia. Available at http://www.elkera.com/cms/articles/case_studies/xml_for_legislative_drafting_and_publishing/ [May 18th, 2007]

Molnár, B., & Benczúr, A. 2013. Facet of Modeling Web Information Systems from a Document-Centric View. *International Journal of Web Portals (IJWP), 5*(4), 57-70.

Molnár, B. & Benczúr, A. 2015. Document Centric Modeling of Information Systems. Procedia Computer Science, 64, pp.369-378.

Molnar, R., Gostojić, S., Sladić, G., Savić, G. and Konjović, Z. 2015. Enabling customization of document-centric systems using document management ontology. In International conference on information society and technology (ICIST). Kopaonik, Serbia (pp. 239-243).

Moskal, J., Kokar, M. and Morgan, J. 2015. Semantic Validation of T&E XML Data. In International Telemetering Conference Proceedings. International Foundation for Telemetering.

Myers, M. D. (Ed.) 2007. Qualitative Research in Information Systems. ISWorld Section on Qualitative Research in Information Systems (IS). http://www.qual.auckland.ac.nz/ [March 21, 2007]

Novakovic, D. & Huemer, C. 2013. Business context sensitive business documents: An ontology based business context model for core components. In Tenth Conference for Informatics and Information Technology (CIIT2013), Bitola Macedonia.

Novakovic, D. and Huemer, C. 2014. A survey on business context. In Intelligent Computing, Networking, and Informatics (pp. 199-211). Springer India.

Nunamaker, J.F., Chen, M., & Purdin, T.D.M. 1990. Systems development in information systems. *Journal of Management Information Systems, 7*(3), 89–106.

Nurmeksela, R. 2007. Facing the challenges in implementing XML: The case of the Finnish Parliamentary documents. In M. Muñoz, A. Freitas, & P. Cravo (Eds.), Proceedings of the IASK International Conference E-Activity and Leading Technologies 2007, Porto 3-5 December, 247-255.

Nurmeksela, R., Virtanen, M., Lehtinen, A., Järvenpää, M., & Salminen, A. 2006. Suomalaisen lainsäädäntötyön tiedonhallinta. Suuntana semanttinen web. Eduskunnan kanslian julkaisu 2/2006. Helsinki: Eduskunnan kanslia.

Nurmilaakso, J.-M., Kettunen, J. and Seilonen, I. 2002. XML-based supply chain integration: a case study. *Integrated Manufacturing Systems 13*(8), 586-595.

Obasanjo. 2003. W3C XML Schema Design Patterns: Avoiding Complexity. Microsoft Corporation. https://msdn.microsoft.com/en-us/library/aa468564.aspx

Ogbuji, P. 2004. Principles of XML design: When to use elements versus attributes. IBM developerWorks. http://www.ibm.com/developerworks/library/x-eleatt/x-eleatt-pdf.pdf

OPPD. 2010. Information and Communication Technologies in Parliament Tools for democracy. Bryssel: European Parliament.

Palmirani, M. and Vitali, F., 2011. Akoma-Ntoso for legal documents. In Legislative XML for the Semantic Web (pp. 75-100). Springer Netherlands.

Palmirani, M. & Vitali, F. 2012. Legislative XML: Principles and Technical Tools. Inter-American Development Bank.

Papas, N., O'Keefe, R. M. & Seltsikas, P. 2012. The action research vs design science debate: reflections from an intervention in eGovernment. *European Journal of Information Systems 21*, 147-159.

Paré, G., 2004. Investigating information systems with positivist case research. *The Communications of the Association for Information Systems, 13*(1), p.57.

Peffers, K., Tuunanen, T., Rothenberger, M.A. and Chatterjee, S., 2008. A design science research methodology for information systems research. *Journal of management information systems, 24*(3), pp.45-77.

Poulin, D., Huard, G. & Lavoie, A. 1997a. The other formalization of law: SGML modelling and tagging. In Proceedings of the 6th international conference on Artificial intelligence and law (pp. 82-88). ACM.

Poulin, D., Lavoie, A. & Huard, G., 1997b. Supreme Court of Canada's cases on the Internet via SGML. JL & Inf. Sci., 8, p.177.

Poulymenopoulou, M., Malamateniou, F. & Vassilacopoulos, G. 2014. Document Management Mechanism for Holistic Emergency Healthcare. *International Journal of Healthcare Information Systems and Informatics (IJHISI), 9*(2), pp.1-15.

Päivärinta, T. & Munkvold, B. E. 2005. Enterprise Content Management: An Integrated Perspective on Information Management. In Proceedings of the 38th Hawaii International Conference on System Sciences.

Päivärinta, T., & Salminen, A. 2001. Deliberate and emergent changes on a way toward electronic document management. Annals of Cases on Information Technology Applications and Management in Organizations 3, 320-333.

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. Boston, M. A.: Idea Group Publishing, Hershey, PA, U.S.A., 163-173.

Raggett, D., Le Hors, A., & Jacobs, I.. 1999. HTML 4.01 Specification W3C Recommendation, W3C Consortium. http://www.w3.org/TR/html401/

Rakenne 2008 –työryhmä. 2008. Ehdotus yhteistyöksi valtioneuvoston ja eduskunnan asiakirjojen tuottamisen ja julkaisemisen uudistamiseksi. Valtioneuvoston kanslian raporttisarja 6/2008.

Rezk, N. G., Sarhan, A. & Algergawy, A. 2016. Clustering of XML documents based on structure and aggregated content. In 2016 11th International Conference on Computer Engineering & Systems (ICCES).

Rickenberg, T. A., Neumann, M., Hohler, B., & Breitner, M. H. 2012a. Enterprise content management: A literature review. In Proceedings of the 18th Americas Conference on Information Systems, Seattle, Washington.

Rickenberg, T.A., Neumann, M., Hohler, B. & Breitner, M.H. 2012b. Towards a Process-Oriented Approach to Assessing, Classifying and Visualizing Enterprise Content with Document Maps. In ECIS (p. 118).

Rockley, A., Kostur, P., & Manning, S. 2003. Managing enterprise content: A unified content strategy. Indianapolis, IN: New Riders.

Routledge, N., Bird, L. and Goodchild, A., 2002, January. UML and XML schema. *Australian Computer Science Communications (24)*2. Australian Computer Society, Inc, 157-166.

S1000D. Issue 4.1. http://public.s1000d.org/

Salminen, A. 2000. Methodology for document analysis. In A. Kent (Ed.), *Encyclopedia of Library and Information Science*, Vol. 67 (Supplement 30). New York: Marcel Dekker, 299-320.

Salminen, A. 2003. Towards digital government by XML standardization: Methods and experiences. In Proceedings of the XML Finland 2003 (pp. 5-15).

Salminen, A. 2005. Building digital government by XML. In R. H. Spraue, Jr. (Ed.) *Proceedings of the Thirty-Eight Hawaii International Conference on System Sciences.* Los Alamitos CA: IEEE Computer Society Press.

Salminen, A., Kauppinen, K. & Lehtovaara, M. 1997. Towards a methodology for document analysis. *Journal of the American Society for Information Science 48*(7), Special Issue on Structured Information/Standards for Document Architectures, 644-655.

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

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

Salminen, A., Lyytikäinen, V., Tiitinen, P., & Mustajärvi, O. 2004. Implementing digital government in the Finnish Parliament. In W. Huang, K. Siau, & K.K. Wei (Eds.), Digital Government: Strategies and Implementation (pp. 242-259). Hersley, PA: IDEA Group Publishing.

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

Salminen, A. & Tompa, F.W. 2001. Requirements for XML document database systems. In Proceedings of the 2001 ACM Symposium on Document engineering (pp. 85-94). ACM.

Salminen, A. & Tompa, F.W. 2011. Communicating with XML. New York: Springer.

Salminen, A. & Virtanen, M. 2005. Semantic web support for business processes. In C.-S. Chen, J. Flippe, I. Secura & J. Cordeiro (Eds.), Proceedings of the 7th International Conference on Enterprise Information Systems, ICEIS 2005 (pp. 468-473). Miami, FL: INSTICC Press.

Sandahl, T., and Jenssen, A. 1997. The First Steps in Designing an SGML-Based Infrastructure for Document Handling". *Scandinavian Journal of Information Systems, 1997, 9*(2), 25–44.

Sapienza, F. 2004. Usability, structured content, and single sourcing with XML. *Technical communication 51*(3), 399-408.

Sedlar, E. 2005. Managing structure in bits & pieces: the killer use case for XML. In International Conference on Management of Data: Proceedings of the 2005 ACM SIGMOD international conference on Management of data (Vol. 14, No. 16, pp. 818-821).

Smith, H.A., and McKeen, J.D. 2003. Developments in practice VIII: Enterprise content management. *The Communications of the Association for Information Systems, 11*(1), 647-659.

Sprague, R.H. Jr. 1995. Electronic document management: Challenges and opportunities for information systems managers. *MIS Quarterly, 19*(1), 29–49.

Su, M., Li, F., Tang, Z., Yu, Y. and Zhou, B. 2014. An action-based fine-grained access control mechanism for structured documents and its application. *The Scientific World Journal, 2014*, .

Sundholm, E. 1998. ODIN: the central web-server for official documentation and information from Norway. INSPEL, 32, pp.120-125. http://www.ifla.org/IV/ifla63/63hole.htm

Susman, G.I. and Evered, R.D., 1978. An assessment of the scientific merits of action research. *Administrative science quarterly,* 582-603.

Svoboda, W. R. 2005. Current state of publication of legislation in the EU Member States. Available at: http://forum.europa.eu.int/irc/opoce/ojf/info/data/prod/html/index.htm [July 18, 2006]

Svärd, P. 2014. Exploring two approaches to information management: two Swedish minicipalities as examples. In vom Brocke and Simons (Eds.) *Enterprise content management in information systems research*, 217-235.

Thiam, M. 2016. A deep and uniform model for semantic annotation of semi structured documents based on SHIRI. In 2016 4th International Conference on Control Engineering & Information Technology (CEIT).

Thompson, H. S., Beech, D., Maloney, M., & Mendelsohn, N. 2004. XML Schema Part 1: Structures Second edition. W3C Recommendation. Available at http://www.w3.org/TR/xmlschema-1/.

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 (pp. 1144-1151). Wien: Wirtschaftsuniversität Wien.

Tough, A & Moss, M. 2003. Metadata, controlled vocabulary and directories: electronic document management and standards for records management. *Records Management Journal 13*(1), 24–31.

Tyrväinen, P., Päivärinta, T., Salminen, A. & Iivari, J. 2006. Characterizing the evolving research on enterprise content management. *European Journal of Information Systems 15*(6), 627-634.

the United Nations & the Inter-Parliamentary Union. 2014. Technological Options for Capturing and Reporting Parliamentary Proceedings. Rome: Global Centre for ICT in Parliament

Usman, M., Muzaffar, A. & Rauf, A. 2009. Enterprise content management (ECM): needs, challenges and recommendations". In *Proceedings of the 2nd IEEE International Conference on Computer Science and Information Technology, Beijing, August 8-11.* doi: 10.1109/ICCSIT.2009.5234541

Valtiovarainministeriö, 2008. Valtioneuvoston ja eduskunnan yhtenäiset metatiedot ja tietorakenteet (VM 24/2008)

vom Brocke, J. & Simons, A. 2014. Enterprise Content Management in Information Systems Research. Berlin, Heidelberg: Springer Berlin Heidelberg.

vom Brocke, J., Simons, A. & Cleven, A. 2008. A Business Process Perspective on Enterprise Content Management: Towards a Framework for Organisational Change. In Proceedings of the 16th European Conference on Information Systems, Galway.

vom Brocke, J., Simons, A., & Cleven, A. 2011. Towards a business process-oriented approach to enterprise content management: The ECM-blueprinting framework. Information Systems and eBusiness Management, 9(4), 475-496.Weitzman, L., Dean, S., Meliksetian, D., Gupta, K., Zhou, N., and Wu, J. 2002. "Transforming the content management process at IBM.com", Case studies of the CHI2002/AIGA Experience Design Forum, ACM Press, New York, 1-15.

Yin, 1994. Case study research: Design and methods. Thousand Oaks: Sage publications.

Zachman, J.A. 1987. A Framework for Information Systems Architecture, *IBM Systems Journal, 26*(3), pp. 276--292.

APPENDIX 1

HAASTATTELULOMAKE: **Metatietojen tuottaminen ja käyttö suomalaisessa lainsäädännön laadintaprosessissa**

Haastateltavan taustatietojen tarkistamisen jälkeen haastattelijat esittävät lyhyesti haastattelun tavoitteet ja kohdealueen asiankäsittelyprosessin yleiskuvauksen, joka on tehty RASKE2-projektin perusselvitysvaiheessa. Tämän jälkeen täytetään haastattelulomakkeen kohdat sekä liitetaulukko yhteistyössä haastateltavan kanssa. Haastateltava tarkistaa lomakkeelle ja liitteeseen tulevan tiedon oikeellisuuden.

