Metadata Enhanced Content Management in Media Companies

Sami Jokela

Helsinki University of Technology
Department of Computer Science and Engineering
Software Business and Engineering Institute
P.O. Box 9600
FIN-02015 HUT
Finland

Dissertation for the degree of Doctor of Technology to be presented with due permission for public examination and debate in Auditorium T2 at Helsinki University of Technology, Espoo, Finland, on the 9th of November, 2001.

Jokela S. **Metadata Enhanced Content Management in Media Companies**. Acta Polytechnica Scandinavica, Mathematics and Computing Series No. 114, Espoo, 2001, 155 pp. Published by the Finnish Academies of Technology. ISBN 951-666-585-3, ISSN 1456-9418.

Keywords: Convergence, semantic metadata, ontologies, domain modeling, advanced content-based products.

Abstract

Media companies are facing new opportunities and challenges. Communications, computing, and content industries are converging into a single, horizontally connected content value chain, where changes are frequent and activities are highly interdependent. However, before convergence and digital content are taken seriously, media companies must understand what is expected from them, how their operations will be affected, and why they should be involved.

The production, distribution, and use of content rely heavily on computers and automation. This requires the content essence to be enhanced with explicit descriptions of semantics, or more specifically, semantic metadata. However, semantic metadata is useful only if its nature is understood clearly, and when its structure and usage are well defined. For this purpose, ontologies are needed to capture the essential characteristics of the content domain into a limited set of meaningful concepts.

The creation and management of ontologies and semantic metadata require skills and activities that do not necessarily exist in traditional print-based publishing or broadcasting. Companies developing ontologies must understand the essential characteristics of available content, user needs, and planned or existing use of content. Furthermore, they must be able to express this information explicitly in an ontology and then reflect changes in the environment back to that ontology.

Content production and distribution should be flexible and able to support the reuse of content. This thesis introduces two abstract models, a component model and a process model. Both models assist in the understanding and analysis of electronic publishing of content for multiple media products and on multiple media platforms.

When semantic metadata, ontologies, and improved publishing processes are available, new advanced content-based products, such as personalized information feeds, are possible. The SmartPush project, for which the author was the project manager and worked as a researcher, has shown that semantic metadata is useful in creating advanced content-based products, and that media companies are willing to alter their existing publishing processes. Media companies participating in the SmartPush project have acknowledged the impact of our work on their plans and operations. Their acknowledgement emphasizes the practical importance of semantic metadata, ontologies, improved electronic publishing process, and personalization research.

Prelude

I have been working full time on metadata and content related issues since summer 1997, but the interest in the field stems further back in time. In the early 1990's I, among a group of other students, was asked to develop a business simulator that would be used in one of the courses at the Helsinki University of Technology. During development the group tried to identify and define the most important elements and relations for manufacturing consumer products and marketing them to different customer segments. The development task taught the team the important connection between structural information and business operations.

My master's thesis work at the Helsinki University of Technology, entitled *Configurator Building Guidelines* exposed me to the practical challenges of knowledge modeling. Although I had touched on knowledge-related issues in the preceding years, the thesis was the starting point of a journey researching the question how to capture important aspects of a certain content domain into metadata, and then to use that metadata for multiple purposes. This work also acted as an early introduction to the world of knowledge modeling frameworks.

After graduating in 1994 I joined my current employer, Accenture. In 1997, I was asked to join the SmartPush project, which concentrated on content production and personalized information feeds. My field of research in the SmartPush project was media companies and their metadata activities. After the SmartPush project, I returned to Accenture and joined Accenture Technology Labs in Palo Alto, California, where the search for new knowledge continues.

The high-paced consulting world has clearly had an impact on this dissertation. I cannot ignore the importance of rooting the ideas and thoughts to the underlying business reality. I hope that you agree on the relevance of this point and find this work useful for your own purposes.

Acknowledgements

This kind of work could not be possible without the help of a group of incredible people. First of all, I would like to thank Professor Reijo “Shosta” Sulonen for giving me the opportunity to work with such a fascinating topic and for all his advice and support during the effort. I also want to express my warmest gratitude to my pre-examiners, Professor Michael Franklin at the University of California, Berkeley, with whom I enjoyed working during my time at UC Berkeley, and Professor Timo Käkölä at the University of Jyväskylä, for their constructive comments and suggestions. Your insights have improved this work significantly.

My former colleagues at the SmartPush project, Teppo Kurki, Eerika Savia, Andreas Anderson, and Virpi Huhtinen, provided a pleasant working environment and fertile ground for scientific work. Other people related to the SmartPush project have also contributed much to my thesis. Mikko Laine, JR Leskinen, Eero Tuomisto, Olli Pitkänen, and Pirkka Palomäki deserve special thanks for their support.

At Accenture Helsinki I want to thank Juho Malmberg, who believed in me and gave me the opportunity to try something that had not been done before. At Accenture Technology Labs, I want to thank Anatole Gershman and Luke Hughes for the great opportunity to participate in both the academic research and the hectic, business-oriented atmosphere of the Silicon Valley. Of my colleagues at Accenture Technology Labs in Palo Alto I would like to thank Anne Donker, Adam Knapp, Sean Smith, and especially Tony Miller, who each did an incredible job in helping me improve the manuscript.

I feel very privileged to have had the opportunity of spending the last two years with the intellects and researchers at the School of Information Management, SIMS, at UC Berkeley. I want to express my gratitude to Hal Varian and others at SIMS for the opportunity and the valuable insights gained during those years.

My co-authors, Janne Saarela and Marko Turpeinen, deserve my warmest thanks and praise for their support. Your guidance has helped me to avoid some pitfalls that I probably would have encountered during the thesis work.

I want to express my gratitude to my family, my sister Jutta and my mother Marja-Liisa, and to my late father Seppo, who unfortunately never had the opportunity to see the completion of this work. Without the stimulation and curiosity they nurtured in me as a small boy, this work could not have been possible. And finally, my warmest thanks to Reetta – without your support and patience none of this would have materialized.

It has been a great pleasure to work with and to know all of you.

Palo Alto, California, August 15, 2001.

Sami Jokela

Contents

Part I

Part II

Summary of publications

Publication 1.

Savia, E., Kurki, T., and Jokela, S. (1998) *Metadata Based Matching of Documents and User Profiles*, in Human and Artificial Information Processing, Proceedings of the 8th Finnish Artificial Intelligence Conference, Finnish Artificial Intelligence Society, August 1998, Jyväskylä, Finland, pp. 61-70.

Publication 2.

Kurki, T., Jokela, S., Turpeinen, M., and Sulonen, R. (1999) *Agents in Delivering Personalized Content Based on Semantic Metadata*, Intelligent Agents in Cyberspace -track, Papers from the 1999 AAAI Spring Symposium, Technical Report SS-99-03, AAAI Press, March 1999, Menlo Park, California, USA, pp. 84-93.

Publication 3.

Jokela, S. and Saarela, J. (1999) *A Reference Model for Flexible Content Development*, Proceedings of the Second International Conference on Telecommunications and Electronic Commerce (ICTEC), October 6-8, 1999, Nashville, Tennessee, USA, pp. 312-325.

Publication 4.

Jokela, S., Turpeinen, M., and Sulonen, R. (2000) *Ontology Development for Flexible Content*, Proceedings of the HICSS-33, IEEE Computer Society, January 4-7, 2000, Maui, Hawaii, USA, *Best Paper Award of the Internet and Digital Commerce track.*

Publication 5.

Jokela, S., Turpeinen, M., Kurki, T., Savia, E., and Sulonen, R. (2001) *The Role of Structured Content in a Personalized News Service*, Proceedings of the HICSS-34, IEEE Computer Society, January 3-6, 2001, Maui, Hawaii, USA.

List of abbreviations

AAAI	American Association for Artificial Intelligence
ACM	Association for Computing Machinery
BBC	British Broadcasting Corporation
BSE	Bovine Spongiform Encephalopathy
CPT	Content Provider Tool
CSLI	Center for the Study of Language and Information
CYC	An attempt to create an immense multi-contextual knowledge base and an inference engine for fundamental human knowledge
DAML	DARPA Agent Markup Language
DVD	Digital Video Disc (also known as Digital Versatile Disc)
EBU	European Broadcasting Union
HICSS	Hawaii International Conference on System Sciences
ICE	Information & Content Exchange
ICTEC	International Conference on Telecommunications and Electronic Commerce
IE	Information Extraction
IEEE	Institute of Electrical and Electronics Engineers
IF	Information Filtering
IPTC	International Press Telecommunications Council
IR	Information Retrieval
ISO	International Organization for Standardization
MACR	Moderately Abstract Conceptual Representation
MARC	Machine Readable Cataloging
MIT	Massachusetts Institute of Technology
MPEG	Moving Picture Experts Group
MPEG-7	MPEG content representation standard for multimedia information search, filtering, management, and processing
MPEG-21	MPEG multimedia framework
NewsML	News Markup Language, an XML-based standard for multimedia news creation, storage, and delivery
NITF	News Industry Text Format
PDA	Personal Digital Assistant
PRISM	Publishing Requirements for Industry Standard Metadata
RDF	Resource Description Framework
RDFS	Resource Description Framework Schema
RSS	RDF/Rich Site Summary
SGML	Standard Generalized Markup Language
SHOE	Simple HTML Ontology Extension
UDDI	Universal Description, Discovery, and Integration of business for the World Wide Web
VAP	Value-Added Publishing
W3C	World Wide Web Consortium
XML	Extensible Markup Language
XTM	XML Topic Maps

Part I

1 Introduction

Before introducing the actual essence of my work, I'll start with a few words of how this work is organized and to whom it is written.

To help find and explain the key terms used in the thesis, I have collected them in Appendix A, where they are introduced in alphabetical order. A more detailed discussion on the key terms is available elsewhere in the work and in the publications.

I have also used a special formatting to help identify the definition and usage of key terms. Term definitions in this dissertation are marked with ***this style***. When a term is marked with this style, the term or its plural form is defined in the Glossary of Terms.

This work spans a number of research fields in computer science, digital economy, and social sciences. Key themes in the work are convergence, ontologies, semantic metadata, creation, reuse, and distribution of content, publishing processes, and new applications based on content. From research perspective, the most important utility of my thesis is its integrative approach to the discussed issues and the rooting of these issues to practice via constructive research. Even though this thesis is first and foremost an academic work, I hope and believe that this thesis is useful also for the people involved in the creation and management of content in their daily activities.

1.1 Motivation

Media companies are facing new opportunities and challenges. Converging industries, the plethora of new media platforms, and increased availability of digital content are major issues that will impact media companies and content management.

Before media companies are willing to change, they must understand what is expected from them, how their operations will be affected, and why they should be involved. A major part of this motivation is related to understanding how different content-related activities create value. If different steps in the production and delivery of content are linked to value creation, the existence of these activities as part of the emerging content value chain are easier to justify.

One of the main factors affecting production, delivery, and usage of content is convergence: Communications, computing, and content industries are merging into a single, interconnected industry [Collins et al., 1997].

Convergence is likely to demand a high degree of co-operation between participants in the content value chain. It also requires standardized methods to manage and describe content essence so that different participants in the content value chain are able to utilize the available content. In the content value chain, content may originate from multiple sources or it can be augmented with other content, requiring additional attention on integration and interaction.

Media products are impacted by new opportunities and challenges as well. In newspaper publishing, one of the determining factors in the past was the physical nature of the delivery media, paper. This limitation is not necessarily valid with electronic publishing. Media companies today are capable of creating more content, but they need better methods for content management. Media companies must be able to operate 24 hours a day, seven days a week, and reuse the same content on different media products and on multiple media platforms. In addition, instead of spending months or even years developing new print-based products, media companies are required to produce new advanced content-based products based on their existing content in a matter of weeks. For example, one of the media companies I worked with was required to produce a new version of its financial newsfeed almost on a weekly basis to meet the complex and changing requirements of its customers.

One result of new operating methods is the increased reliance on computers and software. As many of the activities related to content are performed by computers that do not have the same level of intelligence as humans, content essence must be augmented with explicit descriptions of its characteristics, its metadata. Media companies need flexible methods to manage the production and delivery of digital content, and semantic metadata describing the essential semantics of content essence is a key enabler in making this goal a reality. However, semantic metadata is useful only if its nature is understood clearly, and its structure and usage are well defined. For this purpose, ontology – conceptual models that map the content domain into a limited set of meaningful concepts – is needed.

Creation and management of ontologies require skills and activities that do not necessarily exist in traditional print-based publishing or broadcasting. Companies developing ontologies must understand the essential characteristics of available content, user needs, and planned or existing advanced content-based products. Furthermore, they must be able to express this information explicitly in the ontologies. It is also important that they recognize changes in the environment and reflect these changes back to the ontologies.

The availability of structured content enables new ways to produce, reuse, and deliver content on multiple media products and media platforms beyond merely presenting the content essence to the end-user. Structured content can be used to create interactive services, personalize presentation of content essence based on a user's likings, or augment the content essence with other relevant information. Other possible advanced content-based products include services, in which computers perform actions based on content, such as automatic notifications of urgent information.

1.2 Scope of the study

This thesis builds on the foundations of convergence and discusses how media companies in the content value chain are going to be impacted by new and changing requirements in the operating environment. Although individuals and communities have a major role in creating content, especially for the Internet, this work concentrates on content originating from professional organizations.

The main emphasis of this thesis is on creation, management, and use of content for new kinds of services and applications – advanced content-based products. In addition to discussing the nature of content in detail, this work analyzes some relevant aspects of the production of and product development for content. The examples at the end demonstrate some of the possibilities and challenges that are likely to emerge

when companies participate in the content value chain and begin producing content for advanced content-based products.

This work spans a number of research fields in computer science, digital economy, and social sciences. The main focus areas are convergence, domain modeling, semantic metadata management, and electronic publishing. Some of the relevant fields are discussed only briefly, such as philosophy, linguistics, communications, media studies, knowledge management, and artificial intelligence. Certain fields, such as intellectual rights management or organizational structures, have been excluded altogether, although they do have a high relevance to my work.

The following Figure 1 introduces the main topics of the thesis. The numbers in the figure illustrate the main chapters in Part I where these topics are discussed. Part II consists of the publications that present in detail the research foundation of the thesis.