Rooli: **Haastattelupvm ja paikka:**

1. Haastateltavan taustatietoja

1.1 Nimi:

1.2 Organisaatio / organisaatioyksikkö:

1.3 Asema / työtehtävä:

1.4 Kuinka kauan olette toiminut kyseisessä organisaatiossa / tehtävässä:

1.5 Tietojärjestelmät, joita käytätte lainsäädännön laatimisprosessissa säännöllisesti:

2. Haastateltavan suhde kohdealueeseen

2.1. Organisaationne pääasiallinen tehtävä lainsäädännön laatimisprosessissa on:

2.2. Miten oma työnne liittyy lainsäädännön laatimisprosessiin (lyhyesti ilmaistuna)?

2.3. Mikä on oma paikkanne lainsäädännön laatimisprosessissa (kohdealueen kuvauksesta)?

2.4. Onko teillä huomautettavaa lainsäädännön laatimisprosessin kuvaukseen?

3. Asiakirjojen ja metatietojen tuottamiseen liittyvät työtilanteet lainsäädännön laatimisprosessissa

Kuvailkaa niitä työtilanteita lainsäädännön laatimisprosessissa, joissa tuotatte yleiskuvauksessa esitettyihin asiakirjoihin sisältyviä tai niihin liitettäviä metatietoja.

3.1 Mihin asiakirjoihin tuotatte joko niihin sisältyviä tai niitä koskevia metatietoja? Mikä on kyseisen työtilanteen tavoite tai tulos?

3.2 Minkä tiedon tai tietojen perusteella laatimanne asiakirja liitetään kuuluvaksi tiettyyn lainsäädäntöasiaan?

3.3 Tuotatteko muita dokumentteja, joita pidätte tärkeinä lainsäädännön laatimisprosessissa?

3.4 Mitä liitteessä 1 kuvatuista metatiedoista sisällytätte asiakirjoihin?

3.5 Mitä liitteessä 1 kuvatuista metatiedoista tuotatte johonkin muualle kuin asiakirjojen sisältöihin?

3.6 Tiedättekö muita liitteen 1 luonteisia tietoja, joita tuotatte tai liitätte asiakirjoihin?

3.7 Tiedättekö muita liitteen 1 luonteisia tietoja, joita muut tuottavat tai liittävät asiakirjoihin?

3.8 Mitä tuottamistanne metatiedoista hyödynnetään säädettävänä olevan lain laatimisprosessin lisäksi myös jossain toisessa lain laatimisprosessissa?

Tilanne ja sen tavoite	Tuotettavan asiakirjan nimike tai asiakirjan osat, jota koskevaa metatietoa tuotetaan / metatiedon sisältyminen asiakirjaan	Liitos lainsäädäntöasiaan

4. Metatietojen hyödyntämiseen liittyvät työtilanteet lainsäädännön laadintaprosessissa

Kuvailkaa niitä työtilanteita lainsäädännön laatimisprosessissa, joissa hyödynnätte yleiskuvauksessa esitetyn prosessin asiankäsittelytietoja.

4.1 Mikä on kyseisen työtilanteen tavoite tai tulos?

4.2 Mitä tietoja tarvitsette?
4.3 Koskevatko tiedot säädettävänä olevaa vai jotain muuta lainsäädäntöasiaa?
4.4 Miten saatte käyttöönne tarvitsemanne tiedon?
4.5 Minkä tietojen perusteella tunnistatte saman lainsäädäntöasian eri järjestelmissä?

Tilanne ja sen tavoite	Tarvittavan tiedon nimike / tiedon suhde säädettävänä olevaan asiaan	Tiedon saantitapa / asian tunnistamistieto eri järjestelmissä

5. Metatietojen yhtenäistämiseen liittyvät tarpeet, ongelmat ja kehittämisehdotukset

5.1 Minkälaiset ja minkä tyyppiset tiedot kaikista lainsäädäntöprosessissa tuotettavista asiakirjoista tai asiankäsittelystä tukisivat mielestänne parhaiten lainsäädännön asiankäsittelyprosessia?
5.2 Näettekö esteitä näiden tietojen yhtenäiselle esittämistavalle koko lainsäädäntöprosessissa?
5.3 Mitä kehittämisehdotuksia teillä on lainsäädännössä tuotettavien asiakirjojen tai asiankäsittelytietojen yhtenäistämiseksi?

Lainsäädännön asiankäsittelyprosessia tukevat metatiedot	Metatietojen yhtenäistämismahdollisuus

Tarve asiakirjojen tai asiankäsittelytietojen yhtenäistämiseksi	Kehittämisehdotus asiakirjojen tai asiankäsittelytietojen yhtenäistämiseksi

6. Metatietojen hyödyntäminen tulevaisuudessa

Seuraavassa on kuvattu muutamia esimerkkikyselyjä, joihin lainsäädäntöasian käsittelyn metatietoihin kohdistuvalla tiedonhaulla voitaisiin tulevaisuudessa mahdollisesti vastata. Miten arvioisitte niitä omalta kannaltanne?
6.1 Mitä tiettyä säädöstä koskevia muutoksia on valmisteilla?
6.2 Missä käsittelyvaiheessa tietty lainsäädäntöhanke on? Mitä käsittelyvaiheita tiettyyn lainsäädäntöhankkeeseen on liittynyt?
6.3 Mihin lainsäädäntöasioiden käsittelyvaiheisiin tietty henkilö on osallistunut? Mitä lainsäädäntöprosessissa syntyviä asiakirjoja tietty henkilö on laatinut?
6.4 Mitä asiakirjoja tietyn lain valmistelu- ja käsittelyvaiheissa on syntynyt? Mitkä niiden kohdista dokumentoivat kyseistä lainsäädäntöasiaa? Mikä ovat näiden asiakirjakohtien tietosisällöt?
6.5 Lakitekstiä on muutettu eduskuntakäsittelyn aikana. Mitä pykäliä on muutettu ja miten muuttunut lakiteksti eroaa hallituksen esityksen lakitekstistä?
6.6 Mitkä tietyn lain kohdista liittyvät johonkin muuhun lakiin ja millainen on näiden muiden lainkohtien lakiteksti? Mitkä muut lait rajoittavat tietyn lain kohtia ja millainen on näiden rajoittavien lainkohtien lakiteksti?
6.7 Liittyykö tietyn lain käsittelyyn äänestyksiä? Mikä on äänestyksen tulos?
6.8 Liittyykö lakialoitteeseen rinnakkaislakialoitteita?
6.9 Liittyykö lakiehdotukseen vastalauseita?

APPENDIX 2

KYSELYLOMAKE: **Rakenteisten asiakirjojen käyttöönotto ja laatiminen**
16.5.2006 RASKE2-seminaari

Kyselyssä kartoitetaan rakenteisten SGML/XML-asiakirjojen käyttöönottoon ja laatimiseen liittyviä asioita. Kyselyllä kerätään aineistoa Jyväskylän yliopistossa tehtävään tutkimustyöhön. Toivomme, että mahdollisimman moni RASKE2-seminaariin osallistuja vastasi kyselyyn seminaarin aikana. Kyselyyn vastaaminen kestää muutamia minuutteja. Mikäli haluatte enemmän vastausaikaa, voitte toimittaa täytetyn lomakkeen 31.5.2006 mennessä osoitteella:

Reija Nurmeksela
Tietojenkäsittelytieteiden laitos
PL 35 (Agora)
40014 Jyväskylän yliopisto

1. Taustatietoja

Voitte halutessanne vastata kyselyyn myös nimettömänä

1.1 Nimi: ____
Sähköpostiosoite: ______________________________

1.2 Organisaatio, jossa työskentelette:
Vastausohje: Ympäröi yksi vaihtoehto

1. Eduskunta
2. Kauppa- ja teollisuusministeriö
3. Liikenne- ja viestintäministeriö
4. Maa- ja metsätalousministeriö
5. Oikeusministeriö
6. Opetusministeriö
7. Puolustusministeriö
8. Sisäasiainministeriö
9. Sosiaali- ja terveysministeriö
10. Tasavallan presidentin kanslia
11. Työministeriö
12. Ulkoasiainministeriö
13. Valtioneuvoston kanslia
14. Valtiovarainministeriö
15. Ympäristöministeriö
16. Muu, mikä?
____

1.3 Työtehtävä:

____

____

____

____

1.4 Kuinka kauan olette toiminut kyseisessä organisaatiossa/tehtävässä:
Vastausohje: Ympäröi yksi vaihtoehto

1. Alle 1 vuotta
2. 1 -5 vuotta
3. 6-10 vuotta
4. Yli 10 vuotta

2. Asiakirjojen rakenteistaminen

Tietokoneeseen tallennettujen dokumenttien tehokas hyväksikäyttö edellyttää formaalien sääntöjen sopimista. Dokumentteja, joihin on liitetty tietokoneen tulkittavissa oleva, standardoitu rakennemäärittely, kutsutaan *rakenteisiksi dokumenteiksi*. Dokumenttien rakenne voidaan ilmaista esimerkiksi SGML- tai XML-kielillä ja standardoitu rakennemäärittely esimerkiksi DTD- tai XML Schema -kielillä. Rakenteisten asiakirjastandardien kehittämistä kutsutaan *rakenteistamiseksi*.

2.1 Oletteko osallistunut työtehtävissänne asiakirjojen rakenteistamiseen?
Vastausohje: Ympäröi yksi vaihtoehto

1. Kyllä
2. En

Jos valitsit vaihtoehdon kaksi, siirry kyselylomakkeen kohtaan 3.

2.2 Minkä asiakirjatyyppien rakenteistamiseen olette osallistunut?
Vastausohje: Ympäröi yksi tai useampi vaihtoehto

1. asetus A
2. eduskunnan kirjelmä EK
3. eduskunnan vastaus EV
4. hallituksen esitys HE
5. hallituksen kirjelmä
6. kertomus K
7. keskustelualoite KA
8. kirjallinen kysymys KK
9. lakialoite LA
10. lepäämään jätetty lakiehdotus LJL
11. lisätalousarvioaloite LTA
12. ministeriön päätös MP
13. ministeriön selvitys MINS
14. muu asia M
15. puhemiesneuvoston ehdotus PNE
16. puhemiesneuvoston laatima luettelo PNL
17. Päiväjärjestys PJ
18. pääministerin ilmoitus PI
19. säädösteksti
20. talousarvioaloite TAA
21. talousarviomietintö, talousarviokirjelmä
22. toimenpidealoite TPA
23. toivomusaloite TA
24. täysistunnon keskustelupöytäkirja PTK
25. ulko- ja turvallisuuspolitiikan asiakirja UTP
26. vahvistamatta jätetty lakiehdotus VJL
27. valiokunnan lausunto VL
28. valiokunnan mietintö VM
29. valiokunnan pöytäkirja, valiokunnan esityslista
30. valtioneuvoston kirjelmä U
31. valtioneuvoston kirjelmä U-jatkokirjelmä
32. valtioneuvoston kirjelmä VN
33. valtioneuvoston päätös VNP
34. valtioneuvoston selonteko VNS
35. valtioneuvoston selvitys E
36. valtioneuvoston selvitys E-jatkokirjelmä
37. valtioneuvoston tiedonanto VNT
38. Valtion talousarvioesitys
39. välikysymys VK
40. Y-kirjelmä
41. Y-jatkokirjelmä
42. Muu,mikä?_Valtiosopimus, Muistio, TP vahvistus, VN määräys______________________
____
____
____

2.3 Millä tavalla osallistuitte rakenteistamiseen?

____

____

____

____

2.4 Millaisia ongelmia koitte työtehtävissänne asiakirjojen rakenteistamisen aikana? Millaisia ajatuksia Teillä on asiakirjojen rakenteistamiseen liittyen?

____

____

____

____

3. Rakenteisten asiakirjojen laatiminen

3.1 Tuotetaanko organisaatiossanne rakenteisia SGML/XML-asiakirjoja?

Vastausohje: Ympäröi yksi vaihtoehto

1. Kyllä   2. Ei   3. En tiedä

3.2 Laaditteko itse rakenteisia SGML/XML-asiakirjoja tai niiden osia?

Vastausohje: Ympäröi yksi vaihtoehto

1. Kyllä   2. En   3. En tiedä

Jos valitsit vaihtoehdon 2 tai 3, siirry kohtaan 3.5.

3.3 Mitä asiakirjatyyppejä kysymyksessä 3.2 mainitsemanne rakenteinen tuottaminen koskee?

Vastausohje: Ympäröi yksi tai useampi vaihtoehto

1. asetus A
2. eduskunnan kirjelmä EK
3. eduskunnan vastaus EV
4. hallituksen esitys HE
5. hallituksen kirjelmä
6. kertomus K
7. keskustelualoite KA
8. kirjallinen kysymys KK
9. lakialoite LA
10. lepäämään jätetty lakiehdotus LJL
11. lisätalousarvioaloite LTA
12. ministeriön päätös MP
13. ministeriön selvitys MINS
14. muu asia M
15. puhemiesneuvoston ehdotus PNE
16. puhemiesneuvoston laatima luettelo PNL
17. Päiväjärjestys PJ
18. pääministerin ilmoitus PI
19. säädösteksti
20. talousarvioaloite TAA
21. talousarviomietintö, talousarviokirjelmä
22. toimenpidealoite TPA
23. toivomusaloite TA
24. täysistunnon keskustelupöytäkirja PTK
25. ulko- ja turvallisuuspolitiikan asiakirja UTP
26. vahvistamatta jätetty lakiehdotus VJL
27. valiokunnan lausunto VL
28. valiokunnan mietintö VM
29. valiokunnan pöytäkirja, valiokunnan esityslista
30. valtioneuvoston kirjelmä U
31. valtioneuvoston kirjelmä U-jatkokirjelmä
32. valtioneuvoston kirjelmä VN
33. valtioneuvoston päätös VNP
34. valtioneuvoston selonteko VNS
35. valtioneuvoston selvitys E
36. valtioneuvoston selvitys E-jatkokirjelmä
37. valtioneuvoston tiedonanto VNT
38. Valtion talousarvioesitys
39. välikysymys VK
40. Y-kirjelmä
41. Y-jatkokirjelmä
42. Muu,mikä?_Valtiosopimus, Muistio, TP vahvistus, VN määräys______________________

3.4 Millaisia rakenteisten asiakirjojen laatimiseen liittyviä ongelmia koette työtehtävissänne?
3.5 Mitä rakenteisten asiakirjojen laatimiseen liittyviä kehittämisajatuksia ajatuksia Teillä on?

____

____

____

Kiitos vastauksestanne! Voiko Teihin tarvittaessa ottaa yhteyttä kyselyyn liittyen?

1. Kyllä
2. Ei

APPENDIX 3

Reviewers of the section 3.3 Finnish Parliamentary documents.

Jaana Kaakkola, Senior Specialist, Parliament of Finland
Tuula Kulovesi, Director of Legislation, Head of Central Chancellery, Parliament of Finland
Timo Tuovinen, Deputy Secretary General, Parliament of Finland
Sanna Turpeinen, Assistant at the Committee Office, Parliament of Finland

ORIGINAL PAPERS

I

CONTENT PRODUCTION STRATEGIES FOR E-GOVERNMENT

by

Salminen, A., Nurmeksela, R., Lehtinen, A., Lyytikäinen, V., & Mustajärvi, O.
2008

In A.-V. Anttiroiko (Ed.), Electronic Government: Concepts, Methodologies, Tools, and Applications.

II

XML DOCUMENT IMPLEMENTATION: EXPERIENCES FROM THREE CASES

by

Nurmeksela, R., Jauhiainen, E., Salminen, A., & Honkaranta, A. 2007

In Y. Badr, R. Chbeir, & P. Pichappan (Eds.), Proceedings of the Second International Conference on Digial Information Management, 224-229

XML Document Implementation: Experiences from Three Cases

Reija Nurmeksela Eliisa Jauhiainen Airi Salminen* Anne Honkaranta**
University of Jyväskylä, Department of Computer Science & Information Systems, Finland
** University of Toronto, Faculty of Information Studies, Canada*
***SYSOPENDIGIA Plc, Finland*
rekorhon|raelurja|airi.salminen@jyu.fi, anne.honkaranta@sysopendigia.com

Abstract

Implementing production of XML documents is a rarely discussed topic in academic literature even though it is an important issue in many contemporary organizations. This paper describes findings from three case organizations where different kinds of XML documents were implemented. Our findings suggest that the implementation is a domain-specific task related to various kinds of organizational activities from document authoring to business processes. As expected, the amount and complexity of document types as well as the number of people and organizations involved affect the challenges in the implementation process. Hiding the XML format from the software users and training the end users are important means to reduce the user resistance against structured document authoring and novel tools.

1. Introduction

A great deal of the information resources in organizations consist of documents produced in organizational business processes. Documents serve a number of different purposes, for example, as tools for supporting communication and decision making and as recordings of business activities. Some documents are information carriers meant primarily for human readers, while some others are targeted for software systems. Recently *Extensible Markup Language (XML)* [1] has been adopted in many organizations to support more systematic *Enterprise Content Management (ECM)*, i.e., the management of information content in various kinds of assets like documents, Web sites, intra- and extranets across the organization and between parties involved in business processes [2]. XML is a standard *de* facto by W3C consortium. It is a metalanguage that provides a way to exchange information between software applications in standardized formats.

Adopting the standardized format for documents by means of XML is motivated, for example, by the needs for interoperability, data integration, improved information access, and reuse of information content [2, 3, 4, 5]. The XML *standardization* in an organization refers to the adoption of XML standard which includes agreeing upon rules for the ways information is clustered and represented in documents as well as those for content production and management practices. The implementation of standards hence requires both technical and also organizational solutions, possibly including extensive re-engineering of information systems and document production practices (e.g. [3, 5]). ECM environments of organizations are varying and thus the ways XML is used in the environments, too. Therefore also the efforts needed for implementing the XML documents vary.

This paper describes and compares three XML document implementation cases and focuses on the changes in document production practices. The research is conducted by using qualitative case study method [6]. The case analysis is targeted on finding answers to the following questions: What motivates organizations in document standardization? How the implementation process is realized in different kinds of organizations? What kinds of changes the implementation causes in the document production practices?

The paper is organized as follows. Section 2 describes the concepts related to the adoption of XML in document production. Section 3 introduces the three cases, all of which were realized in Finland. The cases include one private sector and two public sector organizations. In the first case the implementation started by using the SGML (Standard Generalized Markup Language [7]), the predecessor to XML language, and the latter two cases were realized using XML. Findings from the cases are discussed in Section 4. Section 5 concludes the paper.

2. Production of XML documents

This section introduces the core concepts related XML documents, the various ways XML documents are produced in organizations, and a model for XML standardization process. The model will be used as an analysis tool to describe the three cases in Section 3.

2.1. Characteristics of XML documents and their use

SGML and XML documents are *structured documents* where the structure definitions, document instances, and layout specifications can be managed as separate content items. The logical structure of an XML document is hierarchical, the structure and other constraints for document content are described by an *XML schema*. In the document instance the logical structure is indicated by markup which follows the schema rules. The layout specification is typically defined with a *stylesheet*.

XML documents are often divided into data-centric and document-centric (see e.g. [8]) ones, based on their purpose and type of use. *Document-centric* XML documents are designed primarily for human understanding, and the content of these documents usually consists of natural language. *Data-centric* XML documents are primarily designed for software processing and data exchange. They are typically shorter than document-centric ones and without layout specifications. XML documents in both categories may serve as recordings of business activities. An alternative categorization for XML use is proposed by [2], dividing XML into two categories: use as information assets and use as data interchange format. In the first case the XML use results a persistent XML repository. The information assets are further divided into documents and metadata.

2.2. XML document production

XML documents may be produced in a number of ways [9]; by human actors or by software systems. Human authoring may be supported by multiple ways, such as described in the following.

Using document templates with styles that may be mapped to XML document schema by a software application which post-processes the document after the human authoring. For example, a Word .dot document template may have styles “title” and “bodytext”. The author marks the text up by using these styles in Word. A markup application takes the document as input and maps the style definitions into schema elements to produce an output XML document.

Using a generic, schema or syntax directed XML editor such as Altova XMLSpy. This editor allows the document to be produced as valid or well-formed XML. The author types the text into table view in which each cell is a placeholder for an element content, or author types the text in between the element start and end tags directly. The editor may show the schema structure and provide hints of the elements that may be added into the document with regard to the current position in XML document type schema structure.

Using a generic word processor with XML support. For example Microsoft Word 2007 and OpenOffice 2.0 support XML. Word 2007 allows a form of schema-directed editing by using content controls and element markers into which element content may be typed into. Open Office 2.0 Writer uses Open Document Format, which allows content to be typed in a WYSIWYG-interface with styles to produce an XML document which conforms to Open Document Schema, and may be transformed to another XML document markup language such as Docbook or XHTML.

Using a custom-designed interface developed for a certain document type separately. One may for example define one .edd-template for each document type schema for FrameMaker+XML, develope a tailored form according to one’s schema in Microsoft InfoPath (2003 or 2007), utilize InfoPath forms and Microsoft Forms Server and transform the InfoPath forms into Web forms, or utilise Altova’s StyleVision to define a Web form for each document type schema separately for document authoring. Some of the aforementioned solutions may be mapped with specific content controls and sources to facilitate content retrieval and reuse from external sources, such as databases or Web Service-interfaced content sources on the Web. Such functionality is provided by Altova XML Spy, Microsoft Word 2007, Microsoft Infopath and Web Forms transformed from them, for example.

In all of the cases above the document schema sets constraints to the authoring, and the author must be more or less familiar with the schema for the document type. In traditional document authoring the document content is strictly tied to the external presentation visible to the author. The authors may find the schema-guided authoring too restrictive and feel uneasy about the separation between the logical structure and layout [10]. Therefore new solutions for XML editing have been developed (e.g. [11, 12]).

2.3. Standardization process

XML document implementation in an organization results from an XML standardization process. For our case analyses we use the standardization model depicted in Figure 1. The model has been adapted from [13]. The circles represent process phases and the arrows show the order for starting the activities. The small black circle indicates that all of the following three activities may be started either in parallel or in any order.

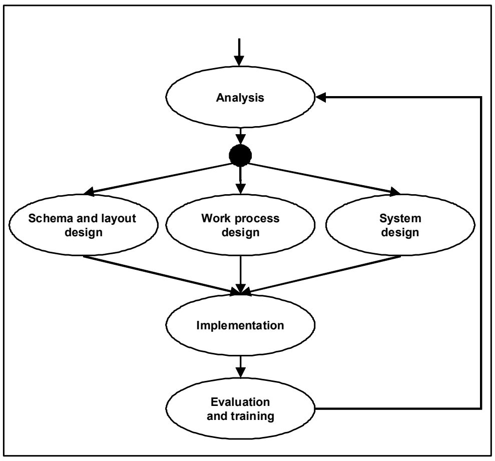

Figure 1. The XML standardization process.

The process starts with an *analysis* phase producing descriptions of the actors, business processes, systems, and documents of the domain as well as a requirements analysis report. The users of the future solutions may be involved in the process from the analysis phase.

The design of the new solutions usually requires a *schema and layout design*, carried out in parallel with *systems design* and possibly with *work process design*. The schema design is accompanied by layout design for facilitating the representation of documents to human authors and readers (e.g. [3, 5]). The system design may include, for example, selection of XML software products, their customization, and designing transformations between different data formats. Collaboration with the future document authors during the design facilitates reactivity to the anticipated problems in the schema design and systems customization [3, 5, 10].

The *implementation* of the new XML-based solution requires both technical and organizational implementation of the new solutions, possibly including major changes in document processing, as pointed out in previously reported cases (e.g. [3, 5, 9, 10]).

Evaluation and training is an important phase for successful adoption of new solutions. Evaluation may lead to further redesign. For example, XML schema design is not completed until schemas have been used by end users [5, 10]. After some time of operational use, standardization may continue by iterating the process.

The type of standardization domain most obviously affects the extent and challenges of the standardization process and the implementation of document production. Following section describes the three cases and compares them.

3. XML document implementation in the cases

In each of the cases the activities in the organizations have been supported by one or two of the authors of the paper. Most of the data were originally collected by participating in the standardization activities, interviewing people involved, and analyzing documents and schemas.

Case 1 concerned standardization of 35 **parliamentary document types** in the Finnish Parliament and 13 ministries. One of the authors was involved in the analysis and evaluation phases of the case. The description of the case is based on data analyzed and reported earlier in [3], which was updated by interviews and schema analysis for the paper. *Case 2* concerned **agendas and memorandums** of the Faculty of Information Technology in the University of Jyväskylä, in which two of the authors were involved throughout the standardization process. *Case 3* concerned standardization of **invoice documents** in an international ICT service provider and its customers. One of the authors participated in the design and implementation phases of this case. Table 1 summarizes the cases and their characteristics.

The Finnish Parliament and Government (Case 1) decree governmental and administrative matters with the President of the Republic. Standardization activities took place during 1994-2007. **The standardization was motivated by** incompatibilities of systems, inconsistencies in representations, heterogeneity in retrieval techniques, and uncertainty of the future usability of archived digital documents. In 1994 XML was not yet published. At the end of the analysis phase SGML was chosen as the basis for standardization and preliminary schemas were designed. Redesign and implementation projects took place 1998-2001. Iteration of the standardization for replacing SGML and style-based authoring with XML was started in 2004 and is still going on. **Major challenges on the implementation have been** in co-operation between different organizations, large amount of document types and instances, strict usability requirements, and user resistance. **As a result**, quality of documents has been improved. Standardization had affected both in and out of the organizations involved.

Table 1. XML standardization in the three cases.

	Case 1: Finnish Parliament and Government, 1994-2007	Case 2: Faculty of Information Technology in the University of Jyväskylä, 2004-2006	Case 3: An international ICT provider company and its customers, 2000-2007
Analysis	The initial analysis phase during 1994-1998, including extensive data gathering. The analysis concerned the Parliamentary documents, people involved and the tools used in the document production. New analysis in the Government during 2004-2006 for adopting XML.	Fall 2004, including interviews and studying existing documents. The analysis was carried out by a student group. The analysis concerned the people involved in document production, the two document types and the tools used in document production. The analysis was iterated in department of the faculty 2006.	2000-2002 by the ICT provider. The core group consisted of project managers with support of technical consultants. The analysis concerned different invoice types as well as people, organizations and systems involved in invoicing activities. Interchange message standard for invoice was modeled. Iteration of the analysis four times during 2003-2007.
Design Schema and layout	20 preliminary SGML schemas were designed by researchers and selected companies designed the final SGML/XML schemas and layouts.	Schemas for agendas and memorandums were designed. Strict requirements for layout.	The selected subcontractor and later on the system analysts of the provider developed schema and layout for the invoice. Strict requirements for layout.
Work process	Work processes were redesigned.	Existing work processes were supported by the adaptation of the system.	Work processes were not redesigned.
Systems	Adobe Framemaker +SGML was selected as the authoring tool at the Parliament, Microsoft Word in the Government.	Microsoft InfoPath replaced Word as document authoring software.	Invoicing system produced XML documents. Significant changes in invoicing, purchase ledger and workflow systems in the first standardization process.
Implemen-tation	During 1998-2000 in the Parliament and 2000-2001 in the Government. Transfer to SGML production in the Parliament, use of a word processor with style editor in the Government.	XML-based document production was implemented in 2005. The department version of the system was implemented in 2006.	Provider implemented incrementally exchange service for invoices during 2000-2007. Parallel related systems were implemented. Implementations in the customer organizations 2001-2007.
Evaluation and Training	Gradual improvements on the systems. Training offered to people whose work changed.	Office personnel were trained briefly before they started using the novel XML application.	Training offered to users and developers of the invoicing, purchase ledger and workflow systems.