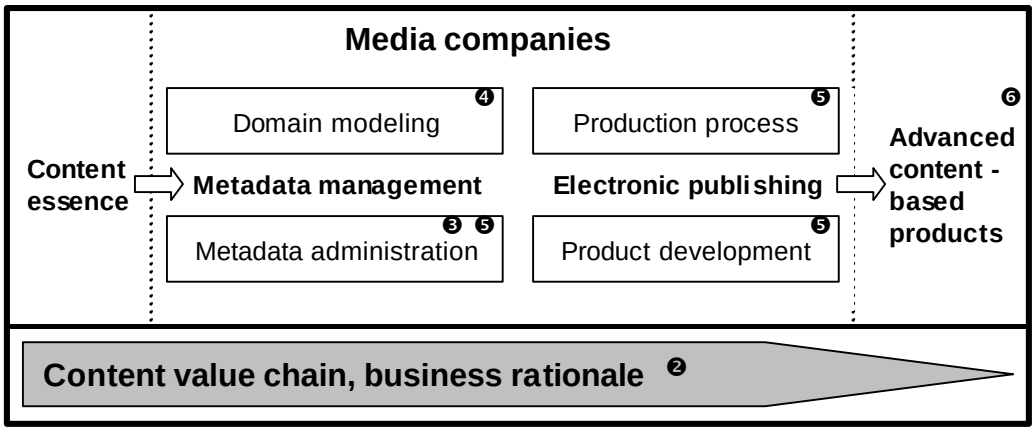

Figure 1. Main topics of the thesis.

1.3 Goals and research methods

The goals of this work are two-fold. The work aims to be a proof of concept and motivator that semantic metadata, electronic publishing, and advanced content-based products have economic and technological justification, and that they can be an integral part of the future of media companies. At a more detailed level, the thesis presents and discusses some of the technology and process-related questions that are likely to appear with the introduction of metadata management and domain modeling into the production process of content. The work also aims to demonstrate some of the characteristics of the potential advanced content-based products that are likely to emerge when media companies start producing content augmented with semantic metadata.

My work does not follow the traditional dissertation format of concentrating on a certain, well-defined problem in isolation from its broader context. The results presented here span a broad range of topics raising the focal point from isolated details to the system and its interrelations. As research integrating different business and technology aspects of metadata enhanced content management is still in its infancy, I consider this and similar works as early trailblazers in the content management for more detailed approaches to follow.

Even if some of the discussed issues and ideas may not be considered completely new, the integration, viewpoint, analysis, and conclusions help make this work unique.

I attended the HICSS-33 conference[1] in January 2000, where the distinguished lecturer, Nobel Prize winner Dr. Arno Penzias, discussed the similarities and differences between academic research a century ago and today. According to the speaker, the 20th century has taught the academic community how to solve problems in isolation, but the time has come to build on the developed foundation and to continue research at systemic level. These words supported my personal thoughts and justified the value of this work.

Most of the empirical part of my dissertation was conducted during the SmartPush project in close collaboration with a number of industrial partners, of which the most influential regarding this work were media companies *Kauppalehti Online*, part of a Finnish media conglomerate, *Alma Media Corporation*, and *WSOY Uudet Mediat*, part of a Finnish media conglomerate, *SanomaWSOY Corporation*. The SmartPush project, which is discussed in more detail later in the thesis and in the publications, has provided me with an excellent insight to the internal operations of large media conglomerates and what kind of issues they are facing in their daily activities.

When I began my research, the SmartPush project had a high-level goal, but its specific focus areas and methods were still relatively unclear. We also initially assumed the existence of suitable standards for semantic metadata as well as that content conforming this standard was available. Unfortunately such standards or content did not exist, so I had to rely much more on constructive research and experiments with a series of prototypes.

My personal goal was to concentrate on the big picture and raise the focal point from individual details to the system and interoperation between different partners in the production and delivery of content. To achieve this, I worked in close co-operation with our industrial partners and refined my goals and research methods iteratively during the project. As a by-product of this approach, I gained quite a detailed understanding of the operations taking place at our project partners, but had fewer possibilities to verify the generic applicability of my findings due to the lack of larger sets of statistically analyzable data. Additionally, by focusing my research at a higher level of abstraction, I did not face some of the hard and unsolved problems related to the everyday details that unarguably need to be solved before presented results can be fully used in practice.

This work is not based on quantitative research methods that typically require a substantial amount of data to be statistically significant. I therefore decided to rely mainly on qualitative data sources and methods, such as discussions, interviews, existing literature, cases, prototype building, analogies, and observations as the basis for understanding and explaining issues related to my work. The used data sources are described in the footnotes and references at the end of the work.

Qualitative research methods (see e.g. [Myers, 2000]) seem to fit best to this work due to the systemic nature and many interrelations between the numerous related fields of research. Additionally, because the research integrating the discussed fields is still in its infancy, quantitative data and suitable methods linking

[1] www.hicss.hawaii.edu/hicss_33/apahome3.htm

the relevant issues is scarce. I am fully aware of the possible weaknesses associated with the used data sources and research methods, such as the applicability and generalization of derived results, so future work should concentrate on generalization and detailed validation of the presented results.

1.4 Key contributions

1.4.1 Author's major contributions

The following list summarizes the author's major contributions during the course of the dissertation. In addition to the described contributions, this work led to several minor contributions described in different parts of the thesis and the publications. The major contributions are:

- **Development of a revised model for describing convergence in the content industry**. The revised model for convergence is based on existing convergence theories, but emphasizes the value of content. It also incorporates customers and discusses how their feedback affects certain parts of the resulting content value chain. This model is described mainly in chapter 2.
- **Identification of desired characteristics for semantic metadata and conceptual models**. The author, together with other researchers in the SmartPush project, identified and developed iteratively desirable characteristics for semantic metadata and conceptual models. These characteristics are discussed in chapter 3 as well as in publications **1**, **2**, **4**, and **5**.
- **Identification of key elements for ontology development**. The author developed principles for ontology development in order to describe the key factors affecting ontologies and to assist in understanding what is needed when ontologies are developed. These principles also discuss how changes might impact ontologies over time. The principles for ontology development are discussed mostly in chapter 4 as well as in publication **4**.
- **Development of a component model for content**. The author co-developed a four-layer framework for media companies to understand and better manage different elements of electronic publishing. This component model is introduced in chapter 5 and in publication **3**.
- **Development of a process model for electronic publishing**. The author was the main developer of the process model for electronic publishing. The process model assists in understanding different activities in the electronic publishing process and allows media companies to reuse content in multiple media products and across different media platforms. The process model is discussed in chapter 5 and in publication **3**.
- **Development of a metadata creation tool**. The author was responsible for the development of a novel metadata creation tool and wrote most of its code. This tool was used in producing semantic metadata for the SmartPush project and helped in developing the principles related to the production of semantic metadata. The metadata creation tool is discussed in chapter 6 and in publications **2**, **4**, and **5**.

1.4.2 Key findings

This chapter summarizes the key findings and lessons learned resulting from my thesis work.

Media companies must prepare themselves for convergence and the flow of digital content in the content value chain.

Communications, computing, and content industries are converging into a single, horizontally connected content value chain, where activities are highly interdependent. However, before convergence and digital content will become a reality, media companies must have suitable tools, methods, and business models to justify the effort.

New requirements call for changes in the production and distribution of digital content. Content providers must be able to cut down the development time for media products, reuse content on multiple media products and media platforms, and be more flexible and proactive towards other participants in the content value chain and end users.

The value of content is a difficult but important factor in determining the acceptable cost of producing content, as well as in assessing the resulting revenue and profit potential. Some aspects and methods, such as predictive utility, customization, and avoiding repetitive information, are useful in estimating the value of content.

Structured content consisting of content essence, semantic metadata, and shared ontologies is essential for producing and delivering content in the content value chain.

Computers play an increasingly important role in the content value chain. Computerization and advanced content-based products require structured content consisting of content essence and metadata, especially semantic metadata. However, before semantic metadata can be produced and used, an ontology is needed. Ontologies capture relevant characteristics of the content domain into dimensions that are as independent as possible.

The SmartPush project and co-operation with media companies has resulted in a list and analysis of desired characteristics for metadata. These characteristics include expressiveness, extensibility, neutrality, immutability, compactness, high value, uniformity, openness, versioning, and unique identification.

Some standards for metadata exist to describe the semantic characteristics of content essence, but in general these standards are yet quite limited and insufficient to cover the semantic aspects of the content essence.

Media companies must alter their processes, methods, and tools to enable domain modeling, metadata-enhanced content management, content reusability, and the creation of advanced content-based products.

Domain modeling is highly dependent on three elements: information feeds, user needs, and intended use. The dynamic nature of the content domain must be taken into account when the ontology is developed and used. If any of the elements for domain modeling change, these changes must be reflected in the ontology.

Automation in domain modeling and in the creation of semantic metadata is important, but to ensure the highest quality, human experts should still be in control of the creation of ontologies and semantic metadata. Only fully routine creation of semantic metadata can be performed without human intervention.

Two abstract models, a component model and a process model, assist in the understanding and analysis of electronic publishing of multiple media products on multiple media platforms.

Advanced content-based products allow the content to be used in ways beyond mere presentation. When companies develop new advanced content-based products, the development should consider the characteristics of the outcome and available content, including quality, relevance, and clarity of content, the amount of content provided, as well as the value of content. The content should also originate from credible sources, be suitable for further use, and be delivered on a timely basis.

The SmartPush project demonstrated a valid and viable approach in which media companies augment content essence with semantic metadata and use that content in producing a personalized information service. Media companies participating in the SmartPush project have acknowledged its impact in their plans emphasizing the practical importance of semantic metadata, ontology, electronic publishing process and personalization of content.

1.5 Future research

In addition to issues outside the scope of this work, such as organizational and legal questions, I have identified and touched on a number of important challenges that are limitations of my work and remain open for future research. These challenges include advanced methods for domain modeling, versioning and mapping ontologies, measuring the quality of the ontologies and semantic metadata, tool support for automated metadata creation, and proper tools for management and visualization of ontologies. In order to avoid the potential weaknesses of the used research methods such as constructive research, cases, and experimentation, future research should also aim at validating and generalizing the presented results. More discussion on future research can be found in the conclusions of this thesis.

2 Business rationale for the metadata enhanced content management

The journey into metadata enhanced content management begins with an introduction to the convergence theory and discussion of its impact on the media companies and the content value chain. After that, I introduce some of the business model alternatives for the media companies and other participants in the content value chain. This chapter ends with a discussion on the value of content and how the value is connected to advanced content-based products and metadata enhanced content management.

2.1 Background

The production and delivery of content essence has changed dramatically in the past years. Digital content and new media platforms have diminished the physicality of the content value chain. In print publishing, the physical nature of paper-based deliverables requires high up-front investments in printing and distribution facilities and makes printing and delivery mostly unidirectional activities emphasizing the existence of deadlines. With digital content investments in physical structures are typically much lower and the updating and distribution of content is easier decreasing the need to have strict deadlines. Digital content and electronic delivery have also shortened product development time. Instead of spending multiple years in planning and building the infrastructure for a new paper-based product such as a newspaper, media companies must today develop new advanced content-based products in a matter of weeks, as our partners in the SmartPush project have witnessed. Customers demand the same content essence in different formats and on multiple media platforms. Content is used for many other purposes than viewing, which sets additional requirements on the reusability of the content.

Electronic publishing is highly dependent on two aspects: capability and motivation. The capability to participate in the content value chain depends on the availability of methods, resources, tools, and processes that allow the content to be created, integrated, packaged into a variety of products, and distributed via multiple media platforms. However, if the participants do not have a business model explaining why they are involved in the content value chain and what kind of expectations and economic returns are likely to appear, their participation is difficult to justify from a business perspective.

2.2 Convergence and media companies

The starting point for this work is the convergence theory, according to which communications, computing, and content industries merge into a single, interconnected industry. The following chapters introduce the convergence theory and the content value chain as well as discuss some business model alternatives and business issues impacting media companies.

2.2.1 Theories of convergence

The idea of converging industries is not novel. One of the early industry pioneers was NEC Corporation in Japan. In 1977 the NEC Corporation began promoting a vision in which computer and communications

industries were destined to converge [Yoffie, 1997]. In the 1980's Nicholas Negroponte, among other early visionaries, incorporated content into the convergence theory and founded the MIT Media Laboratory to research different issues related to content and convergence.

The particular version of the convergence theory that has been a major motivator for my work was an outcome of a colloquium *Colliding Worlds; The Convergence of Computers, Telecommunications, and Consumer Electronics* organized at Harvard in 1994 [Yoffie, 1997]. The goal of the event, which had roughly 80 participants from both industry and academia, was to discuss the findings of multiple convergence-related research projects conducted at Harvard and develop further a vision of networking in a new economy. This vision was captured into a framework for convergence, according to which the consumer multimedia industry will transform from three discrete vertical businesses including personal computing, telephone, and television, into five converging horizontal industry segments: content, packaging, transmission, manipulation, and terminals. This framework, which is the starting point for modeling convergence in the content industry, is illustrated in Figure 2.

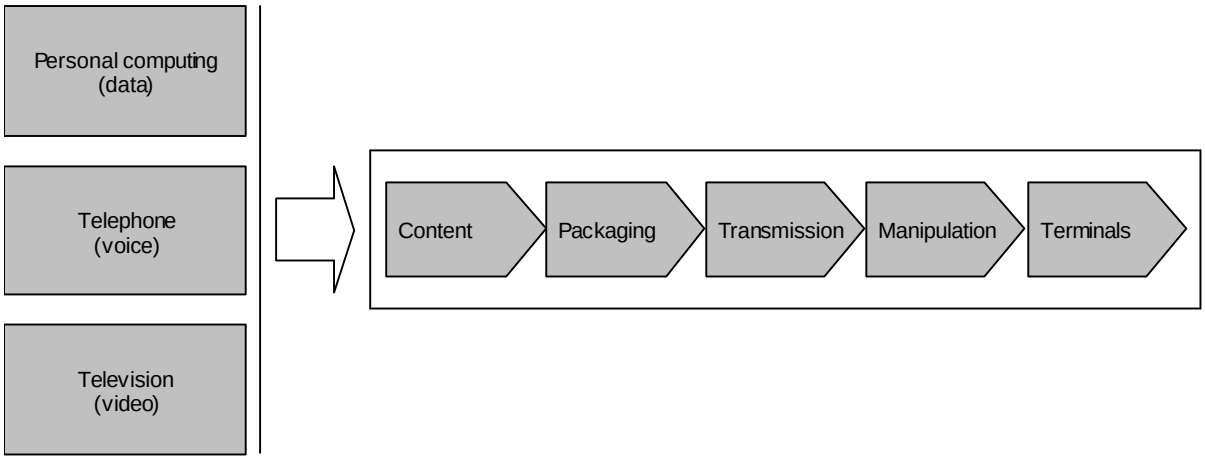

Figure 2. The impact of convergence to the consumer multimedia industry (adapted from [Collins et al., 1997]).