The initial project for the XML implementation in the University of Jyväskylä was carried out during 2004-2005. **The standardization efforts were motivated by** the need to improve content reuse and enhance the laborious document preparation and publishing process. **Major challenges on the implementation were** caused by the limited timetable and the lack of XML competence on the project group. **As a result**, quality of documents i.e. the consistency on the content and layout was improved and document publishing as well as content reuse were enhanced. The project was followed by another standardization project at a department of the faculty in 2006. This time the implementation was smooth and easy, since the solution developed in the faculty was tested and successfully implemented before the decision for its adaptation in the department was done.

Case 3 considered XML implementation for invoice documents in an international ICT service provider company and its customers. **The standardization efforts were motivated by** the need to improve data integrity between invoicing and purchase ledger systems in small and medium enterprises (SME), speed up the handling of invoices, and to reduce the costs. The case was started in 2000 and a new XML-based solution was implemented for the first customers in 2001. The implementation of a new version of invoicing system for the first customer organization took a few days. During 2003-2007 the standardization process was iterated four times, because of the need for new schema versions. **Major challenges on the implementation** have been caused by the large amount of document instances, disagreement of identification standards with different business partners, and importance of layout. **As a result**, quality of invoice data has been improved and the invoicing process streamlined.

4. Evaluation

The motivation for SGML/XML implementation and standardization varied from case to case. In Case 1, standardization was activated by inconsistencies in content management, incompatibilities of tools, and uncertainty of the future usability of archived digital documents. In Case 2, requirements for content reuse and difficulties in document publishing were the main motivations. In Case 3, automation of invoice processing in SMEs was the key motivator.

The SGML/XML implementation in the three cases has major differences. Case 1 was clearly the most challenging. It started at the time before XML, the experiences about the use of SGML in public domain were limited, the SGML tools were expensive, and there were not many choices for the tools. Compared to the other cases, the number of document types and document instances was much bigger, as well as the number of people affected by the standardization. The documents in question were nationally very important and the standardization was expected to have many impacts both in the work environment and in the society as a whole. These expected impacts most probably gave extra motivation to the persons involved. Wide impacts have also been realized as reported in [3].

In Cases 1 and 2 the XML documents were targeted for human consumption due to which the document layout design was an essential part of the standardization process. The analyses conducted in the cases focused on similar aspects; documents and tools used as well as people involved in document production were analyzed. In Case 3, the document type to be standardized (the invoice) had both document- and data-centric characteristics. Thus the analysis was focused on the data stored in the accounting systems and the organizations involved in the invoicing process, instead of people as end users of the system. In all cases the layout requirements had a significant impact on the schema design.

In Case 1 the document schemas were designed incrementally within years. In Cases 2 and 3 the document schemas were designed within a few months. In each of the cases the schema design was an iterative process. The tools utilized for content reuse in Case 2, and the data integration requirements between systems in Cases 1 and 3 had effects on the schema design. In all cases usability requirements had an impact to schemas [e.g. 10]. For example, in Case 2 significant requirements and limitations came from the authoring tool, which provided the visual layout for authors. The layout had to support the functions and easy to use. Therefore schemas were modified several times before the implementation. In Cases 1 and 2 the changes in the schemas reflected further in the authoring tools.

In each of the cases new formats were implemented to support multi-channel publishing. Changes in work practices vary between the cases. In Case 1, the standardization changed significantly work tasks of different groups of people including new publishing practices. In Case 2, work practices remained in essence the same, only the authoring tool changed. Furthermore, because the process of preparing agendas and memorandums was similar in the Faculty of Information Technology and in another faculty, the same XML-based system was soon implemented in the other faculty as well. Only minor refinements in the XML schemas were needed and the change of the document authoring software was not resisted due to the awareness of the user satisfaction gained by the novel system and due to the fact that the benefits were already manifested by the neighboring faculty. Easy and simple launching of new systems based on the same content is also reported in [5].

In Case 3, the redesign of work was also avoided. Only the new functionalities of the new versions of the invoicing, purchase ledger and workflow systems were introduced, together with the new invoice exchange service. As in Case 2, the same versions of the systems were quickly implemented in many other organizations, because the invoicing process is similar and also the same systems are used in those organizations.

The case comparison reveals that if the business process and document types involved in it are long and complex, and the amount of documents is large, as in Case 1, the standardization is a time-consuming and complex task. Some of the complexity in the case was caused by the large number of actors involved. This was also observed in Case 3 where negotiating the agreement of identification standards between several organizations was one of the main challenges. Similar findings are reported e.g. in [4].

If there is a great number of document authors involved, a major emphasis has to be given to the usability of authoring tools. The usability requirements were a significant challenge in Case 1, because the authors had to learn new ways for authoring, guided by logical structure of documents. Minimizing user resistance required some efforts. In Case 2, usability remained important, but building XML support the authoring tools helped the implementation. The form-based user-interface hided the logical structures from the authors. In Case 3 the logical document structures were hidden from the users by generating XML documents automatically from databases.

If the production of XML documents can be embedded in existing systems and processes within organizations, the need to change work practices decreases as in Cases 2 and 3. In Case 1, the authoring tool was new and cumbersome to use. Another major challenge was the adoption of the novel publishing tasks. Thus implementation in Case 1 was more challenging than in Cases 2 and 3 where corresponding changes did not occur.

5. Conclusion

The paper described three cases where the goal was to improve enterprise content management by the adoption of SGML/XML. Consistency in content management practices, automation of business processes, and more effective content reuse were important motivators of the adoption but the emphasis of the goals clearly differed in the cases.

The findings suggest that XML document implementation is domain-specific task that requires co-operation of people and organizations. Cases 2 and 3 show that it is possible to produce XML schemas and embed XML-based production into software and thereby lower the end-user resistance.If the benefits of the XML document production have been earlier demonstrated, the adoption of a novel system may be quite fast and fluent, as demonstrated in Cases 2 and 3. The XML standardization model by [13] was successfully utilized as the framework for analyzing the three cases. The framework may therefore be utilized as an analytic tool both for XML standardization development and for case research.

Our study focused only on document-like content of the ECM environment [2] and neglected the use of XML for metadata. Metadata standardization and implementation for XML document production is a possible avenue for further research, as well as the comparison of the findings with other studies.

6. Acknowledgments

The research was facilitated by the active cooperation and support of a great number of experts in the Finnish Parliament and Government, Faculty of Information Technology in University of Jyväskylä, and the international ICT service provider company of Case 3. The authors are grateful to all of them. Two anonymous reviewers provided invaluable comments for improving the paper.

7. References

[1] Bray, T., Paoli, J., Sperberg-McQueen, C. M., Maler, E., Yergeau, F., and Cowan, J. 2006. *Extensible Markup Language (XML) 1.1. (2nd Edition)* W3C Recommendation, W3C Consortium. http://www.w3.org/TR/xml11/ [February 8, 2007]

[2] Salminen, A. 2005. “Building digital government by XML”. In R. Sprague, Jr. (Ed.). *Proceedings of the Thirty-Eighth Hawaii International Conference on System Sciences.* Los Alamitos, CA: IEEE Computer Society, 122b- 122b.

[3] Salminen, A., Lyytikäinen, V., Tiitinen, P., & Mustajärvi, O. (2004). ”Implementing digital government in the Finnish Parliament”. In W. Huang, K. Siau, & K.K. Wei (Eds.), *Digital Government: Strategies and Implementation* (pp. 242-259). Hersley, PA: IDEA Group Publishing.

[4] Nurmilaakso, J.-M., Kettunen, J. and Seilonen, I. 2002. “XML-based supply chain integration: a case study.” *Integrated Manufacturing Systems 13 (8),* 586-595.

[5] Weitzman, L., Dean, S., Meliksetian, D., Gupta, K., Zhou, N., and Wu, J. 2002. “Transforming the content management process at IBM.com“, *Case studies of the CHI2002/AIGA Experience Design Forum*, ACM Press, New York, 1-15.

[6] Yin R. 1994. *Case Study Research: Design and Method*, Sage, Beverley Hills.

[7] Goldfarb, C. F. 1991. *The SGML Handbook.* Oxford, Oxford University Press.

[8] Smith, H., and McKeen, J. 2003. “Developments in practice viii: enterprise content management”. *Communications of AIS 11 (33)*, 1-26.

[9] Braa, K and Sandahl, T. 1998. “Approaches to standardization of documents”. In Wakayama et al. (eds.) *Information and process integration in enterprises: rethinking documents,* Kluwer Academic Publishers, Cambridge, Massachusetts, 125-143.

[10] Sandahl, T., and Jenssen, A. 1997. “The First Steps in Designing an SGML-Based Infrastructure for Document Handling”. *Scandinavian Journal of Information Systems, 1997, 9(2)*, 25–44.

[11] Müller, U., & Klatt, M. 2005. SCOPE – An XML Based Publishing Platform. In *Proceedings of the 8th International Symposium on Electronic Thesis and Dissertations.*

[12] Sefton, P. 2007. An Integrated Approach to Preparing, Publishing, Presenting, and Preserving Thesis. In *Proceedings of the 10th International Symposium on Electronic Thesis and Dissertations.*

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

III

FACING THE CHALLENGES IN IMPLEMENTING XML: THE CASE OF THE FINNISH PARLIAMENTARY DOCUMENTS

by

Nurmeksela, R. 2007

In M. Muñoz, A. Freitas, & P. Cravo (Eds.), Proceedings of the IASK International Conference E-Activity and Leading Technologies 2007, Porto 3-5 December, 247-255

Facing the challenges in implementing XML: The case of the Finnish Parliamentary documents

Reija Nurmeksela

Abstract — Transfer into production of XML documents in organizations requires collaborative efforts focusing on XML schema development for document types, redesign of work practices, as well as selection and customization of the software systems. The shift into using XML is a challenging move particularly if documents are produced by human authors. This paper describes findings from a case study concerning the implementation of XML on the Finnish Parliamentary documents. Difficulties and solutions for the problems in XML implementation are discussed from a human perspective and impacts of XML implementation to authors examined. Implementation has demanded extensive co-operation between parties involved and adoption of various international and national standards, recommendations and vocabularies. A participatory design method, custom-designed editors that hide document markup from the authors, keeping the structure of XML schemas simple, and organized training have been proved successful solutions to resolving resistance among authors. Most authors have adopted the new work practices, whereas for some the change has been insurmountable. Some of the latter have continued traditional authoring while having assistants convert the documents into structured format; others have changed position.

Index Terms — e-Government, Enterprise Content Management, Structured Documents, XML.

1 Introduction

In e-Government, a common objective is to make the content of public sector information repositories available on information networks, including the Internet, extranets, and intranets of particular organizations. Public sector activities are mainly based on the production and use of documents. Therefore document production practices impact the extent to which the content may be used to support e-Government goals. Development of content management is an important basis for e-Government.

Enterprise Content Management (ECM) considers the management of documents, metadata, actors, activities, and systems in production and use of the documents [1]. In ECM, a shift into using Standard Generalized Markup Language (SGML) [2] and recently its more streamlined subset XML has been thought to offer a technical basis for improving document management in organizations [3]. *XML (Extensible Markup Language)* [4], a core technology for Internet content, is a metalanguage that provides a way to exchange information between software applications in standardized formats. *XML documents* are structured documents where the structure definitions, document instances and layout specifications can be handled as separate content items. Structure definition for a document type is described with an *XML schema* that defines the logical structure and other constraints of the document type. XML schemas may be formalized by using, for example, Document Type Definition (DTD) [4] or XML Schema [5] languages. During authoring, the XML schema controls the logical structure of the document. The structure is marked up into the XML document as elements and attributes among character data.

For organizations, XML provides more efficiency compared to previous formats [6], [7], [8]. However, undertaking the production of structured documents in organizations may require a demanding standardization process closely related to the standardization activities on universal (for example W3C recommendations) and sectoral levels (for example JHS143 metadata recommendations for public sector organizations in Finland) [1]. *Standardization* signifies agreement upon rules which define the way information is represented in the documents of the domain [1]. These rules are expressed by XML schemas.

The implementation of standards demands extensive negotiations and co-operation

▪ *Reija Nurmeksela is with the University of Jyväskylä, B.O.Box 35 (Agora), 40014 University of Jyväskylä, Finland. E-mail: rekorhon@ jyu.fi.*

between different parties in a community (e.g. [1], [9]), and may require a major re-engineering of information systems and document production practices (e.g. [6]). It is obvious that document standardization is challenging when documents are produced by human authors [1]. However, there is not much research on document standardization from a human perspective.

This paper describes the case of the Finnish Parliamentary documents. It focuses on the challenges that may occur during the realization of XML in any organization where XML documents are authored by human beings. In addition, this paper discusses the impacts of XML implementation on document authors. The research was conducted by using a qualitative case study method [10]. The data for the paper were collected during years 2004-2007 by interviewing people involved in the implementation process, collecting data by a questionnaire targeted at people who participated in an ECM seminar organized for people involved in the content management activities of the Finnish Parliamentary documents, and analyzing documents, XML schemas, and previous reports of the case [1], [6], [11]. The aim of the case analysis is to increase knowledge of challenges and solutions for problems in the XML implementation process particularly from the document author's point of view.

The paper is organized as follows: In Section 2 literature concerning the implementation of XML in organizations is reviewed. Realization of XML implementation on the Finnish Parliamentary documents is described in Section 3. The challenges that occurred in implementation and solutions suggested for them are introduced in Section 4. Section 5 includes impacts on the authors with lessons learned and Section 6 concludes the paper.

2 Implementing XML Documents

XML documents may be produced in a number of ways either by human actors or automatically by software systems [12]. From the authors' point of view, transition to production of XML documents can be realized in three different ways: through a soft, guided or enforced standardization process [13]. The result concerning document authoring is different in each approach (see Table 1).

In *soft standardization* text may be authored as before, but other people markup the structure of the document afterwards [13]. In *guided standardization* the author marks up the content of the document with predefined styles by using a style editor included in the word processor. The author must be aware of which styles are allowed and in which logical order, because the style information forms the basis for further conversion of the document into the XML format. Custom-designed software developed for editing a certain document type may help in using the right styles. In realizing *enforced standardization* the document is authored directly in the XML format by using one of four alternative methods A-D (see Table 1). In each method the XML schema controls the document structure during or after authoring and the author must be aware of the logical structure of the document. In methods A and B the author marks up the text, whereas in method C the custom-designed editor hides the markup from the authors and the structure is defined by selections in the graphical user interface of the software. In method D, content is created into an XML database typically with XML editors [14] by using any of the methods A-C.

TABLE 1

Authoring Methods in Different Standardization Approaches

Standardization approach [13]	Authoring method [12],[13]
Soft	Converting content from other document format into XML format manually or partly automatically.
Guided	Using a word processor with style editor and converting document into XML format manually or automatically.
Enforced	A. Using a generic syntax directed editor. B. Using a generic word processor or Web browser with XML support. C. Using a custom-designed editor developed for certain document type. D. Creating the content into an XML database.

Direct human authoring in the XML format as a result of enforced standardization is different than traditional authoring. The main difference is that authors must be familiar with the logical structure of the document instead of layout. At first this might be confusing for authors [15]. Authors also have to be aware of the predefined structure when writing documents. This may be felt as restriction in their work [15].

The new XML-based ECM solution may require major changes in document processing, as has been noticed in reported cases (e.g. [6], [8], [13], [15]). Therefore, transition to structured authoring may cause many difficulties and resistance among authors. However, easy acceptances have also been reported (e.g. [7]).

3 STANDARDIZATION OF THE FINNISH PARLIAMENTARY DOCUMENTS

3.1 Finnish Parliamentary Documents

Finnish Parliamentary documents are produced and mainly used in the Finnish Government and the Parliament of Finland. The Government decrees governmental and administrative matters with the President of the Republic and the Parliament. Currently the Government consists of 12 ministries lead by the Prime Minister and 19 other ministers involving hundreds of people participating governmental activities. The Parliament consists of 200 Members and about 650 civil servants. The main task of the Parliament is to enact laws. It also prepares the state budget, handles EU affairs, and oversees the Government. All these activities demand interaction between the Parliament, the Government, and ministries.

Governmental and parliamentary work is deeply intertwined with the production and use of documents, which act as evidence, carriers of knowledge, and means of content sharing between the parties. The most important author groups in the Government are the ministers, civil servants, and secretaries. The major author groups in the Parliament are the Members of Parliament and their assistants, the committee councilors and their secretariats, and people in the Records Management Office. Published Parliamentary documents are records about reading of parliamentary matters. The documents are available to all citizens free of charge in register offices, libraries and on the Internet.

Parliamentary documents have a centennial history dating back to the establishment of the nation. During the years, some record types have been renounced and thus only exist in document archives, while some new ones have emerged. Currently the Parliamentary documents consist of 35 different record types. Annually tens of thousands of different record types are produced by hundreds of people. For example, in each parliamentary year the Government gives ca. 200-400 Proposals to the Parliament and ministers reply ca. 500-2000 Written Questions that are authored by Members of the Parliament. Documents produced during a parliamentary year are collected as part of printed publication series. In addition, some record types are available only in a digital format on the Internet. The process of producing parliamentary document types has been quite constant for decades. Major changes have been caused by the adoption of information technology: digitalization of the Parliamentary documents with word processors during the 1980s, and standardization of the record types with SGML in the 1990s and later on with XML.

3.2 The Standardization Process

Researchers at the University of Jyväskylä have participated in the standardization of Finnish Parliamentary documents. This section briefly reviews the standardization process reported earlier in [1], [6], [11] and updated for the paper by interviews and document analysis.

The standardization of Parliamentary documents began in 1994 through a project called RASKE, a collaborative effort with the Finnish Parliament, a software company, and researchers at the University of Jyväskylä. Also the Ministry of Foreign Affairs, Ministry of Finance, Prime Minister’s Office, and a publishing house participated in some phases of the project. The standardization was motivated by several problems in the management of Parliamentary documents. The standardization started in the RASKE project with analysis of parliamentary work during 1994-1998. SGML was chosen as the basis for standardization at the end of the analysis, and preliminary SGML DTDs were designed in the RASKE project.

In the Government, soft and enforced standardization [13] were selected as alternative approaches for standardization in the budgetary domain. A new SGML-based budgetary system was implemented in each ministry in 1998. Almost all authors selected the soft standardization approach instead of the enforced one. Various quality and efficiency problems during the manual conversion of word processor files into the SGML format led to a restriction in the standardization approach in 2004 to only allow enforced standardization, along with a replacement of the SGML-based budgetary system with an XML-based one. Currently, custom-designed software is used for authoring the State Budget Proposal. This software is based on Microsoft Word with XML support that is extended with a script-based user interface for hiding the XML markup from the authors. In the other three major domains (legislative work, handling of EU matters, and overseeing the Government), guided standardization has been selected as the approach in the

Government. Microsoft Word with a custom-designed style editor was implemented in each ministry during 2000-2001 for supporting structured authoring of nine record types.

In the Parliament, enforced standardization [13] was selected as the standardization approach in all domains. The implementation began in 1998 with the Report of a Special Committee as a pilot record type. This type was selected because there was a limited number of people involved in its authoring and because some committee councilors had shown strong interest and commitment in the improvement of document management.

Based on an analysis carried out in the RASKE project, three separate design projects were launched concerning (1) SGML DTDs and layout specifications, (2) re-design of work practices, and (3) selection and customization of authoring software and a new archiving system for structured documents. In addition a new tracking system was designed and implemented in the Parliament for supporting committee work. At the time, Finland's accession into the EU and reformulation of the constitution required changes in work practices in parliamentary work as well. The implementation process was quite fast: during 1998-2000 the authoring of all record types produced in the Parliament shifted into the SGML format. Currently, 25 record types are authored in the SGML format by using nine custom-designed applications of Adobe FrameMaker+SGML. In addition, some types of records authored in the Government are transformed into the SGML format for further editing or including them as parts of documents authored in the Parliament.

The standardization of Parliamentary documents has taken over ten years from 1994 and is still in year 2007 going on: Standardization of the record types produced in the Government, shift from SGML to XML in the Parliament, and design of a common authoring system for the Written Question had been started. The main reasons for the continuation are various positive impacts of the standardization such as unified document structures, improved layout and more correct content of the documents, decreased dissemination of paper versions of the documents, saved publishing costs and improved accessibility to the parliamentary information. The positive effects are more deeply reported e.g. in [6], [11]. In addition, huge conversion costs caused by soft standardization, the recent development of XML editors particularly concerning usability issues, and the need for more consistent ECM practices motivate to continue the standardization although many difficulties have been faced and solved during past standardization activities.

4 Facing the Challenges of Standardization

In this section, experiences and challenges of the standardization of Finnish Parliamentary documents, and solutions for difficulties that arose are analyzed. The following analysis is based on the four major components of the ECM environment [16] i.e. actors, activities, content, and systems.

4.1 Actors

In an ECM environment an actor may be an organization, a person, or a software agent acting behalf of a person in an organization [16]. In this case, the first two are concerned.

Organizations. The main challenge in standardization has been co-operation between the organizational actors. Most of the record types are shared between the organizations, but wishes and work practices have varied particularly between the ministries. In addition, XML and related technology is not equally streamlined for each actor. As a result, different document production technologies and practices have been adopted in different organizations. This means that technical corrections are sometimes needed when, for example, documents, which are first produced in the ministry and then handled in the Government, are further passed from the Government to the Parliament. In order to harmonize different ECM and other IT-related practices, a new State IT management organization was established for defining common policies, practices and technologies. Also several cross-organizational projects have been initiated.

Persons. In addition to co-operation at the organizational level, willingness and commitment of people particularly at the managerial level has been seen an important means for adopting common practices. However, lack of SGML/XML knowledge among participants and finding a common language between the developers of ECM practices (such as IT experts and consultants), the managers and the authors has been difficult in the beginning of both SGML and XML standardization activities. Accumulating experience alongside the progress of the SGML/XML adoption has improved knowledge and communication.

Particularly, understanding the idea and benefits of structured documents has been difficult for some authors. The difficulties concretize in the training of new authors. Motivation for structured authoring is also problematic to some authors if they do not benefit from the markup at all. The concrete benefits are realized, for example, in information retrieval which is not a typical task of these authors. Improvement of usability issues concerning XML editors alongside the transfer from SGML to XML has been considered a solution for supporting and motivating authors. In addition, training has been organized.

4.2 Activities

Activities in an ECM environment consist of development and deployment activities [3] concerning business process and content management activities [17]. This paper focuses on development and deployment of content management. This section includes citations of the answers to the questionnaire. The author has translated them from Finnish to English.

Development activities. In this case, several changes concerning work practices have been going on parallel to standardization. For example, in the Parliament a new tracking system was introduced and work processes were re-designed because of the new constitution and Finland's accession into the EU. The situation was frustrating for the authors and reflected as resistance concerning the re-design of work practices that resulted from standardization. Also the authors' commitment to and participation in the standardization process was troublesome. The manager of the Records Management Office of the Parliament illustrated the situation:

> *"In the beginning a work model for transforming previous work practices into the one needed in the production of structured documents was problematic. Organizing workshops, where we first discussed current document production practices and then chose the activities that were involved in the production of structured documents, improved the situation."*

One document author, a secretary working in the Records Management Office, saw the situation as follows:

> *"There was no introduction to the idea of standardization, but a total lack of information concerning the background of the change. We participated in development work a long time without knowing what we were doing and why…*
>
> *…The process was hard, but as the understanding increased, benefits of the structured documents became clearer, too. The process incurred significant changes into the document production practices."*

The problematic situation was reflected in the work of IT consultants, as one consultant illustrated:

> *"Sometimes it took a while in the beginning of the workshops until general debate about the standardization process between different parties subsided. The only thing you could do was wait until it was possible to continue development work in the session."*

The main reason for the problems confronted was a lack of knowledge about the standardization process. This reflected as difficulties to separate the right and the wrong decisions beforehand and to estimate other possible faults. The lack of knowledge also caused waste of development work during the standardization. Accumulating experience together with participatory design method improved the situation.

Document production activities. The main problem concerning the document production practices arose because a representative of the printing office was ignored from the workshops. This reflected as a need for manual work in editing. Also the quality of prints was very poor initially after implementation. Manual work is still needed since all Parliamentary documents published during the parliamentary year are compiled into a printed part of a collection series. Renumbering the pages of an enormous amount of documents is the most resource-demanding activity. This problem could have been avoided if the whole document management process involving all activities to the very end of the process had been focused on.

4.3 Content

Content in an ECM environment consists of addressable parts of stored data, such as documents, web pages, and content of databases. Content may be clustered into documents and metadata. [16]

Documents. Currently only one record type, the State Budget Proposal, is authored in the XML format in the Government, whereas guided standardization is the approach in other domains. Conversion problems sometimes occur because wrong styles are included in the documents. As a solution, all record types should be authored in the XML format. However, authors of the Finnish Parliamentary documents are a heterogeneous group of people. Thus, it has been difficult to define structures of SGML DTDs and XML schemas clearly enough for all authors to understand them the same way. Written guidelines for the most important record types are created and training organized in order to increase common understanding.

One major problem in marking up correctly structures of the documents has arisen due to excessively complicated logical structures defined into XML schemas. This has also reflected to difficulties in design of customized XML authoring tools. Thus in standardization, the number of elements in the XML schemas should be minimized. This will make both the work of the authors and maintenance of the XML schemas more efficient. In a standardization process, organizations should also be prepared that XML schemas are only ready after some time of operational use. For example in this case, some content that was difficult to include in any structure of the XML schema was found only after implementation. This required a restructuring in some XML schemas. However, changing the XML schemas after implementation has caused even more problems: resistance against the change from the authors and a question of whether the structure of existing documents should be changed as well to conform to the changed XML schema?

However, although many record types of the Parliamentary documents are currently produced in a structured format, utilization of the structural characteristics of the documents in its full potential is limited. For example, in the Internet services, navigation into the sections of the documents could have been offered for users. Also digital signatures that would allow delivery of documents in the digital format instead of paper could have been adopted.