One of the obvious challenges in the presented convergence theory was the unidirectional flow of content essence through different industry segments. Although the role of consumers was mentioned briefly in the theory, consumer-generated content and the role of consumers to create, participate, choose, initiate, and control the flow was mostly neglected. As one can see from the horizontal industry segmentation in the previous figure, the coverage of the original convergence theory was far too limited. Computing today contains much more than just personal computing, communications focuses on data delivered over multiple platforms, and content consists of much more than plain video signal for television.

Since its introduction, the idea of converging industries has received a lot of publicity and spawned a number of variations in regards to which industries are to merge and what the resulting horizontal segments will be. For example, according to [Rappaport, 1997]:

> *"The technological convergence of computing, media, telecommunications, and consumer electronics and the rise of the information highway are irreversible trends bundled in one phrase: the digital convergence."*

[Dowling et al., 1998] discuss different aspects related to the merging of online services and television and define convergence as:

> *"Convergence describes a process change in industry structures that combines markets through technological and economic dimensions to meet merging consumer needs. It occurs either through competitive substitution or through the complementary merging of products or services or both at once."*

As the definition of convergence by Dowling et al. demonstrates, convergence theory has implications in the role of the traditional ***media industry*** concentrating on print, radio, and television. Changed requirements to the media industry are discussed e.g. in [Bruck, 1997], where the term *content industry* has wider coverage than the media industry that provides the content essence. I define the ***content industry*** as those companies that take part in content production, refinement, and distribution, beginning at the creation of content and ending with the companies producing equipment and software for end-users to consume the content.

Some variations of the convergence theory concentrate more on communications and computing, some emphasize more the role of content. I do not treat these aspects separately, but simply see the convergence taking place over communications, content, and computing, thereby forming a horizontally-connected content industry in which content is processed and managed with computers and delivered via multiple different communication methods and media platforms to the users. If a company wants to operate in any segment of the content industry, it must take convergence into account and understand the roles and relations of other participants in that industry. Furthermore, companies must define and use common standards for describing content essence before many of the advanced content-based products and reuse of content essence will become reality.

2.2.2 Content value chain

In this work, the content value chain is a fusion of different aspects of convergence and companies participating in the content industry. The content value chain used in this thesis consists of the different participants and their activities in which the content is processed and its value increased starting from the creation of content and ending when the content, i.e. either metadata or content essence, is used for the last time. This converged content value chain is illustrated in Figure 3. It shows how content flows through different interconnected process steps and participants of the content value chain and how customers impact the different steps in the content value chain through feedback and communication.

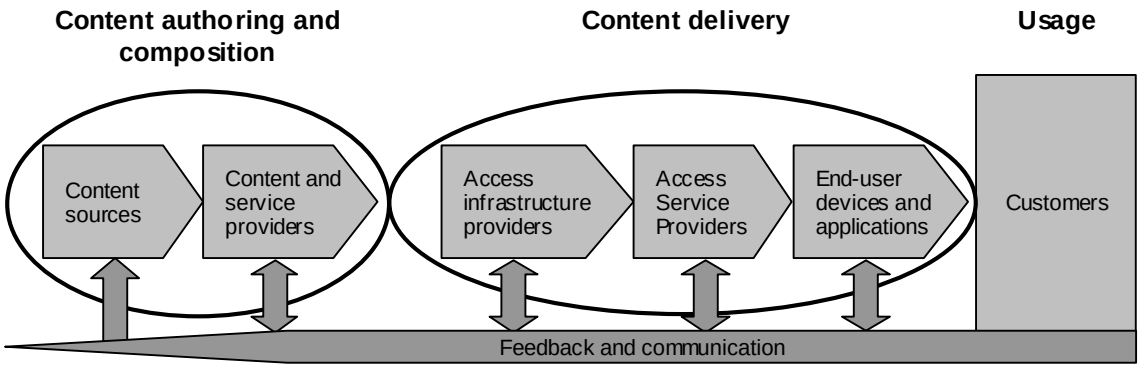

Figure 3. Content flow through different participants in the content value chain.

Although individuals and communities are an important source for content, feedback, and control in the content value chain, the role of customers is not discussed in detail in this work. For readers interested in further information on the role of customers and communities in the content value chain, [Turpeinen, 2000] contains a detailed analysis on individuals and communities in regards to news content customization.

Different segments in the content value chain must produce value; otherwise it is difficult to justify the existence of those segments. Figure 4 presents some examples of how different segments of the content value chain can increase the value of content.

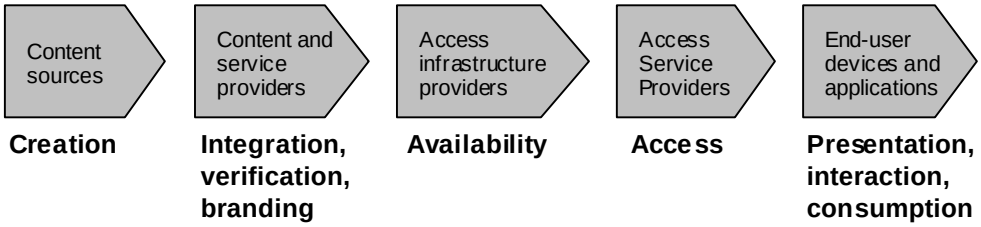

Figure 4. Examples of value creation in the content value chain.

Although there will always be a need for new professionally created content, it remains unclear, whether dominating players will develop to control the content value chain. If this happens, a further question is, will there be only one dominant company or is the content value chain going to be divided into horizontal or vertical segments that have their own dominant companies? Furthermore, if one company dominates, from which segment will this company come? Even though access service providers and end-user terminal manufacturers have been quite successful in dominating and profiting in the content value chain, especially in regards to wireless communications, it is questionable, whether they will maintain their commanding position in the longer run, once other companies outside of the lucrative access service segment decide to enter the competition and offer their own products bundled with their connectivity solution alternatives. For example, in the near future, power companies are likely to combine electricity distribution with Internet access services using their existing infrastructures.

In the past, content providers, service providers, and access providers have been the most active participants in the content value chain, with companies like AOL[2] and Yahoo[3] acquiring players in other segments. However, the stock market downturn of 2000-2001 has decreased the market capitalization of many technology companies, making them less likely to continue acquiring smaller players and in some cases has turned them into potential targets for takeover activities.

Independent of stock markets and acquisitions, companies in the content value chain seem to be increasingly interested in having exclusive access to content. This observation is supported also by [Kauffman and Riggins, 1998].

2.2.3 Convergence based business models

The characteristics of the content value chain are strongly linked to the business models companies pursue in the content industry. ***Business model*** describes a company's motivation: why company is involved in the business, how it is pursuing its goals, and what kind of returns it is likely to achieve. Business model is often a good indicator of the kind of strategies the company is likely to execute.

[O'Reilly, 1996] considers traditional print publishing to be a suitable example for business models related to digital content due to a number of similarities and motivators, including multiple alternatives of revenue sources, open platforms, a number of successful players, and universal non-exclusive access to content.

Despite the similarities, media companies focused on print publishing have been generally slow in adapting their products to electronic publishing. [Berghel, 1999] has identified two main reasons for this: The technology side lacks secure and trusted access and payment solutions and the business side is missing a sound business model. Based on my experience, these reasons only partially explain the slowness. Intellectual property questions, lack of proper methods, tools, and knowledgeable personnel needed for content management, as well as the need to support, not cannibalize, the existing core business, have all been major contributors to the slow adaptation.

However, in the past few years, significant progress has been made. Business models adopted from other industries, more integrated solutions to content management, improved technological methods for secured transactions, as well as the increased demand for content on a variety of media platforms such as wireless devices – all of these have contributed to the reassessment of electronic publishing and digital content as critical to the content industry.

Business models can be defined at different levels of abstraction. This work does not discuss business models for individual products or services, such as various subscription, pay-per-use, licensing, revenue sharing, and advertisement-based business models. I am more interested in the role of media companies in the content value chain as a whole, where some of the applicable business models for the production and packaging of content include the transaction-based cybermediary and manufacturer business models [Smith

[2] www.aol.com

[3] www.yahoo.com

and Jutla, 1999]. Although many media companies use a combination of these two business models, they indicate two possible strategic directions that media companies may pursue.

In the ***cybermediary business model***, a company operates as an intermediary between the source and the customer. When this business model is applied to the content value chain, the media company acts as an intermediary between sources of content essence and the customer, adding value to the process by allowing comparisons, repackaging content, or by forming communities. The cybermediary business model argues for the strengths of outsourcing the actual production to external participants. Many current portal and syndication companies follow the cybermediary business model by refraining from the actual creation of content, acting instead as an integrator of different sources of content.

The ***manufacturer business model*** is based on the assumption that manufacturing, marketing and distribution are manufacturer's core business activities. This business model is applicable to both advanced content-based products and traditional products such as newspapers and television programs. Examples of organizations using this business model are large media companies with strong online presence such as AOL Time Warner[4] and Disney[5]. The manufacturer business model requires good customer service processes, easily configurable products and skillful marketing. Media companies with a strong presence in traditional media platforms such as print or television can gain advantage from their existing business, because they can use their established media platforms to drive traffic to their alternative media platforms or media products. Both European and U.S. media conglomerates have lately used this cross-media strategy extensively by advertising their online products through their print media and TV-broadcasts [Sacharow et al., 1999].

One of the main consequences of the convergence is that companies in the content value chain are moving away from reactive *make and sell* business models towards a more customer focused and proactive *sense and respond* operating mode [Bradley and Nolan, 1998]. This requires the ability to co-operate between different segments in the content value chain, to specialize and produce complex products in much shorter timeframes than what was possible before, and to be more flexible to new and quickly changing conditions in supply and demand.

Not all companies in the content industry are willing to co-operate and share revenue between multiple participants specializing in their own segments of the content value chain. [O'Reilly, 1996] sees two trends in maturing online publishing: ***differentiation*** and ***consolidation***. Some players differentiate themselves from competitors by becoming experts in a certain niche, whereas other companies concentrate on gaining total control of a certain segment of the content value chain.

Media companies have faced consolidation and horizontal acquisitions, where large companies especially in the content packaging and access service business have bought or been interested in other participants of the content value chain (see e.g. [Mack, 2000], [Schenker, 2000]). This approach would ensure the company exclusive access to content and distribution, but would require substantial size and power to achieve and

[4] www.aol.com

[5] www.disney.com

maintain a commanding position. A contradictory development is that the distribution of digital content is becoming more and more syndicated, as the emergence of new companies and products such as *iSyndicate*[6], *NewsEdge*[7], or *Moreover*[8] seem to suggest.

I have not conducted a detailed analysis on the optimal role and position that media companies should pursue in the content value chain. In most of the cases that I have encountered, media companies perform some parts of the creation of content in-house, combining their content with material from other sources and media companies, after which the media company typically packages and brands the resulting content. With certain media products, such as online newspapers, media companies often host the online service themselves, whereas with more complex media platforms, the media companies may outsource the distribution to a third party. As an example, media companies may use outsourcing with some mobile phone-based services, where the telecommunications service providers manage the delivery, billing and interaction with the customer. In some cases, media companies syndicate and co-brand their content with other media companies, and the delivery of content is simply a bulk file transfer initiated by either the media companies or the customers. In this way, the media companies are able to share their marketing and promotion costs with other players and receive revenues from sources that might not be accessible otherwise. Media companies will quite likely continue using their brand to demonstrate the quality of their content. It seems therefore natural that media companies own, or are at least in control and responsible for, both the content essence and its augmentation with metadata.

No matter which business model is used, new advanced publishing methods and lowered production costs do not necessarily guarantee success. The publications avoided in hardcopy will be avoided in electronic form as well [Berghel, 1999]. Additionally, if media companies do not understand and master the combination of technologies, business models, and content, the results might not be commercially viable. Problems with understanding and integrating content, technology, and business models have recently led to failures and problems in many content-based startup companies [Wilson, 2000].

2.3 Value of information

Content can be valuable in many different ways. For example, content may have directly measurable economic value as information, but it might also have social, entertainment, artistic, or aesthetic values that are harder to measure. Although all of these value categories are important, this work concentrates on the discussion of such value of information that can be estimated and converted into monetary measures. A typical content domain, where information has directly measurable economic value, is *stock market*.

The idea of using monetary measures in information management is not new. For example, economic filtering was discussed already in the 1980's as a method to compare the costs and value of information to the user [Denning, 1982; Malone et al., 1987].

---

[6] www.isyndicate.com

[7] www.newsedge.com

[8] www.moreover.com

Media companies must understand how content creates value and how content essence, metadata, and their management generate costs. Automation may reduce the costs of creating metadata, but it may lower the quality of content, especially in complex cases where creation and processing of semantic metadata would require human intelligence. If the quality of the content essence or metadata is low, the value of content is diminished and in extreme cases may lead to very expensive errors. For example, if advanced content-based products are dependent on semantic metadata and they are used as a basis for investments, false or missing information may have extremely expensive consequences.