Layout of the Parliamentary documents is very important. Although the idea of XML schemas is to describe only the logical structure of the document, in this case definition of structures for formatting purposes only is required for generating the desired layout. Negotiations were needed for including these “layout structures” into XML schemas because some parties resisted the inclusion.

Metadata. Although some local level standards for metadata are defined (such as metadata recommendation JHS 143, an application of Dublin Core [18] for the Finnish public sector), these are not widely adopted in Parliamentary documents. In addition, varying values of the same metadata items between documents and systems constitutes a major problem. For example, an identifier of a document in the Government may be “K1/2006 vp”, whereas in the Parliament the value of the same identifier is “K5/2006 vp”. As a solution, metadata recommendations should be adopted, and values of the metadata elements should be harmonized between organizations. These are also the key issues when adopting the Semantic Web [19] for content management.

Some authors find the creation of metadata for documents additional and frustrating work, particularly if they think certain metadata elements are unuseful. Thus the number of metadata elements should be minimized and automatic creation of metadata supported. This is possible by data integration, for example if metadata is included in parts of the documents.

4.4 Systems

Systems consist of hardware, software, and standards used to support the operation of ECM activities [16]. The last two of them concern document authoring.

Software. The main problem in late 1990s, when the first standardization activities took place, was a few alternatives of editors for SGML. The selected tool, Adobe FrameMaker+SGML, has several problems: it has not been localized into the Finnish language, its support functionalities are clumsy to use, and it does not support the content management process. Currently nine custom-designed applications of FrameMaker are used for authoring support. The customization includes, for example, connection to databases for selecting reused content, such as names of the Members of the Parliament, semi-automatically. However, authoring with the tools should be even easier. With the emergence of XML, the selection and usability of available tools has evolved significantly. Currently an extended version of Microsoft Word customized for editing the State Budget Proposal is used in the Government. However, two different editors in a common document production process cause sometimes problems. For example, it is not possible to automate the combining of content produced with different editors. Instead, manual editing and converting is needed. Ongoing design of common authoring systems and adoption of XML will automate the process.

Standards. Standardization of the record types and metadata, as well as development of recommendations and common vocabularies is central for ECM development efforts concerning the Finnish Parliamentary documents. Although much work has been done regarding these issues, problems still arise in some areas. For example, there has been a lack of recommendations for naming conventions concerning elements of XML schemas and metadata. Thus, ad hoc values are used. In addition, difficulties have been faced in adapting international and national standards and recommendations that cannot be adopted directly. As a consequence, standardization work at the local level has proved to be an enormous and continuing process where changes in IT, organizations, and processes cause constant updating needs and develop new ones [1].

5 Discussion and lessons learned

The standardization of the Finnish Parliamentary documents has effected positively inside the organizations involved as well as in the national level (see e.g. in [6], [11]). From the citizen point of view the standardization has improved availability of parliamentary information and thus supported e-Government goals. The wide impacts in the legislative domain are discussed in [6]. However, the standardization has impacted the internal work of the Finnish Parliament and Government.

From the organizational perspective, standardization has created new content management tasks, particularly concerning publishing, whereas some tasks have disappeared through automation [6]. In addition, new roles have emerged as in cases reported in [13]. However, standardization has not impacted the work practices of all document authors: Although guided and enforced standardization are the standardization methods currently used in the Government and the latter in the Parliament, some authors continue to produce text in the non-structured form. In fact, there have been three alternatives for authors in facing the changed work practices:

1. The majority of authors has adopted the new work practices through learning to use new custom-designed editors.
2. Some authors continued to use the same word processor they had learned to use before the beginning of standardization. As in soft standardization [13] their secretaries convert the documents manually into the SGML/XML form with editors developed for structured editing. The authors who selected this alternative belong to a group that is responsible for handling parliamentary matters. Examples include the Members of the Parliament and some civil servants in the ministries.
3. A few people have not adopted the new work practices and have not had the possibility to continue traditional authoring. As an impact they have changed position.

Authors who joined the organizations after the standardization have selected the first alternative because in the beginning of their career in the new organization they had to learn all of the work practices of the organization including structured editing.

In this case, standardization progressed parallel in different domains and iteratively from soft and guided to enforced standardization. However, the progression has varied between organizations. Therefore, this case reveals that different standardization approaches may be adopted in parallel with each other in a community depending on organizations but also the roles of people. Thus, in contrast to cases reported in [13], different standardization approaches are not mutually exclusive alternatives. Another contrast to the case reported in [13] is the emergence of new roles also in soft standardization. In this case, some secretariats adopted new role of technical writer, whereas in the other case [13] the parallel organizational unit for technical editing was allocated.

Implementation of SGML/XML on Finnish Parliamentary documents has demanded extensive co-operation between parties involved. The findings are similar to [9] who reported on XML implementation for data integration standards between software systems. However, if document types authored by human beings are standardized as in this case, co-operation and common language between the participating organizations as well as people acting in different roles are essential for successful solutions.

The implementation of structured documents has not been easy, as in the case reported in [7]. Instead, several difficulties have been encountered during the over decade-long process from the first analysis phases to current operational use. As in the case reported in [8], a participatory design method was used as a solution for resolving difficulties with document authors. It is possible to decrease the resistance toward structured authoring also by data integration between systems concerning reused content. Despite of a participatory design method and custom-designed editors supporting structured authoring, the redesign of work practices has been a difficult process for many authors. Even after a decade from the beginning, some authors continue to regard standardization difficult.

Regarding documents, keeping the logical structures of the XML schemas simple supports document authoring and the maintenance of XML schemas. The first finding is similar as in [15]. However, also guidelines and training are needed. For the desired layout some “layout structures” may be needed. Therefore, the fundamental separation of the logical structure and the layout is not always possible.

Finally, a major lesson learned from this case is that the development of local standards for document types requires adoption of numerous international and national standards, recommendations, and vocabularies. Provision of adoption guidelines helps to avoid ad hoc solutions and is central to successful and unified implementations of XML.

6 CONCLUSIONS

Implementation of XML, a core technology for Internet content, in an organization requires demanding standardization process if documents are produced by human authors. This paper presented analysis of the standardization of Finnish Parliamentary documents. Experiences and difficulties concerning entities of an ECM environment were discussed particularly from the document author’s point of view. These include issues that may be confronted in any XML implementation case where XML documents are authored by human beings.

The XML implementation has been a process that has lasted for over a decade and it has iteratively changed the work of the authors significantly. During the process, many difficulties have been faced regarding entities of the ECM environment [1] i.e. actors, activities, content and systems. The emphasis of the paper has been on the social issues because of the characteristics of the data, but also technical problems have been faced. By finding out solutions for the confronted difficulties through negotiations, common experience and training, the implementation of XML has improved collaboration and unified content production practices among organizations involved in the production and use of Finnish Parliamentary documents.

ACKNOWLEDGMENT

The research would not have been possible without the contribution of the RASKE2 project [20] that involved active collaboration and support with numerous experts working in the Finnish Parliament and the ministries. The author is grateful to all of them.

REFERENCES

[1] Salminen, A. 2005. “Building digital government by XML”. In R. Sprague, Jr. (Ed.). *Proceedings of the Thirty-Eighth Hawaii International Conference on System Sciences.* Los Alamitos, CA: IEEE Computer Society, 122b- 122b.

[2] Goldfarb, C. F. 1991. *The SGML Handbook*. Oxford, Oxford University Press.

[3] Tyrväinen, P., Päivärinta, T., Salminen, A., & Iivari, J. 2006. Characterizing the evolving research on enterprise content management. *European Journal of Information Systems, 15 (6)*, 627-634.

[4] Bray, T., Paoli, J., Sperberg-McQueen, C. M., Maler, E., Yergeau, F., and Cowan, J. 2006. *Extensible Markup Language (XML) 1.1. (2nd Edition)*. W3C Recommendation, W3C Consortium. http://www.w3.org/TR/xml11/ [February 8, 2007]

[5] Thompson, H. S., Beech, D., Maloney, M., & Mendelsohn, N. 2001. *XML Schema Part 1: Structures.* W3C Recommendation. Available at http://www.w3.org/TR/xmlschema-1/. [February 16, 2007]

[6] Salminen, A., Lyytikäinen, V., Tiitinen, P. and Mustajärvi, O. 2001. “Experiences of SGML standardization: The case of the Finnish legislative documents”. In R. Sprague (Ed.). *Proceedings of the Thirty-Fourth Hawaii International Conference on System Sciences.* Los Alamitos: IEEE Computer Society.

[7] Ray, D. and Ray, E. 2001. “Maintenance Procedures for a Class of Warships: Structured Authoring and Content Management”. *Technical Communication 48 (2)*, 235-247.

[8] Weitzman, L., Dean, S., Meliksetian, D., Gupta, K., Zhou, N., and Wu, J. 2002. “Transforming the content management process at IBM.com“, *Case studies of the CHI2002/AIGA Experience Design Forum*, ACM Press, New York, 1-15.

[9] Nurmilaakso, J.-M., Kettunen, J. and Seilonen, I. 2002. “XML-based supply chain integration: a case study.” *Integrated Manufacturing Systems 13 (8)*, 586-595.

[10] Yin R. 1994. *Case Study Research: Design and Method*, Sage, Beverley Hills.

[11] Salminen, A., Lyytikäinen, V., Tiitinen, P., & Mustajärvi, O. (2004). Implementing digital government in the Finnish Parliament. In W. Huang, K. Siau, & K.K. Wei (Eds.), *Digital Government: Strategies and Implementation.* Hersley, PA: IDEA Group Publishing, 242-259.

[12] Salminen, A., Nurmeksela, R., Lehtinen, A., Lyytikäinen, V., & Mustajärvi, O. 2006. Content production strategies

for e-Government. In A.-V. Anttiroiko & M. Malkia (Eds.). *Encyclopedia of Digital Government.* Hersley, PA: IDEA Group Publishing.

[13] Braa, K and Sandahl, T. 1998. Approaches to standardization of documents. In Wakayama et al. (eds.) *Information and process integration in enterprises: rethinking documents,* Kluwer Academic Publishers, Cambridge, Massachusetts, 125-143.

[14] Salminen, A., & Tompa, F.W. 2001. Requirements for XML document database systems. In E.V. Munson (Ed.), *Proceedings of the ACM Symposium on Document Engineering (DocEng '01)* (pp. 85-94). New York: ACM Press.

[15] Sandahl, T., and Jenssen, A. 1997. "The First Steps in Designing an SGML-Based Infrastructure for Document Handling". *Scandinavian Journal of Information Systems, 1997, 9(2)*, 25–44.

[16] Salminen, A. 2003. Document analysis methods. In C.L. Bernie (Ed.), *Encyclopedia of Library and Information Science, Second Edition, Revised and Expanded.* New York: Marcel Dekker, 916-927.

[17] Salminen, A. 2006. Sisällönhallinnan menetelmiä. In Nurmeksela, R., Virtanen, M., Lehtinen, A., Järvenpää, M., and Salminen, A. *Information management in the Finnish legislative work – Towards Semantic Web of legislative information resources.* The Parliamentary Office.

[18] *Dublin Core Metadata Element Set, Version 1,1.* 1999. Dublin Core Metadata Initiative.

[19] Berners-Lee, T., Hendler, J. & Lassila, O. 2001. The Semantic Web. *Scientific American 284(5),* 34–43.

[20] Nurmeksela, R., Virtanen, M., Lehtinen, A., Järvenpää, M., and Salminen, A. 2006. *Information management in the Finnish legislative work – Towards Semantic Web of legislative information resources.* The Parliamentary Office.

Reija Nurmeksela is a PhD student in the Department of Computer Science & Information Systems at the University of Jyväskylä. She works as an XML integration specialist at an international ICT service provider company. The topic of her dissertation is XML support for Enterprise Content Management.

IV

TOWARDS CONTENT INTEGRATION IN DOCUMENT PRODUCTION

by

Honkaranta, A., & Nurmeksela, R. 2007

In K. Soliman (Ed.), Information Management in the Networked Economy. 8th IBIMA Conference on 20-22 June in Dublin, Ireland.

Towards Content Integration in Document Production

Anne Honkaranta, Ph.D, lecturer (*), anne.honkaranta@it.jyu.fi
Reija Nurmeksela, M.Sc., Ph.D student (*), rekorhon@jyu.fi
(*) University of Jyväskylä, Department of CS &IS, Jyväskylä, Finland.

Abstract

In document-oriented business processes effective document production requires integration of metadata as well as other existing content into the document to be produced. Based on findings from literature and two research projects on e-Government this paper describes requirements for content integration in document production. It also proposes a model for integrated document production. The model consists of document architecture and integrated document production process which may utilize the Semantic Web technologies. The model is demonstrated by an exemplar process of making a statement. XML is used as an enabling technology for integrated document production.

1. Introduction

Contemporary enterprises move towards adopting the Semantic Web [1] for content management. Processes and content may be considered as the two sides of the same coin, the development of content management requires enhancements on business processes, too [2, 3]. Metadata management is an integrated part of content production [4] and provides the glue for bridging the content and process management. Enterprise content management (ECM) is an integrated perspective for the management of content and metadata together for document production, storage, publication, and utilization in organizations [5].

A major portion of business processes are based on production, capture, and use of documents, which act as containers of knowledge for administrative actions in e-Government [6]. Thus, content management during document production provides a base for efficient management of these document-oriented business processes. While persons produce documents they typically integrate some existing content with the new one to produce a new document. In document-oriented business processes effective document production requires integration of metadata and other existing content into the documents to be produced.