Although the value of content is a key driver in the content value chain, measuring the value of content is far from being understood. According to [Teece, 1998]:

> *"The information/knowledge/competences dimensions of inputs (especially intangibles) used to create products remain almost completely unexplored in economics and strategy."*

Value can occasionally be directly linked to transactions producing measurable results in fields like finance, but in many cases the value is indirect and/or difficult to measure. [Fahey and Prusak, 1998] noted the same problem in regards to information:

> *"…the value of data and information is often anything but obvious. Sometimes it is only after considerable discussion and dialogue that the decision relevance and usefulness of data and information becomes evident."*

[Berghel, 1999] sees content essence as utilitarian, where the value of content depends on the ability of customers and applications to read, understand, and use it. The utilitarian approach can be interpreted such that if content essence cannot be found or is not suitable for consumption, it is useless. Additionally, from utilitarian viewpoint content has value only if it contains new information or affects our view of the environment. Utilitarian approach is thus applicable mostly for informative content such as news and does not explain other reasons, such as mere ownership, as basis to acquire content.

Information has an interesting threshold characteristic [O'Reilly, 1996]. Up to a certain point, people and organizations demand more information and choice, but after a certain threshold the situation flips. Users become overwhelmed when information requires too much attention, after which less information is better. Information threshold has been one of the main reasons for building filtering applications for content such as the SmartPush project, which is introduced later in the thesis.

Context linking, communities, and interaction increase the value of information, which has led many media companies to augment their content essence with other relevant content. [Berghel, 1999] sees the payoff in electronic publishing to be the deployment of new technologies for the integration of digital documents into the network fabric of associated ideas, texts, times, and people. Publishers must supply the customer with more than just the pure document by connecting the content essence to its context. Motivated by these ideas, Berghel defines five components that assist building value in electronic publications. The five components are: enhanced content essence, metadata, feedback allowing learning of user preferences, interactivity, and support in the form of assisting technologies and tools.

Value of information cannot be measured without estimating the involved costs related to false and missing information. [Konstan et al., 1997] used predictive utility and comparison of costs and benefits to estimate how much value Usenet news articles would create to the user. ***Predictive utility*** is the difference between potential risk and benefit. It estimates how valuable a prediction on an item's utility is before the user decides to invest time and money on that item. A content domain that has high predictive utility is one where users pay a great deal of attention to the predictions. In content domains with low predictive utility, the predictions have little impact on user decisions. A factor related to predictive utility is the total number of desirable and undesirable items. If most of the items are valuable, the potential for advanced content-based products, such as filtering to predict good choices and to increase value, is only modest. If the source contains a lot of noise – the number of desired items is low compared to the total number of items – predictive utility is important. Figure 5 illustrates how the importance of predictive utility varies over different content domains. It shows what kind of benefits and costs typically emerge due to correct and false predictions of the value of information in exemplary content domains.

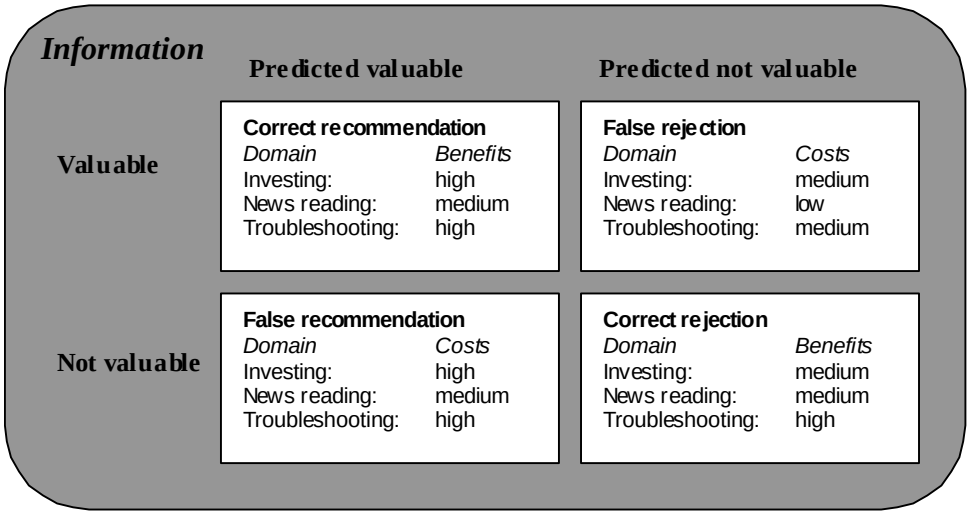

Figure 5. An example of using predictive utility to estimate information value in different content domains for a generic user (adapted from [Konstan et al., 1997]).

In practice, predictive utility and domain characteristics are a good starting point, but alone they are insufficient to estimate the value of information. If we want to predict the value of information, we must understand the values and risks of individual users and their activities in regards to content.

The value of ***repetitive information***, where the same information is offered multiple times to the same customer, can be argued and should be taken into account when media companies are developing and producing advanced content-based products. Although some advanced content-based products such as search engines typically aim at finding similar content essence, multiple results containing the same facts produce little additional value, at least in regards to the novelty of information. [Shapiro and Varian, 1999] claimed that the revenue of digital products tends to approach its variable production costs and can be almost zero, independent of the development and other pre-production costs. They suggested that, in the

long run, versioning is the only reasonable way to sell information and that content providers could avoid the value depreciation by customizing digital products into a unique product for each delivery.

[Varian, 1999] presented an interesting example to illustrate the value of repetitive information. In his example, a person is asked to pick one of two envelopes, one containing 100 dollars and the other being empty. If the person is risk-neutral and knows nothing that could help to determine the right envelope, the person would be willing to invest a maximum of half of the money to know which envelope contains the money. Similarly, if the person is offered the same information a second time, there would be no reason to pay anything for the information unless it changes the odds to pick the right envelope.

[Skyrme, 1999] discusses different characteristics that raise the value of information in regards to knowledge management. The characteristics include timeliness, accessibility, quality, utility, customization, contextualization, linking, flexibility, reusability, marketing, and meta-knowledge. Content and content management, including metadata management, address and improve most of these qualities. For example, semantic metadata has a major role in building reusable content for customized and contextualized advanced content-based products. The role of semantic metadata and reusable content in the electronic publishing is discussed in more detail in the publications, especially in publication 3.

In addition to the generic principles on how the value of information is determined, each content domain has its own characteristics that affect the value of content. [Turpeinen, 2000] discusses the methods used by news organizations to select stories to be published and introduces an analysis of newsworthiness based on a news value categorization by [Hartley, 1982]. According to Turpeinen, the most important characteristics affecting newsworthiness include event frequency, scale of importance, clarity of an event, cultural proximity or relevance, consonance, unexpectedness, continuity, balance of different events, reference to persons and social proximity, and the tendency to prefer bad news to good. These characteristics can be used as a starting point for identifying and evaluating valuable characteristics of other kinds of content as well.

2.4 Summary of findings

The communications, computing, and content industries are converging into a single, horizontally connected content value chain where participants are highly interdependent and integrated. In the new marketplace boundaries and roles in the content value chain will be dynamic and dependent on the conditions of the particular market segment. In some cases the media companies will opt to cover the whole content value chain from the creation to distribution of content. In others, they may choose to co-brand the content or to act as a bulk source of content. Convergence and digital content will provide new opportunities and challenges, but before media companies can participate in the content value chain, they need suitable tools, methods, and a business model to justify the effort.

Content providers must be able to cut down the development time and develop new advanced content-based products. They must become more flexible and proactive towards new and quickly changing conditions caused by supply and demand. The media companies must develop new methods for content management that will allow reuse of content and participation in the content value chain.

A likely role for media companies in the content value chain is to act as a provider, quality controller, packager, and brand for content. Branding and responsibility of the quality of content calls for media companies to own, or at least to be in control of, the content essence and the production of metadata. A yet to be resolved question is, what will be the balance between the ownership of content and end-customer relationship. Although access and end-user terminal businesses seem to be the most profitable and powerful segments in the content value chain, the situation may change when more alternatives for these segments become available.

The value of content is a difficult, but important factor in determining how much the production of content essence and metadata may cost, and what the likely revenue and profit will be from the content. Some aspects and methods, such as predictive utility, customization, and avoiding repetitive information, are useful in estimating the value of content.

3 Metadata and conceptual models

This chapter concentrates on different aspects of metadata as well as how metadata and ontologies can be used to capture semantics of the content essence.

The definition for metadata used in this thesis is information about content essence, where information means relevant data associated with its content essence in such format that computers are able to use and process the data. My interpretation of metadata is a synthesis of a number of similar definitions, including for example [Berghel, 1999], who defines metadata as information about an electronic document, resource, or the operation of a computer system and [Boll et al., 1998], who define metadata as data or information about data.

My work concentrates especially on the semantics of content essence, i.e. semantic metadata. Before semantic metadata can be created and used, a model for the semantic metadata, an ontology, is needed. The latter part of this chapter concentrates on ontologies in relation to semantic metadata.

The word ontology often causes confusion due to its multiple interpretations depending on its context. The online dictionary Merriam-Webster[9] gives two alternative definitions for ontology: 1.) *a branch of metaphysics concerned with the nature and relations of being*; and 2.) *a particular theory about the nature of being or the kinds of existence*. Similarly, different fields of science such as philosophy and artificial intelligence have their own interpretations of the word ontology (see e.g. [Gruber, 1997]). In this work, ontology means a set of conceptual models consisting of agreed on concepts and relations between the concepts. The concepts bear a limited sense of meaning within them, which enables the ontology to capture the semantics at such a level of detail that content essence can be produced and delivered to the customer and used in advanced content-based products. Ontologies are often structured as hierarchies and their semantics can cover multiple different angles, dimensions, of the content essence.

For readers interested in the different aspects of ontology, [Carrara and Guarino, 2000] contains an exhaustive list of related bibliography.

3.1 Background

When we communicate, information, data with relevance, flows between different participants involved in the communication. We use language to formalize and convey intended meaning and deliver the resulting message to other participants with methods such as speech. Receivers then interpret the message according to their own understanding of the world.

Unfortunately, the receiver's interpretation of the message never completely matches the sender's meaning. The sender and receiver do not share exactly the same mental model needed for interpretation and the message cannot express all possible interpretations of the message. Moreover, the communication may

[9] www.m-w.com

contain multiple intermediate steps causing noise and resulting in interference, distortion, or fading of the original message. Figure 6 illustrates the flow of information from sender to receiver.

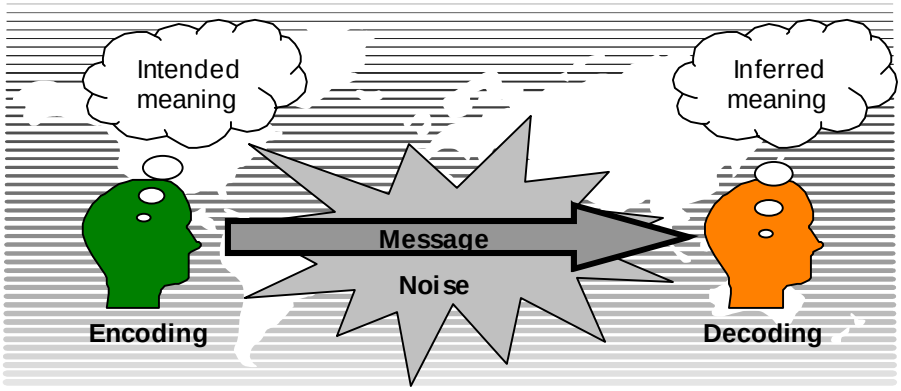

Figure 6. Flow of information between sender and receiver (adapted from [Cruse, 2000]).

If both the sender and receiver have a common method of encoding and decoding the message using a common language with agreed grammar, and shared understanding of the most important concepts in the message, a shared vocabulary, the intent of the message can be transferred from sender to receiver at a satisfactory level. Transferring the intent of the message becomes more complicated if computers participate and control the processing and interpretation of the message before it is delivered to the receiver, who might either consume the message, modify it and pass it on, or respond to it (Figure 7). We still have the same requirements for common language and shared understanding, but we also need to agree on information that is used to support the computerized creation, processing, and delivery of the message. Due to the insufficient capability of computers to extract information from free-formatted content essence such as human language, we need to be much more precise by describing systematically and explicitly the essential semantics of the message. All these requirements call for content essence augmented with semantic metadata.

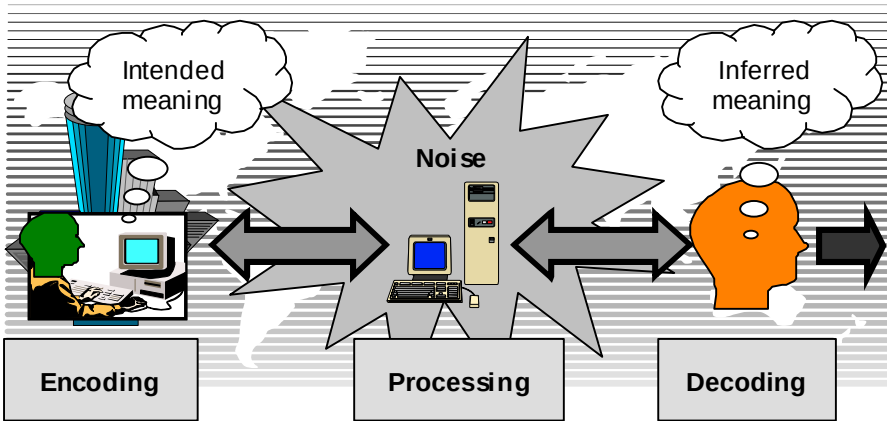

Figure 7. Complex flow of information between sender and receiver.

3.2 The nature of metadata

Metadata describes different qualities of the content essence such as format, semantics, or status. The division between content essence and metadata is not always clear and depends on the context; what may be metadata for one purpose might be considered as part of the content essence for another.

Two important agreements of metadata must be reached before it can be used in the content value chain. Firstly, there must exist an agreed format for metadata, the ***grammar***, and methods to make different formats of metadata compatible so that organizations can access and use metadata in their activities. Secondly, the semantic interpretation of the metadata forming the ***vocabulary*** must be agreed upon, so that content can be processed intelligently [Curtis et al., 1999].

Applications and users have different needs that must be reflected in the metadata. For example, if the metadata is used for producing advanced content-based products, metadata typically describes qualities related to semantics, authoring, formatting, status, and intellectual property rights of the content.