The paper describes the requirements for integrating metadata and other content into documents in e-Government organizations and proposes a model for integrated document production. Requirements are based on the findings from literature and two research projects in different e-Government organizations. Section 2 presents the model for content management and discusses the role of metadata for ECM. Research methodology is described in section 3. Section 4 introduces requirements for integrated document production. Section 5 presents the model for integrated document production in which the Semantic Web technologies such as XML [7] and RDF [8] are used as an enabling technology for the model. Section 6 concludes the paper.

2. Enterprise Content Management and Metadata

ECM may be examined by the content management model [5] depicted in Figure 1. The model presents content management environment as a construction of two types of entities, activities and information resources, and information flows between them. Activities are depicted by the oval, information resources by rectangles, and information flows by arrows.

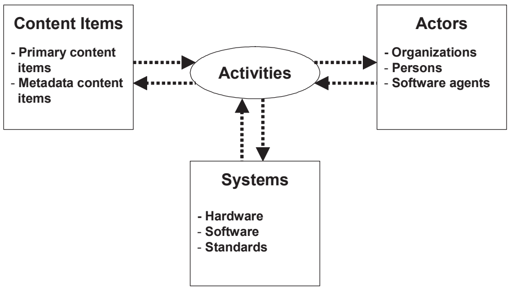

Fig 1. Content management model

An *activity* consists of actions performed by one or more actors in a business process. The process may be for example a legislative process or preparing a statement. The information resources are divided in three types according to their different roles in the activities: actors, systems, and content items. An *actor* may be an organization, a person, or a software agent acting behalf of a person on an organization. In a legislative process, for example, Parliament is one organizational actor and a Member of Parliament may be a role of a person. *Systems* consist of hardware, software, and standards used to support operation of the activities. *Content items* are addressable parts of stored content, such as documents or web pages. Content items may be clustered into *primary content items* and *metadata content items*. *Metadata* offers information about primary content items and about their production, storage, and use environments.

In e-Government metadata management has a long tradition. Official documents have strict requirements for their accuracy and integrity. Commonly in e-Government the documents are managed as records which consist of the documents and metadata about them [9]. In ECM metadata is intended to support system integration, information retrieval, and collaboration of people in work processes [10].

In ECM, *contextual metadata* provides information about the context where the document is produced or used [5]. Thus, it contains information about the entities of content management model [11], and may be further divided into document metadata, process metadata, actor metadata, and systems metadata according to the entities. Terms of metadata element sets in metadata recommendations used in public sector involves metadata about these four entities. Examples of these recommendations are Dublin Core [12] onto which a number of national metadata recommendations such as e-Government metadata standard in UK (http://www.govtalk.gov.uk/), AGLS Recordkeeping metadata standard in Australia (http://www.naa.gov.au), and JHS 143 in Finland (http://www.jhs-suositukset.fi/) are based on. Contextual metadata types related to content management model with examples from Dublin Core and JHS 143 are depicted in figure 2.

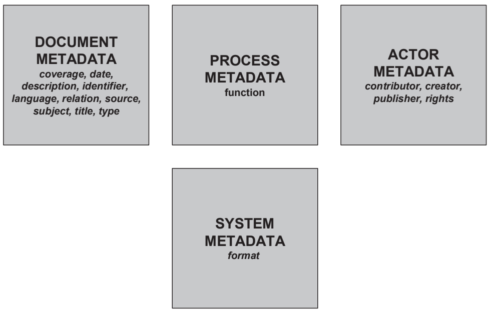
DOCUMENT METADATA
coverage, date, description, identifier, language, relation, source, subject, title, type
PROCESS METADATA
function
ACTOR METADATA
contributor, creator, publisher, rights
SYSTEM METADATA
format

Fig 2. Metadata classification for content production

Document metadata provides information about the document types and document instances. It is typically stored in ECM systems. *Process metadata* links a document to a certain activity in an organization. It provides means for identifying the documents created in a certain business process. Thus it offers the possibility to integrate the processes with documents and vice versa. *Actor metadata* links the actors with processes and documents. For example, an actor may act as a creator, editor, receiver, or publisher for a certain document type, or as an owner of a certain kind of process or activity. Actor metadata may be stored in personnel or ECM system.

In contemporary ECM environments metadata needed in document production and content items to be reused may be fragmented into a number of systems. *System metadata* provides information about the systems in which the primary content and metadata are stored.

In e-Government a lot of metadata overlaps between documents and systems as shown on the following sections. Thus, systematic metadata solution offers possibility to automate collecting, combining and extracting the metadata for document production. *Integrated document production* refers to a processing system in which documents are composed of metadata content items and, reused and new primary content items. If metadata is available via systems such as process and ECM systems, it may be collected and combined automatically in integrated document production. At the end of the process metadata may again be extracted to the process and ECM systems from the documents.

In the following we describe two action research projects where we have developed a general model for integrated content production.

3. Methodology

The research is constructive by its nature. The requirements for the integrated document production model are based on the findings of two action research projects; RASKE2 in the Finnish Government and Parliament of Finland on 2003-2006 and RAKE on 2005-2006 at Finnish Centre for Pensions (FCP). Following subsections describe the organizations and the research carried out with respect to the analysis and findings presented in the paper.

Case Organizations

The Finnish Government convenes governmental and administrative matters with the President of the Republic and the Parliament. Currently the Government consists of 12 ministries lead by the Prime Minister and 19 other ministers. The main task of the Parliament is to enact legislation. It also prepares the state budget, handles EU affairs, and acts as a forum for political debate demanding interaction between Parliament and the Government. The Parliament consists of 200 Members and about 650 civil servants. Legislative and governmental work is deeply intertwined with the production and use of documents, which act as evidence, carriers of knowledge, and means of content sharing between the parties. Annually thousands of documents of different document types are produced. Published governmental documents are available to all citizens free of charge in register offices or libraries and also on the Internet.

The Finnish Centre for Pensions (FCP) acts as a central body for numerous private pension institutions in Finland. It is overseen by the Ministry of Social Affairs and Health and supervised by the

Insurance Supervision Authority. There are nearly 400 employees working at FCP, primarily lawyers, pension schema experts, and pension register system experts. FCP is an expert organization carrying out multiple types of tasks. FCP provides its expert assistance for the preparation of pension-related laws and norms, and produces estimates on the pension use and coverage. Private pension institutions carry out pension-related tasks in a decentralized way, and need FCP's advices, Internet service and circulars for acting out in a consistent and coherent way. PCP also provides information for private persons, as well as guidance via telephone, e-mail, paper-print documents, and on the Internet.

The Two Action Research Projects

The research presents an analysis of similarities of the requirements for content integration and for the model of integrated document production. The analysis was carried out by the authors by comparing and analyzing the findings of the two separate research projects; RASKE2 and RAKE. The first author of the paper was involved on the RAKE project, and the second one on the RASKE2 project. Albeit the original goals of the projects were differing, the authors were able to identify similar patterns of document use, and requirements for content integration in document production on the both organizations. The requirements for content integration on document production were further defined as a preliminary model for integrated document production.

The research projects were carried out as action research [13]. Therefore the research projects are described by the phases of the action research cycle: diagnosing, action planning, action taking, and evaluating. The last phase of the action research – specifying learning – considers the findings for the paper and is therefore explained in more detail on the remaining sections of the paper.

The RASKE2 originally aimed for developing methods for integrating information resources by developing schemas and practices for metadata management in organizational networks [5] [14]. Selected domain where methods were development was Finnish legislative process. For demonstrating benefits of consistent metadata practices a prototype of Semantic Legislative Portal [15] was developed. However, the project also considered other aspects, such as how to integrate document and metadata production. Table 1 summarizes the activities carried out at RASKE2 project with respect to the aim of this paper.

In the diagnosing phase metadata related to Finnish legislative process and use of XML documents was analyzed. The analysis revealed that metadata of a legislative process was actually a source of content for a multitude of documents and provided business rules for document content integration.

The action taking phase considered testing the finding by preparing a demonstration with the legislative organizations exemplar process and document content, and proposing an integrated document production model.

The evaluation considered analysis of the demonstration and the responses received from the legislative organizations. The evaluation supported the preliminary findings for the requirements for integrating metadata about business processes and documents into document content, thus illustrating the requirements for a model of integrated document production.

Table 1: The Action Research Cycle on the RASKE2 project

Phases of Action Research	**Research and Development in RASKE2**
Diagnosing	Analysis of metadata on the domain. Defining the current use of XML schemas.
Action planning	Analyzing the metadata and its use for XML documents. Planning a demonstration for illustrating the benefits of integrating metadata of processes and documents into document production.
Action taking	Suggesting import and extraction of metadata to and from XML document contents. Presenting a demonstration of metadata extraction to documents.
Evaluating	Content is largely reused across document types, thus content integration is essential. A great deal of some document type's content consists of document and process metadata.

FCP has been active in developing its content management, which has also included document redesign [16]. The RAKE project was a part of this continuous development. It was originally targeted for analyzing the benefits of XML and the use of document-type specific schemas for document management at FCP, and for proposing a method for XML document management and development.

One of the requirements posed by the FCP was that the production of XML documents should not require expensive, XML-specific editor. Thus, a part of the project also considered evaluation of Office 2007 Beta for XML Document production at FCP. The phases of the RAKE research with respect to the aim of the paper are described on Table 2. On the table the content component refers to the topical content or element-type level content of a document type. For example, a memo document may consist of information components of "list of participants", "agenda", a number of "items" and "signature".

Table 2: The Action Research Cycle on the RAKE project

Phases of Action Research	**Research and Development in RAKE**
Diagnosing	How may FCP utilize XML documents? Findings from previous studies on XML implementation as well as the document management efforts carried out at FCP.
Action planning	Making a list of document types relevant for the project. Planning the workshops for analyzing the needs for document content reuse and for redesigning the document types.
Action taking	Organizing workshops for analyzing the document types and their content components, as well as the requirements for content component reuse. Preparing a demonstration of integrated XML document production.
Evaluating	Document types have a multitude of information sources containing reusable content for documents, such as law texts, address databases, document metadata and process metadata. There are requirements for reusing metadata about business processes, document types and documents, i.e. the metadata stored on the Process and Document Management systems.

As described on the table 2 the RAKE project analyzed the content of a number of (8) document types used at FCP to find out requirements for the use of XML and possible benefits and pitfalls for XML implementation to the organization. The contemporary document production environment as well as the ECM system (Hummingbird) was studied. The content components of the document types were analyzed and partially redesigned for more consistent document type use, and for analyzing the requirements for content reuse between documents and across information resources available at FCP.

4. Requirements for Integrated Document Production

Following subsections describe the business and content management requirements by providing an example of a business process in e-Government; a process of making a statement.

An Example of a Process of Making a Statement

Figure 3 depicts a process of making a statement – a common process within e-Government. The notation used follows the RASKE modelling method [17]. In the figure the activities of the process are depicted by ellipses. A dashed arrow connects the metadata ("md") or document ("D") required or produced on a work activity. The solid arrows between the ellipses define the starting order of the activities.

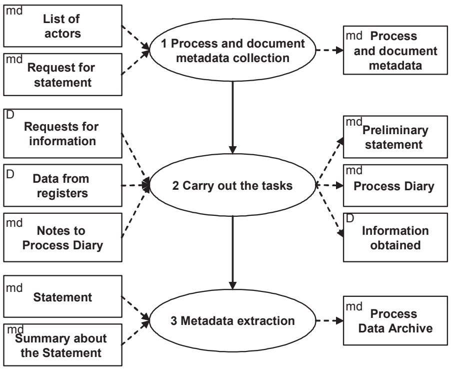

Fig 3. A Process of Making a Statement

The process of making a statement starts when a person or an organization makes a request for a statement. The process has three activities: metadata collection, an activity consisting of several business tasks and metadata extraction. There are two main kinds of business tasks: to produce or revise a document, or to collect and save information needed on the process. The information needed on the tasks may be received from a number of other authorities or it may reside on an internal or external database. For example, a process owner may look up the amount of pension paid for a person, and store it as a document (pension account) along with snippets of information dealing with, for example, who has been contacted (via mail or phone) to fill information into the process diary.

Several kinds of metadata about a process may be described. For example, an owner of a process (the actor taking care of a matter) may be identified by the expertise needed. Pension provisioning matters may involve persons on the accounting. A majority of document types used and produced on the process may be known as well. For example, the Figure 3 lists the document types used and produced on the process of giving a statement. This information may be stored as metadata of a process type. A list of actors of a certain process may depend on the process instance, but as the process is instantiated, the list should contain the names and addresses of the persons and organizations related to the matter, i.e. the potential receivers of documents and information requests. The business requirements for integrated document production therefore command that actor, document, and process information used and produced on the processes should be recorded and made available as metadata if the information needs to be reusable for ECM.

Content Management Requirements
E-Government document types used and produced on the processes have a multitude of content sources that are needed on document production, such as law texts, addresses (accessed via Web Services or databases on the intranet), repeating phrases, and references to normative guidelines. The task of a document producer is mainly to combine information from the sources to produce a document that fits to a task at hand. The finding was underlined by the document producers by their statements considering document production "a great deal of content for a document is available in registers and on the diary in which the activities of the process matter are registered…it is error-prone and tedious to copy and paste the information time and time again into documents…could it not be automatized?" (a snippet of a conversation on the document information component analysis workshop). Therefore, the document production requires integration of existing content available on other content sources with the novel content when producing documents.