The following Table 1 describes metadata fields used in the Dublin Core standard [Dublin Core], which was developed by the library community for resource discovery on the World Wide Web. Although Dublin Core is at the time of writing one of the most widely accepted and adopted standards for describing content essence, its usefulness and expressiveness, the capability to express different semantic aspects of the content essence, is highly limited due to its simplicity and generality. Many of its fields do not have agreed vocabulary leaving a lot of room for different interpretations of the meaning of the field. For example, the *subject* -field does not have any further structure and merely reserves that field for the *topic of the content* [Dublin Core]. Limited usefulness of Dublin Core and other existing standards for semantic metadata have been a major source of motivation for metadata related work in this thesis.

Table 1. Metadata fields in the Dublin Core metadata standard (derived from [Dublin Core]).

Field	Description
Title	A name given to the resource.
Creator	An entity primarily responsible for making the content essence of the resource.
Subject	The topic of the content essence of the resource.
Description	An account of the content essence of the resource.
Publisher	An entity responsible for making the resource available.
Contributor	An entity responsible for making contributions to the content essence of the resource.
Date	A date associated with an event in the life cycle of the resource.
Type	The nature or genre of the content essence of the resource.
Format	The physical or digital manifestation of the resource.
Identifier	An unambiguous reference to the resource within a given context.
Source	A reference to a resource from which the present resource is derived.
Language	A language of the intellectual content essence of the resource.
Relation	A reference to a related resource.
Coverage	The extent or scope of the content essence of the resource.
Rights	Information about rights held in and over the resource.

Different kinds of metadata refer to the content essence at varying levels of ***granularity***. Some kinds of metadata describe overall qualities of the content essence (e.g. *content length*), while others describe just certain parts of the content essence (e.g. *opening paragraph keywords*). Metadata can be kept with its content essence (***tightly-coupled*** or ***implicit metadata***) or kept elsewhere (***loosely-coupled*** or ***explicit metadata***). The latter can be stored and transmitted separately, whereas implicit metadata is stored and transmitted together with the content essence.

Metadata is essential if the content essence cannot be used without its metadata. An example of ***essential metadata*** is information related to the compression or decryption of content essence.

Metadata can be dynamic or static, depending on its usage and the nature of the content domain. ***Static metadata*** remains unmodified from the creation to the last time it is used. An example of static metadata is author information. ***Dynamic metadata*** changes over time and requires periodic refreshing or recreation, such as updating the number of available stories in an online news portal. ***Temporary metadata*** is created only for a certain phase in the content value chain so that it may be intentionally destroyed after it has served its useful purpose. Examples of temporary metadata include status and workflow scheduling information.

Not all metadata is public. ***Private metadata*** is well defined, but its detailed description is not publicly available, although it could be obtained, for example, through licensing. Private metadata might also have been publicly defined but not yet been accepted as an identified kind of metadata by other participants of the content value chain.

The participants in the content value chain should preserve and distribute all such metadata that may have use for some other partner in the content value chain, even though the participant itself does not anymore need the metadata. For example, after creating a web page using compressed content essence, there is no need for preserving metadata related to compression, but if the compressed content essence is available to other participants in the content value chain, that metadata should be passed to other participants as well. Likewise, if a media company does not use or recognize a certain kind of special metadata incorporated in the incoming content, that special metadata may still be important for some other participant in the content value chain, and as such should not be discarded from the output.

In some cases the metadata is not needed for the actual deliverable, but it can be used to produce advanced by-products. An example of this can be seen in automobile manufacturing. Even though the physical product, a car, can hardly be categorized as an advanced content-based product, its content essence and metadata can offer new business possibilities, and used, for example, on the manufacturer's web site to promote the car.

3.3 Categorization of metadata

As we saw in the last chapter, metadata has a variety of characteristics leading to categorization from different points of view. For example, categorization can be based on the usage of metadata, stages in the life cycle of metadata such as creation, usage, and maintenance, or by the characteristics of metadata. If we

use ***role-based categorization*** as suggested by [Boll et al., 1998] for classifying different kinds of metadata, a possible division could be structural, control, and descriptive metadata.

Structural metadata describes the structural characteristics, the format, of the content essence, but does not contain information about what the content essence actually means. Structural metadata has therefore no relation to previously discussed structured content, which emphasizes the co-existence of metadata and content essence. Examples of structural metadata include decoding information such as video, audio, and graphics formats, compression data, composition and synchronization information, as well as information on sequencing the content essence. Structural metadata is often tightly-coupled with the content essence and is essential for its usage.

Different media platforms and advanced content-based products have their own requirements on structural metadata. These media platform and product specific requirements include, for example, the support for audio and video using formats like *QuickTime*[10].

Control metadata is often created and used for controlling the flow of content in the content value chain. Control metadata is used to determine whether content is ready for the next step in the content value chain. Control metadata is quite often temporary in nature as opposed to more permanent semantic metadata. Some examples of control metadata are machine control, quality of service, and error management.

Descriptive metadata can be divided into two subcategories, ***contextual metadata*** and content-based semantic metadata. Contextual metadata describes the environment and conditions of content essence and its creation. This includes geospatial information, timing information, as well as information on the equipment used to produce the content essence. Semantic metadata describes semantic qualities of the content essence answering the question what the content essence means. Semantic metadata is needed for the processing or usage of the content essence. It describes such qualities as the subject, location, names, and style of the content essence. The keywords of a news story are an example of semantic metadata. Semantic metadata is typically used in advanced content-based products, such as in a personalized news service, where the metadata is used to determine whether the user might be interested in the content essence or not. Semantic metadata requires an agreed semantic interpretation before it can be used by different participants of the content value chain. If the content provider does not produce semantic metadata according to agreed standards, or if a common agreement for interpreting semantic metadata is missing, computer-based processing that relies on the semantics of the content essence is likely to fail, or the quality of the outcome is likely to suffer.

Descriptive metadata can also contain information about how the content essence can or should be used and is thus closely related to control metadata. Examples of this kind of usage information are intellectual property and access rights, as well as information on supported media platforms.

More detailed description of metadata and their characteristics is available in the publications, especially in publication 4.

---

[10] webopedia.internet.com/TERM/Q/QuickTime.html

3.4 Characteristics of semantic metadata

This chapter introduces a number of desired characteristics for metadata and explains their relevance in describing the semantics of content essence. Although this thesis is restricted to semantic metadata, many of the presented characteristics are valid for other kinds of metadata as well. The following characteristics and their linkage to content essence result from the work done in the SmartPush project as well as from research and co-operation with a number of media companies. The SmartPush project is discussed more in detail in the chapter 6 and in the publications. Although most of the described characteristics are identified and discussed in the research literature (see e.g. [Boll et al., 1998]), the presented analysis and linkage to semantic metadata for content essence makes this list unique.

The desired characteristics for semantic metadata include ***expressiveness***, ***extensibility***, ***neutrality***, ***immutability***, ***compactness***, ***high value***, ***uniformity***, ***openness***, ***versioning***, and ***unique identification***.

Semantic metadata should have enough expressiveness to cover all the needs it will be used for. If future needs cannot be anticipated or the content domain changes frequently, metadata should be extensible to meet future requirements. Extensibility calls for the capability to describe new and unanticipated qualities of content essence without changing the underlying premises, such as the used ontology, per each required modification. This quality is essential especially in open and dynamic content domains such as *news*, where changes in the content domain are often frequent and unanticipated.

Neutrality calls for the metadata to be applicable to as many media products and media platforms as possible. If the neutrality is not achieved in the content value chain, the reuse of content might not be possible or could require major additional effort. Publication 3 discusses in detail how neutrality is linked to ***content reusability***, the reuse of content, and how media companies should modify their content -related processes to better support content reusability.

Immutability states that semantic metadata should be preserved throughout the content value chain without modifications. Although some exceptions exist, such as when metadata is rendered obsolete during its usage, participants in the content value chain do not necessarily know how the content is finally used. They should therefore pass on the content essence and metadata intact to the next step in the content value chain. In some situations, the authors of metadata are not willing to distribute the original metadata, whether it be for legal, political, business, or other reasons. As an example, stating explicitly what the content essence is about may lead to legal and other actions against the authors and distributors of content. For example, in some countries newspapers cannot in some cases state explicitly who they are writing about, even though that instance could be recognized quite easily by examining the published content essence. Likewise, distributing semantic metadata together with the content essence might allow the end-user to use the content essence for purposes other than originally planned and/or in violation of intellectual property rights. For example, end-user might reformat and repackage the content and distribute it further without respecting the original intellectual property rights. In a sense, this latter case is analogous to the software industry where software companies are typically reluctant to provide source code for their programs unless customers are willing to pay for the additional information.

Compactness states one of the major advantages of semantic metadata, namely its ability to sufficiently represent the key characteristics of potentially a much larger amount of content essence. These advantages are not necessarily obvious with textual content essence, but in the case of video or audio content, a compact representation, used e.g. for previewing or pre-processing content essence, can lead to considerable savings in required labor, computing, bandwidth, and time.

The requirement for semantic metadata to create value is derived from the business rationale. Metadata must be valuable and useful for the participants in the content value chain. Content essence that cannot be found or used is worthless, as stated in [Berghel, 1999]. Similarly, only valuable and useful metadata should be produced. According to my experience, media companies often ignore the requirement for metadata to have high value. The reason for this ignorance is at least partly related to the difficulty in estimating the value of semantic metadata as described previously in this work. Media companies do not know how to measure value, or find it difficult and/or expensive to measure, and thus cannot use that information to guide the production of metadata. Although I have scoped out more detailed analysis on methods for its measurement, value of information is such an important aspect of metadata and business rationale that it requires further research.

Uniformity states that all participants in the content value chain should have standardized and formalized methods to describe metadata. This requirement does not restrict the usage of content for different purposes, but assists in preventing the metadata from having multiple and possibly unknown formats. Uniform semantic metadata allows computerized processing and may lead to better quality of the advanced content-based products by requiring less intelligence from the automated process steps in the content value chain. Uniformity is connected to openness, which requires the standard for semantic metadata to be open and its detailed description and usage to be available to others. Uniformity and openness lead to the need to have standards for describing semantics such as Dublin Core, which was introduced earlier. Using a standardized mechanism for transmitting and sharing content essence such as Extensible Markup Language, XML [Bray et al., 1998], or its derivatives is not enough. If semantic metadata is not made public or if different partners in the content value chain interpret the meaning of semantic metadata differently, the usefulness of such metadata is greatly reduced. On the other hand, concealing the meaning of semantic metadata can be used as a form of control. If the authors of metadata do not want to allow others to use semantic metadata without their permission, they might decide not to release semantic metadata or its interpretation to others in the content value chain.

Versioning tracks changes in an ontology and allows capturing the currently essential semantic characteristics of the given domain. Versioning requires that the corresponding version of ontology is identified and linked with the metadata when semantic metadata is created. When premises in the corresponding domain change, a new version of the ontology for content essence is created. Versioning allows for an easier expansion of ontology and the correct functionality of advanced content-based products.

Unique identification is closely connected to versioning. All content should have mechanisms for unique identification so that required content essence and metadata can be identified and used. If metadata is

loosely coupled with content essence, each of them will require unique identification. If metadata is implicit to the content essence, the unique identification of content is enough.

To conclude the discussion on characteristics of metadata, the following Table 2 introduces some of the pros and cons of semantic metadata. Despite of the existing challenges, semantic metadata clearly has an important and justified role in the content value chain. Publication 4 discusses in more detail the pros and cons of using semantic metadata.

Table 2. Pros and cons of semantic metadata.

Pros	**Cons**
Compact and independent representation. Semantic metadata can capture the essential semantics of the content essence without the need to use a full version of the content essence, leading to computational and bandwidth savings and allowing additional functionality, such as previewing the content essence. When semantic metadata is treated separately from the content essence, it is possible to have different kinds of metadata for different purposes. In addition, that metadata can be modified, processed and distributed without the need to access the original content essence.	**Expressiveness**. Semantic metadata is able to describe only a subset of the original meaning of content essence, which may not be sufficient for unanticipated needs and future use of the content essence.
Explicit statement of author intention. Semantic metadata helps to state explicitly key facts of the content essence while it remains independent of its representation, such as choice of wording or use of language.	**Explicit statement of author intention**. Semantic metadata may force the author to state explicitly a certain message the author wanted to hide between the lines.
Uniform format. Semantic metadata can be used to describe content essence that exists in a variety of types and formats, such as text, images, video, or audio.	**Extra effort and costs**. Additional effort and expertise are required to produce semantic metadata, leading to higher costs. Effort is also required to create ontologies and to keep them updated.
Systematic and standardized representation. Semantic metadata structures the content essence and allows easier computerized processing of it. This leads to more accurate processing of the content essence and better management of media products. Semantic metadata may allow for process simplifications and a higher level of automation of the existing process to produce content.	**Process impact**. Semantic metadata changes the publishing and distribution of content, requiring additional process steps, technologies, time, skills and resources. Metadata also requires closer co-operation and standardization between different participants in the content value chain.
Higher quality. Semantic metadata has the potential to raise the quality of advanced content-based products by reducing the need for the automated analysis of content essence within the processing and delivery of content.	
New advanced content-based products. Semantic metadata allows advanced content-based products such as personalized information feeds.	

3.5 Conceptual models

Before semantic metadata can be created and used, a suitable ontology for the content essence is needed. An ontology comprises a set of conceptual models, also called ***metadata structures***, which are representative of the content domain. Each conceptual model consists of concepts and their relations. These concepts and relations define the semantics that each concept in the conceptual model may possess, and how the concepts are related to each other in the conceptual model.

The SmartPush project, which is discussed later in the thesis, concentrated on building ontologies for news content domain in textual format, where ontologies are largely derived from the print media categorizations. However, direct derivation is often not optimal for digital content. Media companies have often tailored their ontologies with a printed publication in mind. The assumptions of the printed media do not fully match the possibilities of digital content and advanced content-based products, such as the augmentation of the original content with other related information or the reformatting of the content based on user interaction. Publication 4 discusses the methods for building ontologies for content related to news, later called ***news content***.

One possible approach for defining an ontology for a particular content domain is to analyze relevant linguistic elements used in language and then to apply them in building the ontology.

Some of the semantically defined linguistic elements impacting ontologies include ***homonyms***, ***synonyms***, ***taxonomic*** (is-kind-of/is-a) and ***meronymic*** (part-of) ***hierarchies***, nets, and clusters [Cruse, 2000].

The main focus of my work, derived from the SmartPush project, has been on hierarchically organized conceptual models. Hierarchies are based on two main associations, ***dominance*** and ***differentiation***. Dominance defines the relation between parent and children, and differentiation is the difference between siblings in the hierarchy. A characteristic of a well-formed hierarchy is that the branches of the hierarchy never merge again; i.e. the hierarchy forms a tree.

The SmartPush project utilized a number of advantages of hierarchies in the personalization of content essence. Human intuitively use hierarchies to navigate through a multitude of elements. Concepts organized into hierarchies allow the reporter to describe the content essence at different levels of detail. Hierarchies help to create summarizations of the more detailed concepts in the ontologies, and to group concepts based on their relation to others. Hierarchies seem also to be a suitable method for decreasing the amount of required calculations in personalization based on semantic metadata. They can also be used to navigate a multitude of concepts during the creation of semantic metadata, as long as the ontology in general is not overly complex.

Hierarchies have negative characteristics as well. They may be too restrictive and force top-down structuring of concepts, which might not be a natural way to describe semantics. Because hierarchies do not allow multiple parents, avoiding these situations may lead to complex and unnecessary duplicate structures in ontologies.

Most of the conceptual models used in the SmartPush project were taxonomic hierarchies. In taxonomic hierarchies, dominance is based on ***hyponymy***, i.e. the super-type / sub-type relation between parent and

one or more children. Differentiation in taxonomic hierarchies uses exclusive ***co-taxonomies*** starting from each child, so that each child represents a different type of a common parent [Cruse, 2000].

A well-formed taxonomic hierarchy offers an efficient and orderly set of categories at different levels of specificity. Figure 8 shows a simplified example of a taxonomic hierarchy based conceptual model. It helps to understand what kind of levels might be included in a typical taxonomic hierarchy, and how those levels could be connected in a conceptual model. As one can see from the example, concepts in the conceptual model overlap easily leading to further requirements and restrictions regarding the use and interpretation of the conceptual model and semantic metadata.

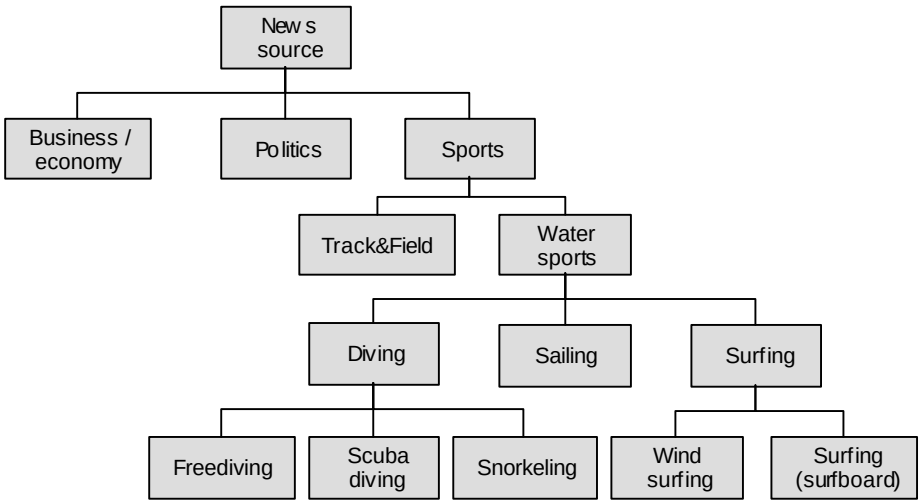

Figure 8. A simplified example of a taxonomic hierarchy based conceptual model.

Humans have occasional gaps in the language so that we do not have a well-defined concept to describe a certain entity or aspect. In this case, humans tend to use ***autotaxonomy*** to extend the meaning of the concept above or below the gap. This is the case with surfing in Figure 8, where the same term is used as a category name for certain type of *water sports* as well as for a specific type of *surfing* with a *surfboard*.

Hierarchies can also be built by using meronymic dominance, where elements are sub-parts of larger entities such as *finger/palm/hand*. In meronymic hierarchies, differentiation is based on ***co-meronymy***, i.e. sister parts do not overlap. A typical example of a meronymic hierarchy is the *body part* ***hierarchy***, where different parts of the body do not overlap. Meronymic hierarchies do not have different levels of detail without the context defining a specific one. If two candidates are on a same detail level, such as the *carburetor* and *steering wheel* in the *car-part* hierarchy, it is difficult to determine which one is higher.

As these examples show, elements in the hierarchies can be interpreted from different points of view. Some possible alternatives include: 1.) seeing something as a whole, consisting of parts; 2.) seeing something as a kind of a larger group, in contrast with other kinds; 3.) seeing something as having a certain function; and 4.) seeing something from the point of view of its origins. It is difficult to build a hierarchy if the purpose and usage of the hierarchy is not clear. This confusion can be greatly reduced by keeping the viewpoint constant and defining a separate hierarchy for each viewpoint, thus creating a dimension. If the viewpoint is

not clearly defined, the concepts in the hierarchy may have different interpretations and complex interrelations between dimensions.

The goal of the dimension must be clear, and it must capture the relevant qualities from that viewpoint. Typical dimensions that were used in the SmartPush project include *Locations*, *Industries*, *Companies*, and *Subject*.

Figure 9 shows an example of ontology and dimensions as well as how the concepts are mapped to corresponding terms in the content domain.

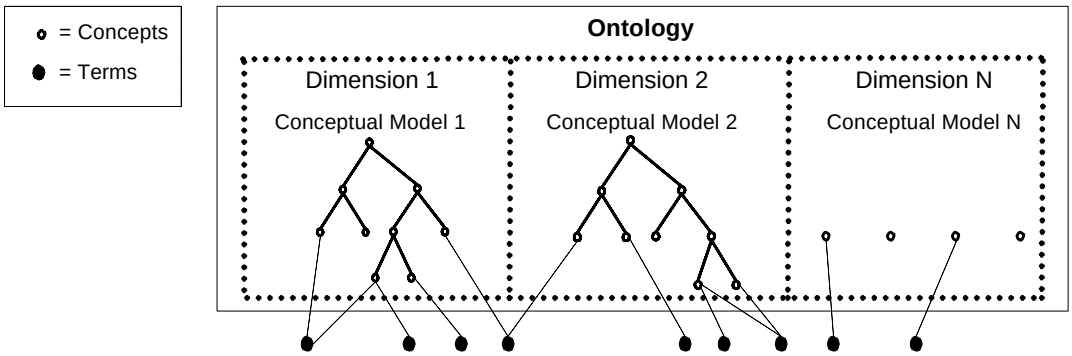

Figure 9. An example of ontology, dimensions and their relation to external terms in the content domain (adapted from publication 4).

Figure 10 shows a simplified example how dimensions, concepts, and terms could be used in practice.

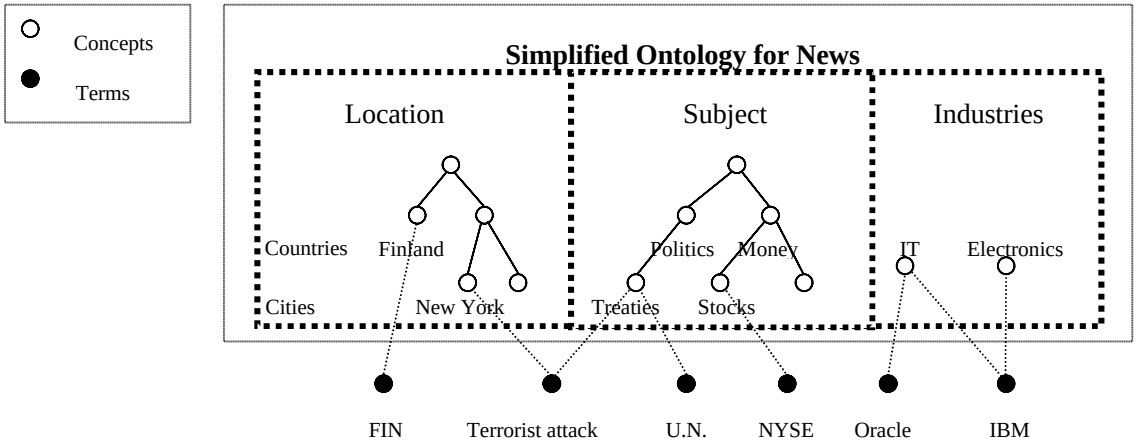

Figure 10. A simplified example of an ontology in practice.

By keeping the conceptual models moderately abstract and isolating the concepts from the terms in the underlying content domain, we can avoid some of the challenges related to changes in the content domain.

One of the initial goals of the SmartPush project was to divide the content essence into separate conceptual models that had little or no relationships with other dimensions. Although concepts in conceptual models represent a contextually limited interpretation of their full meaning, further work with ontologies showed that dimensions are more or less interrelated so that independence between dimensions was a controversial goal. From the perspective of metadata management, keeping the dimensions independent helped to reduce

complexity. On the other hand, independence limited the expressiveness of the ontology, because, in reality, the concepts are often dependent on and interact with concepts in other dimensions.

3.6 Metadata standardization

This chapter contains a brief overview of some standards related to semantic metadata. As these stardards are emerging, evolving, and changing rapidly making the comparison difficult and the results quickly outdated, I have excluded detailed and exhaustive analyze from my work and present just few of the standards that are related to news content and relevant to this work. Comments on these standards are partially my own observations, partially derived from references and other materials discussing standards on news content such as [Dumbill, 2000].

The standards presented in the Table 3 have varying degrees of ambitiousness and are at varying levels of completeness and implementation. They also more or less overlap each other. Development in the standardization of metadata is fast, and changes take place frequently, so it is important to check the validity of presented information from the references. Readers interested in more detailed information on the standardization of metadata should see for example [Saarela, 1999], which contains detailed information on XML and associated standards.

Standard Generalized Markup Language, SGML [ISO8879] and Extensible Markup Language, XML form the basis for most metadata standardization efforts. They define basic methods for using tags for marking semantic metadata in electronic documents, but leave the definition of those tags to the users.

Although the two core standards, SGML and XML, are relatively mature and established, standards related to semantic metadata are still insufficient to describe content essence in greater detail, even though some alternatives, such as NewsML[11] and industry specific efforts[12], have lately focused their efforts towards this goal.

The creation of semantic metadata related standards is a challenging task and requires participation and agreement of impacted participants in the content value chain. The lack of sufficient standards has motivated media companies to develop their own proprietary standards, or to enhance the existing standards with proprietary extensions. However, this development is likely to be a temporary step on the way to more shared and open standards. For example, the *Reuters* news agency has developed its own proprietary standard for describing content. If a content provider does not adhere to the standard, its content is excluded from distribution. Consequently, *Reuters* has lately been shifting away from proprietary solutions and into more open alternatives, such as NewsML [Reuters, 2000].

[11] www.iptc.org/site/NewsML/NewsMLSpec.htm

[12] www.xml.org/registry/index.shtml

Table 3. Some standardization efforts related to semantic metadata for news content.

Standard	**Purpose and brief description**	
	Pros	**Cons**
Dublin Core[13]	Dublin Core is a standard for metadata developed by the library community for resource discovery on the World Wide Web. It captures both semantic metadata and contextual metadata about the content essence. Dublin Core was designed to enable searching of documents across heterogeneous databases and schemas by capturing bibliographic and other descriptive information of content essence. It provides a simplified set of 15 metadata fields that form the core ontology of the standard.	
	- Is widely used for archival purposes in libraries. - Has international consensus as a mature and robust standard. - Is simple and open standard that has been used as a core in other standardization efforts.	- Has limited support for semantic metadata. Dublin Core requires additional agreements before it is suitable for describing semantic metadata that can be used automatically by computers. - Extension beyond the agreed set of metadata fields requires changes in the standard.
ICE[14]	Information & Content Exchange, ICE, is an XML-based standard defining a vocabulary and protocol for the delivery of content and the management of relationships in syndication. The ICE protocol defines the roles and responsibilities of media companies and their customers, defines the format and method of the delivery of content, and provides support for management and control of relationships in the content value chain.	
	- Has some commercial implementations and relatively large support among media companies and their customers[15]. - Manages an important part of the content value chain, namely the delivery of content and relationships between media companies and their customers.	- Is not developed by standards body W3C, which in turn weakens its acceptance outside of the originating companies. - Public version does not have support for describing semantic metadata. - The future of ICE is unknown; it will most likely be merged with other standards.
NewsML[16]	NewsML is an XML-based standard for combining, representing and managing news content irrespective of its media platform, format, or encoding. NewsML is developed by the International Press Telecommunications Council (IPTC). NewsML extends beyond newswire providers and newspapers and focuses on a wider variety of participants in the content value chain. The first version of the NewsML, *NewsML v1.0*, was released to public in October 2000.	
	- A number of large companies and organizations have participated in the standardization effort. - Semantic metadata in NewsML is based on the widely accepted NITF standard (although it can utilize other standards as well). - Is independent of media platform.	- Suffers from overlapping and compatibility issues, especially in relation to semantic metadata, with other standards such as PRISM. - Is not yet in widespread use and implementations are rare.