E-Government organizations typically follow international and national standards for their document content as well as their layout. There are also norms and governmental organization networks that mutually define the requirements for document content and layout as well. For example, the names of the actors involved on a process (matter) should be included in documents below the document identifiers, such as document name and document creator.

Metadata such as document identifiers and names as well as other document metadata may be stored on ECM systems, and the metadata about process and actors on the process management system or other kind of systems. When appended with metadata about reusable content items, such as normative text for a document, a majority of the content needed to produce an instance of a document is available via systems.

5. A Model for Integrated Document Production

Following subsections describe the proposed model for integrated document production. First composite architecture of e-Government document is presented, and then integrated document production process is described. Finally, use of the model is illustrated by presenting an example of document production where XML is used as an enabling technology for the model.

Document Architecture
In ECM environment (figure 1) content items are divided in primary and metadata content. In e-Government the content of a document consists of primary and metadata content that are retrieved from multitude of content sources as discussed above. Thus, content may also divided in reused and new content. We consider metadata as reused content that may be further divided into process, actor, and document metadata according to the metadata classification discussed in section 2. Primary content is either reused or new one. Figure 4 presents the composite architecture of a document.

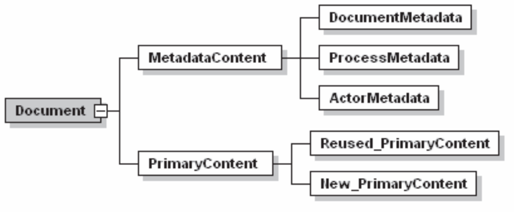

Fig 4. Composite architecture of a document

Process metadata consists of information related to a process, such as process id, name, type, and information of the document types used and produced on process activities. It may be accessible from a process management system as a metadata and presented for example as XML or RDF/XML format.

Actor metadata consists of the information about the actors related to the produced document. It may contain possibly a list of persons who may accept, publish, review, or create it be accessible from the ECM system. It may also contain, for example, a list of Members of the Special Committee who have participated to the creation of a Committee Statement [11].

Document metadata consists of metadata related to document types, such as a name for a document type, XML schemas and content components related to it, and of a list of content sources for it. Document metadata may also contain metadata of a document instance; for example the name or role of the actor who checked the document out from ECM system, and metadata values (s)he defined.

A document template may utilize multiple content sources in order to combine reusable content components into a document while it is being created or manipulated. As an example, the primary content to be reused may involve a model text for a document type containing commonly used statements and prototypical text for a document type. The model text may be attached to a document template or pre-filled to a document instance from an external text database at the time a document instance is created. There may also be other primary content types such as law texts, which may be made accessible for a person editing the document via external databases or Web Services.

Together the metadata content, reused primary content and new primary content construct a model for composite document architecture.

Integrated Document Production Process

Figure 5 shows a process model for integrated document production. The circles in the figure depict activities and the arrows show the control flow specifying the order for starting the activities. The small black circle indicates that all of the activities may begin in any order.

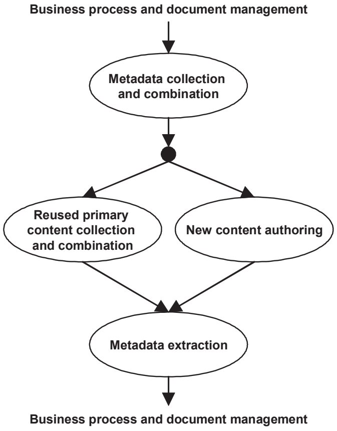
Business process and document management
Metadata collection and combination
Reused primary content collection and combination
New content authoring
Metadata extraction
Business process and document management

Fig 5. Integrated document production process

Production of documents that include reused content involves collection and combination of the reused content with the new content [18]. Metadata included in document to be produced is mainly known in the beginning of the document production process and stored in systems. Thus, the process begins with metadata collection and combination phase. In the phase, process, actor and document metadata is retrieved from systems, transformed into the structure of the document type to be produced, and integrated into the right document structures. Then a pre-filled document is generated and provided to the author to be completed with primary content. The author collects and combines reused content parallel with creating a new content. When the document is completed, metadata needed in systems used in the domain is extracted in metadata extraction phase.

An Example of Integrated Document Production

Utilization of structured content production strategy, where documents are produced in a structured form such as XML documents, [10] facilitates semi-automatic integration of reused content items (e.g. [19]) and creation of metadata (e.g. [20]) in a single system. In XML documents, the logical structure is explicitly marked up in the document. The logical structure of a document type may include content originating from metadata sources and unique text content for the document (figure 4). The mark up of XML document follows the one defined by XML schema [21] for the document type. XML schemas may define both primary and metadata content items [5]. Thus transformation between primary and metadata is possible, for example, by using XSLT [22] and XPath [23] languages.

The following example of integrated document production model illustrates how the reusable text components for certain document types, metadata, and novel text content are tied into the process of integrated content production of XML documents. In a legislative process, the document production model may be applied, for example, in the production of agendas and memorandums. It also shows how portions of reusable content components for a document type are utilized to pre-fill the document content. The phases of integrated document production model are illustrated in Figure 6. The notation follows the RASKE modelling method [17].

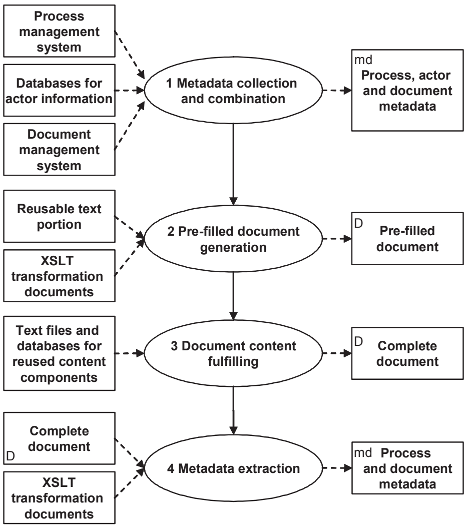
Process management system
Databases for actor information
Document management system
1 Metadata collection and combination
md
Process, actor and document metadata
Reusable text portion
XSLT transformation documents
2 Pre-filled document generation
D
Pre-filled document
Text files and databases for reused content components
3 Document content fulfilling
D
Complete document
Complete document
D
XSLT transformation documents
4 Metadata extraction
md
Process and document metadata

Fig 6. Example of integrated document production

The first phase considers collecting and combination of process metadata from process management system and databases covering actor information. Then, metadata about the document type and related XML schemas are extracted from document management system. Document containing process and document metadata is produced as an output.

The second phase considers generating a pre-filled document on the grounds of the process, document type and document metadata. According to the document type to be produced a relevant reusable text portion and the XSLT transformation documents for the document type may be picked up. Finally a pre-filled XML document instance, including metadata and text portions in the right structures, is created via XSLT transformation, and opened into the software end-users use.

In the third phase document producer modifies the pre-filled document and completes it with reusable primary content, such as law texts or phrases, through pointers into the text files or databases on the document template. Content producer's possibility to change metadata may be limited, because the metadata is inherited from systems, to prevent erroneous content due to possible typing mistakes.

In the fourth phase metadata is extracted from the document through XSLT transformation. As a result a complete document is ready to be checked-in back to the ECM system and metadata document is ready for importing into the systems or published on the Semantic Web of the organization.

6. Conclusion

This paper described the lessons learned from two e-Government research projects: RASKE2 and RAKE. The paper demonstrated that in e-Government a great deal of document content may be pre-filled automatically by integrating metadata and reused primary content instead of copying and pasting the content manually. Copying and pasting was found as error-prone and tedious task by the content producers. The requirements posed for the e-Government documents are also strict; the normative text and references to processes and actors must be correct.

The model for integrated document production consisting of process, document, actor and system metadata, as well as reusable primary content components on the domain was described. We demonstrated that the model is applicable by using the Semantic Web technologies for integrating different software systems or creating metadata for the Semantic Web of the organization. The integration provides means for ensuring content consistency across the documents and systems thus enforcing the integrity of the documents and supporting content management activities taken.

For research the study illustrates the requirements for content reuse based on the findings of the two research projects in different e-Government organizations. The interrelated relationship of business and content management processes [24] were described from the document producers perspective. The findings support the requirement for automatic metadata generation [25].

Potential avenues for further research have been considered. FCP has started a Proof-of-Concept project examining the integrated XML document production model in more detail. The Finnish Government and Parliament have planned to continue migration into XML document production.

7. Acknowledgments

The research would not have been possible without the active collaboration and support from the Finnish Parliament and the ministries as well as from the Finnish Centre for Pensions (FCP).

8. References

[1] Berners-Lee, T., Hendler, J. and Lassila, O. "The Semantic Web," *The Scientific American* (284:5), pp. 34-43.

[2] Glushko, R. and McGrath, T. *Document Engineering: Analyzing and Designing Documents for Business Informatics and Web Services*, MIT Press, Massachusetts, 2005.

[3] Mancini, J. "State of the ECM Industry," *AIIM E-DOC Magazine* July/August 2006.

[4] Murphy, L. "Digital Document Metadata in Organizations: Roles, Analytical Approaches, and Future Research Directions," in *Proceedings of the 31st Hawaii International Conference on System Sciences (HICSS)*, R. Sprague (ed.), IEEE Computer Society, Los Alamitos, 1998, pp. 267-276.

[5] Salminen, A. "Building Digital Government by XML," in *Proceedings of the 38th Hawaii International Conference on System Sciences (HICSS)*, R. Sprague (ed.), IEEEE Computer Society, Los Alamitos, 2005, p. 10.

[6] Lenk, K., Traunmuller, R., and Wimmer, M. "The Significance of Law and Knowledge for Electronic Government," in *Electronic Government - Design, Applications and Management*, A. Grönlund (ed.), Idea Group Publishing, Hershey, 2002, pp. 61-77.

[7] Bray, T., Paoli, J., Sperberg-McQueen, C., Maler, E., Yergeau, F., and Cowan, J. *Extensible Markup Language (XML) 1.1 (2nd ed.)*. W3C Recommendation, 2006.

[8] Herman, I., Swick, R., and Brickley, D. Resource Description Framework (RDF). Retrieved May 7, 2007, from: http://www.w3.org/RDF/

[9] *Moreq Specification - Model Requirements for the Management of Electronic Records.* European Union, 2001.

[10] Salminen, A., Nurmeksela, R., Lehtinen, A., Lyytikäinen, V., and Mustajärvi, O. "Content production strategies for e-Government", in *Encyclopedia of Digital Government*, A.-V. Anttiroiko and M. Mälkiä (eds.) IDEA Group Publishing, Hersey, 2006.

[11] Lyytikäinen, V. *Contextual and Structural Metadata in Enterprise Document Management.* PhD Thesis. University of Jyväskylä.

[12] DublinCore, *Dublin Core Metadata Element Set, Version 1,1*. Dublin Core Metadata Initiative, 1999.

[13] Susman, G. and Evered, R. "An assessment of the scientific merits of action research,"

Administrative Science Quarterly, (23:4), 1978, pp. 582-603.

[14] Nurmeksela, R., Virtanen, M., Lehtinen, A., Järvenpää, M., and Salminen, A. *Information management in the Finnish legislative work – Towards Semantic Web of legislative information resources.* The Parliamentary Office, 2006.

[15] Järvenpää, M., Virtanen, M., and Salminen, A. "Semantic portal for legislative information", in *Proceedings of the Fifth International EGOV Conference*. Springer Verlag, Wien, 2006.

[16] Honkaranta, A., Salminen, A. and Peltola, T. "Challenges in the Redesign of Content Management: A Case of FCP," *International Journal of Cases on Electronic Commerce* (I:1), 2005, pp. 53-69.

[17] Salminen, A. "Methodology for document analysis", in *Encyclopedia of Library and Information Science*, A. Kent (ed.), Marcel Dekker, Inc., New York, 2000, pp. 299-320.

[18] Levy, D. "Document reuse and document systems," *Electronic Publishing - Origination, Dissemination and Design* (6:4), 1993, pp. 339-348.

[19] Goldfarb, C. "SGML: The reason why and the first published hint," *Journal Of The American Society For Information Science* (48:7), 1997, pp. 656-661.

[20] Brun, C., Dymetman, M., Fanchon, E., Lhomme, S., and Pogodalla, S. "Semantically-based text authoring and the concurrent documentation of experimental protocols," *ACM Symposium on Document Engineering*, 2003, pp. 193-202.

[21] Murata, M. Lee, D., Mani, M. and Kawaguchi, K. "Taxonomy of XML schema languages using formal language theory," *ACM Transactions on Internet Technology*, (5:4), 2005, pp. 660-704.

[22] Clark, J., *XSL Transformations (XSLT) Version 1.0.* W3C Recommendation, 1999.

[23] Clark, J. and DeRose, S. *XML Path Language (XPath) Version 1.0.* W3C Recommendation, 1999.

[24] Wang, J. and Kumar, A. "A framework for document-driven workflow systems," *Lecture Notes In Computer Science*, 2005.

[25] Greenberg, J., Spurgin, K. and Crystal, A. "Functionalities for automatic metadata generation applications: a survey of metadata experts' opinions," *International Journal of Metadata, Semantics and Ontologies*, (1:1), 2006, pp. 3-20.

V

A LIFE CYCLE MODEL OF XML DOCUMENTS

by

Salminen, A., Nurmeksela, R., & Jauhiainen, E. 2014

Journal of the Association for Information Science and Technology 65 (12),
2564-2580, DOI: 10.1002/asi.23148

http://urn.fi/URN:NBN:fi:jyu-201412313597