[13] www.dublincore.org

[14] www.icestandard.org

[15] For an example of a tool supporting ICE, see *Vignette Content Syndication Server V6* at www.vignette.com

[16] www.newsml.org

Standard	Purpose and brief description	
	Pros	**Cons**
NITF[17]	News Industry Text Format, NITF is a widely used XML-based standard for marking up the semantic metadata of news content. NITF was developed in co-operation between two major standards organizations in the news industry, *the International Press Telecommunications Council* (IPTC) and the *Newspaper Association of America*. The following characteristics are covered in the newest version of NITF, version 3.0: Who the news is about, who owns the copyright, and who may republish it. What subjects, organizations, and events the news covers. When the news was reported, issued, and revised. Where the news was written, where the action took place, and where it may be released. Why the news is newsworthy.	
	- Powerful organizations and companies are involved in the standardization effort. - Is used as a basis in other standards such as NewsML.	- Support for other than news-based media products is unclear. - Does not contain methods for defining ontologies.
PRISM[18]	Publishing Requirements for Industry Standard Metadata, PRISM is an XML-based standard for describing metadata, which is needed in syndicating, aggregating, post-processing and reusing content in magazines, news, catalogs, books, and journals. PRISM aims to provide standardized vocabularies for metadata in order to enable the interoperability of all kinds of content. The PRISM authoring group released version 1.0 of PRISM to the public in April 2001.	
	- Has a wide coverage over a multitude of characteristics required in the content value chain. - Reusing already existing standards and alliances such as RDF and Dublin Core speeds up development and acceptance. - Has some support for defining ontologies.	- Is not in widespread use due to its novelty. - Acceptance outside of originating companies might be limited due to the closed nature of how the standard was developed. - Semantic capability is dependent on external standards used as part of PRISM. - Compatibility and overlapping with other existing standards might cause issues. PRISM supports only a subset of RDF. Additionally, PRISM seems to overlap with NewsML. Even though the PRISM group is working on compatibility with NewsML, the outcome is still unclear.

[17] www.nitf.org

[18] www.prismstandard.org

Standard	Purpose and brief description	
	Pros	**Cons**
RDF[19]	Resource Description Framework, RDF [Lassila and Swick, 1999] is an XML-based standard providing a common syntax to describe metadata about a resource such as a document on the World Wide Web. RDF is a recommendation from the *World Wide Web Consortium* (W3C) and is used as the core mechanism for Semantic Web at the W3C enabling the exchange of knowledge on the World Wide Web.	
	- Has loose coupling, which allows for the independent processing of semantic metadata and content essence. - Allows the use of multiple metadata standards through the XML namespace mechanism. This makes RDF flexible and extensible for future needs. - Adds structure to content and improves the computerized use of metadata as long as the semantics of the ontology are shared. - Has potential to become the default structure for describing semantic metadata.	- Simple metadata without complex relations can be described without RDF questioning the need to learn or use RDF. - RDF specific tool support is still limited. - Syntax is moderately complex, at least for non-programmers. - Does not define any ontology for semantic metadata. An external ontology is required before RDF can be used to describe semantics.
RSS[20]	RDF/Rich Site Summary, RSS is a simple format for metadata related to online news. RSS has evolved through different versions without backward compatibility such that versions 0.91 and 1.0 are in some cases considered as different standards. Version 0.91 is inspired by RDF, but not strictly conformant to it. Current version, *RSS 1.0*, is compatible with RDF and has roughly the same descriptive abilities as PRISM and NewsML.	
	- A lightweight and flexible standard that is easy to understand and simple to use without large investments in resources.	- Compatibility problems between different versions. - Future of RSS is unknown. *Netscape*, the main advocate for RSS, dropped support for RSS recently, and its role as the host for the standard has not yet been filled.
XML News[21]	XMLNews is an XML-based standard for describing the semantics and context of content essence. XMLNews consists of two parts: 1.) *XMLNews-story*, designed as a subset of the older 1998 version of NITF, describing the semantics of content; and 2.) *XMLNews-Meta*, an RDF-based extensible vocabulary describing news resources.	
	- Used in commercial applications and newsfeeds from companies such as iSyndicate[22]. - Relatively simple.	- Availability and quality of semantic metadata varies in implementations. Some sources use only some of the possible tags, and/or their semantic metadata has low quality. - Contains only basic support for semantic metadata, which is then extended with non-standardized, source-specific semantic metadata. - Standard is outdated and not necessarily compatible with other standards. XMLNews is likely not to be developed further as other standards are taking its place.

[19] www.w3.org/RDF

[20] www.xmlnews.org/RSS/, purl.org/rss/1.0/spec

[21] www.xmlnews.org

[22] www.isyndicate.com

In addition to the presented standards, ***Semantic Web***[23], originated by the W3C, may have an impact on the content value chain. The Semantic Web aims to create a machine-usable network of trust in which rich metadata is machine-usable, semantically flexible, and derived from trusted sources. As the creators of the Semantic Web state[24]:

> *"The Semantic Web is a vision: the idea of having data on the web defined and linked in a way that it can be used by machines not just for display purposes, but for automation, integration and reuse of data across various applications."*

There are other existing and upcoming standards that are not discussed further in this work. These include DAML[25], Dewey decimal classification[26], MARC[27], MPEG-7 [Martínez, 2000], MPEG-21 [Bormans and Hill, 2001], RDFS [Brickley and Guha, 2000], SHOE[28], TopicMaps[29], UDDI[30], XTM [Pepper and Moore, 2001], and a multitude of other standards. In addition to these mostly XML -based efforts, different companies and organizations have been active in defining industry-specific ontologies. Examples in this area include BizTalk[31], CommerceNet[32], the Open Directory Project[33], and RosettaNet[34].

As standards for metadata emerge and evolve quickly, without clear indication how successful these standards will be, I will refrain from further analysis and advise the reader to have a further look at standards bodies and organizations, such as *IDEAlliance*[35], IPTC[36], *ISO*[37], and W3C[38].

3.7 Summary of findings

Computerization increases the need to have structured and formalized communication between the sender and the receiver. Advanced content-based products require content consisting of content essence and metadata, especially semantic metadata.

---

[23] www.semanticweb.org

[24] www.w3.org/2001/sw

[25] www.daml.org

[26] www.oclc.org/oclc/fp/index.htm

[27] www.loc.gov/marc

[28] www.cs.umd.edu/projects/plus/SHOE

[29] www.topicmaps.org

[30] www.uddi.org

[31] www.biztalk.org

[32] www.commercenet.com

[33] www.dmoz.org

[34] www.rosettanet.org

[35] www.idealliance.org

[36] www.iptc.org

[37] www.iso.ch

[38] www.w3c.org

Metadata can be categorized based on a number of different qualities. Semantic metadata is essential to describe and communicate the semantics of the content essence.

My work at the SmartPush project, in close collaboration with media companies, has resulted in a list and analysis of desired characteristics for metadata. These characteristics include expressiveness, extensibility, neutrality, immutability, compactness, high value, uniformity, openness, versioning, and unique identification.

I have concentrated in this work especially on semantic metadata. Although semantic metadata complicates the content value chain, its importance in the content value chain and advanced content-based products justify its existence.

Ontologies define the roles and relations of the various semantic aspects of content essence. Based on our experiences in the SmartPush project, ontologies should capture different characteristics of the content domain in dimensions that are as independent as possible in order to avoid complex dependencies between concepts in different dimensions. A taxonomic hierarchy seems to be a suitable conceptual model for capturing the semantics of content essence, but selection of the most suitable structure for each individual dimension depends on the internal characteristics of that particular dimension.

If the metadata is to be effectively used in the content value chain, participants in the content value chain must agree upon the structure of the ontology, or how to map between different ontologies. Some standards for metadata exist to describe certain semantic aspects of content essence, but standards in general do not yet sufficiently cover the semantics of the content essence.

4 Domain modeling

The last chapter discussed the importance of content essence, semantic metadata, and ontologies. This chapter builds on that foundation and discusses domain modeling, where the important semantics of a certain content domain are formalized into an ontology.

This discussion begins by approaching domain modeling from a linguistic perspective and by introducing how researchers in artificial intelligence have approached modeling expertise. After that the chapter continues with more detailed information on aspects related to the modeling of content domains.

4.1 Using metadata to represent meaning

Researchers in linguistic psychology have been intrigued by the creative use of human language, such as how children are able to manage complex structures in our language without formally learning them from their environment. For example, the well-known researcher Noam Chomsky states in his book [Chomsky, 1988]:

> *"The speed and precision of vocabulary acquisition leaves no real alternative to the conclusion that the child somehow has the concepts available before experience with language and is basically learning labels for concepts that are already part of his or her conceptual apparatus."*

This quote supports my view according to which it is possible to model a content domain and form a pre-defined set of domain-related concepts and relations between them as well as to use these concepts to describe the semantics of content essence.

Humans use words and their meanings in many different ways to express their thoughts [Cruse, 2000]. The words we use may have one or more different meanings, depending on the context in which they were used. We use ***superordinates*** and ***metonymies*** in our communication. Superordinates use one member of a group to represent a larger entity, such as using *a man* to represent all humans. Metonymies are figurative expressions in which one word or phrase is substituted for another with which it is closely associated, such as in the use of *Washington* for the *United States government*.

Before we can start building an ontology, we must have sufficient understanding how meaning is defined and how these ideas can be applied to the semantic metadata.

Conceptual and contextual nature of meaning is important. According to some linguists, language elements have a range of interpretations, the correct interpretation being based on the surrounding elements of language, such as related nouns and verbs [Cruse, 2000]. This same view is supported in the *use theory of meaning* proposed by [Wittgenstein, 1953], who stated:

> *"The meaning of a word is its use in the language."*

Figure 11 illustrates some of the challenges in defining semantics for content essence.

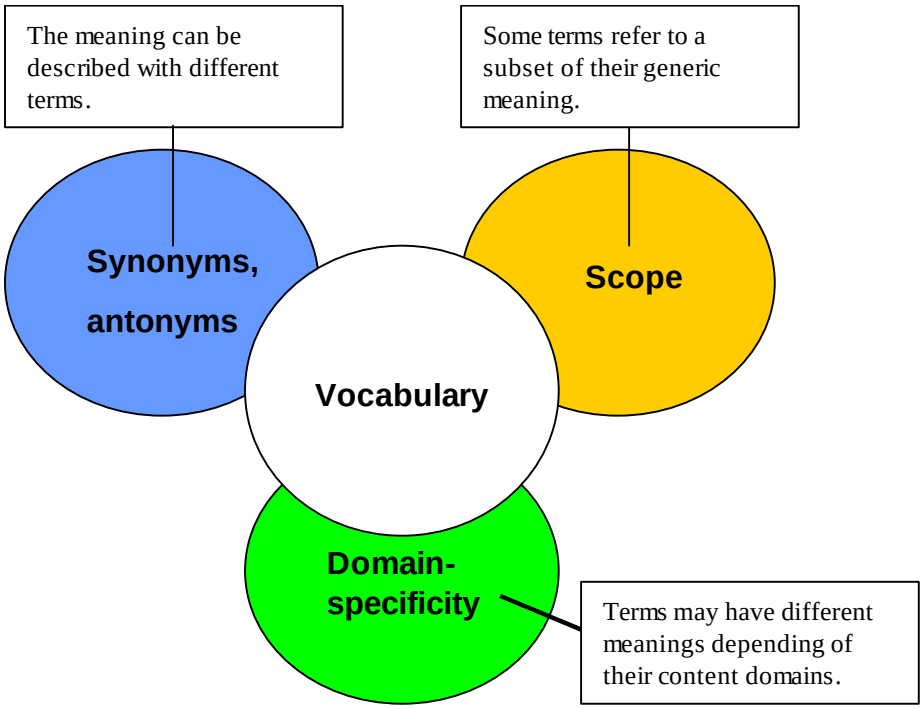

Figure 11. Challenges in defining semantics of a term (adapted from [Wiederhold, 1995]).

In a study researching readers' interests of the Usenet News, the readers seemed to share a sufficiently common conceptual model for building individual models of readers' interests [Stadnyk and Kass, 1992]. Although it is impossible to achieve a common agreement and interpretation on all possible matters in the world, I believe that essential semantics of the content essence can be captured by structuring and defining the important characteristics of content essence into an ontology, and by limiting the descriptions to a certain content domain and detail level. These semantics can then be used to describe content essence at a sufficient level to be used in various advanced content-based products.

4.2 Modeling expertise

The construction of an ontology is laborious and requires close co-operation with experts in the content domain. A number of artificial intelligence researchers have spent a considerable amount of time trying to define different generic models to be used as templates for expert systems (see e.g. [Shadbolt and O'Hara, 1997]). In essence, these models try to codify the expertise of the experts, and then generalize that knowledge for later use.

[Feltovich et al., 1997] present a number of issues that impact the complexity of modeling a certain domain of expertise. Based on my experience, some of the issues are very relevant to the creation of ontologies for content domains. The most relevant of these issues, reflected to building ontologies for content essence, are:

- **Dynamic nature of the domain**. Can the important aspects of the content essence be captured by describing the current state of the content domain, or are the most important aspects contained in the changes in the content domain? For example, the content essence used for business intelligence should concentrate on tracking changes in the competitive landscape, such as management moves

as well as mergers and acquisitions, instead of merely stating the current situation. If changes are important, an ontology must be created with this requirement in mind.

- **Dependence of the context**. Does the creation and use of metadata have specific context-dependent rules and operations, or is the creation and use relatively uniform independent of the context? If some characteristics of context have an impact on the metadata and its use, these characteristics must be captured as part of the ontology. For example, content essence regarding geographic information, such as a country, may be captured without further information about the context, whereas other characteristics, such as quality, might require further information on the source and method by which the quality was measured.
- **Homogeneity of the content value chain**. Do different participants use shared, uniform vocabulary and methods across the content value chain, or are their methods and vocabulary diverse? This issue impacts the need to have multiple ontologies and conversions between them at different steps of the content value chain.
- **Surface/deep knowledge**. Can a single line of explanation convey a concept, or are multiple and overlapping lines of explanation required for adequate coverage, making the automated creation of metadata complex and difficult?
- **Number of perspectives of the content domain**. Do elements in a situation afford a single interpretation or categorization (or just a few of them), or are multiple representations required possibly leading to having multiple dimensions in the ontology?

[LaFrance, 1997] compares four different approaches used to capture expert knowledge: ***excavation***, ***capture***, ***courtship***, and ***creation***. Although challenges in the expertise modeling are to some extent different and more complex than in building ontologies for content, approaches used in the expert systems provide useful alternatives to model semantics of content essence as well. Table 4 illustrates these four approaches for modeling expert knowledge.

Table 4. Different approaches for modeling expert knowledge ([LaFrance, 1997]).

Approach	**Emphasized features**	**Features with lesser attention**
Excavation	Expertise is a thing, requiring straightforward, labor-intensive elicitation effort.	Expertise is dynamic and context-dependent, calling for attention to personalities and context.
Capture	Expertise is often tacit and hidden, calling for attention to take charge and claim the relevant body of knowledge.	Much expertise is public and articulated. Also, expertise is subtle and experts are sometimes willing to co-operate.
Courtship	Expertise is highly personal and idiosyncratic so that elicitation depends a lot on the relationship and co-operation with the source.	Expertise is often transferable and equivalent across experts. This approach also neglects the fact that knowledge might actually be retrieved from a team of experts.
Creation	Expertise emerges in the process of eliciting it. This calls for special attention in order to avoid the modeling of inaccurate or specious knowledge.	In some content domains, there exist a priori criteria for recognizing expertise. The same expertise may also originate from multiple sources and through different methods.

One method used to model knowledge in expert systems is to let the expert think aloud and explain the steps taken while he or she performs the task requiring expertise [Ericsson and Charness, 1997]. Therefore, we could ask the persons authoring content essence to explain what kind of ideas are being incorporated in the content essence and how they are going to do that. As a result, we might be able to capture both the implicit semantic message of the content essence as well as the methods with which that message is typically expressed as long as the content domain is not overly complex. By reverse engineering the process we might also be able to improve the methods for extracting semantic metadata automatically from the content essence.

According to the research of expertise, the primary mechanism for creating expert-level performance is deliberate practice, wherein trainees perform representative tasks, receive immediate feedback and have the possibility to correct or refine their ontology. Furthermore, the skills related to encoding and accessibility in specific activities are very domain-specific, and thus are unlikely to be transferable from one content domain to another [Ericsson and Charness, 1997]. In regards to content essence, these findings could be interpreted that skills learned to codify and categorize content essence for a certain content domain take time and they might require substantial rework to be applicable to another content domain. My own experience supports this theory, although I have not performed any measurements or analysis of the involved difficulty.

According to [Zeitz, 1997], an important part of becoming an expert in a certain content domain is to build a sufficiently high-level abstraction of that domain. Zeitz calls this a *Moderately Abstract Conceptual Representation (MACR).* The main advantages of *MACR,* as opposed to a more detailed representation of expertise, are:

- **Stability**. Detailed information may deteriorate quickly due to frequent changes at a certain level of detail.
- **Accessibility**. A *MACR* can be more accessible since it can be retrieved by a broader range of retrieval cues.
- **Schematic nature**. The *MACR* helps to process and manipulate ill-structured domains because of its schematic usage.
- **Elimination of irrelevant details**. Processing nonessential details in a concrete representation may not produce any accuracy gains and may actually interfere with successful reasoning.

By applying these steps to domain modeling and the creation of semantic metadata, a higher-level structure should assist in capturing the longer-lasting qualities of the content essence, and it reduces the need to concentrate on non-essential details and to constantly update conceptual models.

The problem with abstractions is that they represent a mere portion of the reality from which they were derived. In addition, they can also be reductive of the reality to which they are applied. If this is not the case and each experience is treated as unique and in its full complexity, abstractions would lose their value [Feltovich et al., 1997]. In the context of my work this statement means that an ontology should cover the

content domain in such a way that the resulting semantic metadata is a condensed, yet representative, description of the content essence.

Borrowing from the field of artificial intelligence, the notion of negative knowledge has an impact in the creation of ontology. Competence requires one to know what one must do in a certain situation, but it often requires that one knows what not to do as well [Minsky, 1986].

Although we did not implement any functionality for managing negative knowledge in the SmartPush project, we did identify and discuss the importance of using negative feedback on recommendations as part of the personalization of content essence.

In conclusion, there are useful methodologies and principles in expert systems and expertise modeling that should be taken into account when creating ontologies for content essence.

4.3 Ontology development

The goal of the ***ontology development*** is to formalize and model the essential semantic characteristics of the content domain. During the ontology development the key assumptions, vocabulary, and principles of the content domain are made explicit, represented, and described. Explicit representation and descriptions reduce the possibility of ambiguity, which often results from different assumptions, vocabulary, and principles between users of the same content essence.

There are two relevant questions in defining ontologies to describe content essence: 1.) is it possible to formalize the conceptual models for the content essence; and 2.) are reporters able to produce the same semantic metadata uniformly independent of the person and the creation time of semantic metadata? Although I have not been able to fully validate these questions, discussions with media companies and observations made during the SmartPush project indicate that semantic metadata can be produced uniformly and independent of the time or person, although this process requires a substantial learning effort, constant communication with other persons creating the semantic metadata, explicitly stated and shared semantic explanations of the used ontology, as well as continued participation in the production of semantic metadata.

It is important to note the iterative nature of ontology development and the need to reflect changes in the environment back to the ontology. Based on my experience, the development of an ontology is difficult and requires constant revisions before the ontology becomes useful. Furthermore, even if an organization is finally able to produce a satisfactory ontology for a certain content domain, the content domain will inevitably change. These changes, such as emerging new concepts or semantic changes in old concepts, must be reflected back to the ontology. If the ontology does not adapt to the changes in the environment, the ontology will gradually deteriorate and no longer be able to capture properly the semantics of the content essence.

Ontology development should concentrate only on those qualities of content essence that are relevant and valuable for the ***intended use***, taking into account the number of users, how much value metadata produces for these users, and how much costs the creation and use of this metadata generates. If the metadata serves no real purpose or if the costs of creation exceed its value, there is no reason to capture such qualities into

the ontology. Unfortunately, some of the qualities that make content essence valuable may be difficult to obtain or are contradictory to the production of content. For example, many users value the ***timeliness*** of information, but if the distribution requires metadata, this requirement introduces additional delays in the production of content. One solution to this kind of a problem is to produce the metadata in two phases, initially providing just a minimal set of metadata needed to direct the content essence through distribution, then later augmenting the metadata with more extensive semantic metadata.

Even though the main emphasis in this work from reusability perspective is on the reuse of metadata, ontologies can also be at least partially reused. [Kim, 2000] introduces useful principles for developing ontologies that are built upon existing data. These methods support limited reuse of ontologies and are also applicable to my work. Kim separates application-specific and generic ontology engineering efforts, of which the former concentrates on specific applications and the latter generic sharing and re-use of content essence. The author also states the need to understand the content essence, the actual ontology, and context, consisting of the existing environment and constraints, during the engineering of the ontology. The main steps of his methodology are:

1. **Motivating scenario development**. This is a detailed narrative, a business case, which describes the tasks to be performed and the problems that exist related to these tasks. For example, this could be a description how a company tracks its competitive space and what kinds of difficulties are encountered during these activities.

1. **Stating ontology requirements**. Keywords and phrases are extracted from the above-mentioned scenario and then tested via business and technical questions for their role and importance. If a term is related to a significant question, it is taken further in the process. For example, this step could identify concepts, such as *competitors*, *products*, *technology x* from the scenario described in step 1.

2. **Content domain analysis**. During the analysis of a content domain, the key concepts of the modeled world are distilled and their internal structure is developed. In our example, this would mean further analysis in *competitor tracking* leading to key concepts, such as *competitors*, *suppliers*, *alliances*, *products*, *technologies*, and *influencers*.

3. **Application characterization**. Relevant facts about the world that affect the context of an ontology are stated so that they can be taken into account in designing the ontology. Continuing with the given example, this could result for example in how frequently the given key concepts, such as *technologies*, change in the *competitor tracking* content domain and how the application will be used.

4. **Application ontology and shared ontologies development**. Application characterization is used to develop a system architecture, and an analysis of the content domain is used to define the representation of the ontology. In the given example, this step could lead to a categorization of identified concepts as being unique for this application, such as *alliance x*, or being usable in a number of applications.

5. **Verification of the model**. The resulting application and shared ontologies are verified against the business case, requirements of the ontology, key concepts, and application characteristics.

Kim's methodology has important ideas, such as the identification of different types of ontologies, and the rooting between ontologies and underlying business reality, but it is weak in the iterative nature of ontology development. Furthermore, the methodology could be more specific in identifying the needs of users and describing how characteristics of the available information impact the resulting ontology. These issues are discussed more in detail next.

4.4 Principles of the domain modeling activity

Early in the SmartPush project I received a valuable piece of advice to 'know the domain'[39]. If the content domain is not limited in any way, it seems to be extremely difficult to define any general-purpose ontology that could represent the content essence for all possible uses. [Lenat, 1998] discusses the same quality, the need to divide world into smaller, internally coherent content domains, as being the most important lesson his team has learned in the 15 years of gathering and codifying common knowledge. The focus on content domain is also evident in the language structures for human communications, as we saw earlier in connection to semantic metadata.

Based on the results of the SmartPush project, I have defined domain modeling to be affected by three related elements: ***information feeds***, ***user needs***, and intended use. Information feeds require domain modelers to understand what kinds of content are or will be available, and what that content contains. User needs call for understanding why customers want the content and what kind of characteristics are important in content. Intended use in turn describes what kinds of characteristics are needed in the production, delivery, and use of content. If these issues are neglected, the resulting ontology may be insufficient to express the semantics of content essence. Interaction between these elements, the ontology, and the environment are illustrated in Figure 12 and discussed in more detail later in this chapter, as well as in the publications. ***Environment*** in the following figure represents changes that must be reflected in the ontology.

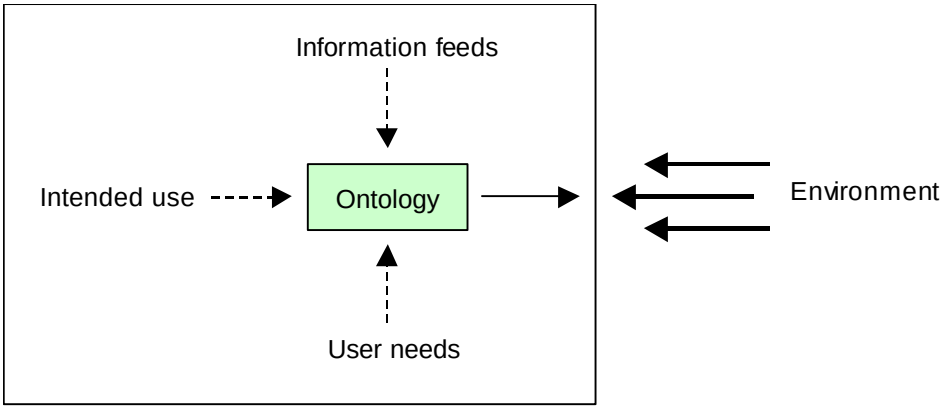

Figure 12. Main elements of domain modeling.

[39] Discussion with Dr. Anatole Gershman, Partner, Accenture Technology Labs, August 1997.

4.4.1 Information feeds

Domain modelers must understand not only what kinds of content essence exist or will be available, but also what kind of information, format, and other characteristics the content essence has. Similarly, if the incoming content essence already contains some kind of metadata, understanding the information feeds helps to estimate the reusability of metadata. Since both the available information feeds and the content domains themselves are often changing rapidly, the possibilities to identify new concepts and bind them to content essence, or to redefine the ontologies altogether, has to be supported [Wiederhold, 1995].

Understanding the information feeds covers also the ***genre*** of content essence, i.e. whether the content essence is fact-based and objective, or if it is subjective and/or interpreted. Genre has an important impact on domain modeling. If the content essence consists of more interpretative elements such as poetry or storytelling, its semantic metadata is likely to be indirect and much more difficult to formalize and extract.

The format of content essence affects the ability to use automated methods in the creation of metadata. Textual content essence is relatively easy to process with existing methods, but the successful creation of semantic metadata for graphics, audio-, and video-formatted content essence requires a combination of advanced methods and expensive manual work.

Another factor impacting domain modeling is granularity of the content essence, especially with larger works such as books and magazines or other kinds of non-textual content essence, such as video or audio. The question in regards to granularity is how to express various levels of representation and generality with a single set of metadata, i.e. what is the suitable semantic metadata for a single paragraph or segment, chapter or program, or the complete work? If we simply combine all semantic metadata, we end up having a vast and overly complicated, detailed representation of semantic metadata at the top level, without identifying the most important and emphasized characteristics of the complete work. For example, with content essence in video format, it is important to understand both the semantics of the content essence as well as when one segment is over and another one has started. Furthermore, when these segments are combined to form a complete program or even a series of programs, authors must define the corresponding metadata at that level. Some aspects of granularity are discussed in [Heeman, 1992], [Rust, 1998], and [Kubala et al., 2000].

[Salton et al., 1994] have encountered the granularity issue during their experiments with a large encyclopedic article collection. Their research shows that the decomposition of long texts into components of content essence with variable levels of granularity, such as full text, text sections, and text paragraphs, increases the accuracy in retrieving the most relevant content essence. However, their approach is based on relatively static content domain and large amounts of similar content essence, which limits the usefulness of their results in regards to frequently changing content domains such as news. Additionally, statistical document word frequency based methods, used for example by Salton et al., suffer from the inability to use the structure of the content domain in the retrieval, which leads to the problems of words having different meanings in different contexts.

4.4.2 User needs

User needs answers the question of why the end-user needs the content essence. User needs consist of both explicit user requirements as well as implicit characteristics of the user. User needs impact the content domain in two different ways. Firstly, user needs define the rationale for producing a certain kind of content essence. Secondly, user needs and characteristics partially define the actual format that the content essence and semantic metadata should have. Some factors related to user needs that impact the ontology include the value of information to the user, the timeliness of the information, the nature of user need itself, and the user behavior pattern with a certain media platform.

Timeliness describes the importance of receiving the content essence as fast as possible. This quality has two different aspects. High emphasis on timeliness means that the media companies should optimize time-consuming steps in the process, such as human-assisted creation of semantic metadata. On the other hand, delivery of content might be performed faster if semantic metadata were available during production and processing. In tasks related to filtering the content essence, timeliness is often of overriding significance. This is not often the case, however, due to ad-hoc needs of information during information retrieval [Belkin and Croft, 1992].

The characteristics of user needs describe how mature and stable can we assume the user needs to be. If the user needs change frequently or on an ad hoc basis, semantic metadata should concentrate on providing good searching capabilities. If the user needs reflect more long-term interests, semantic metadata should concentrate on assisting automated filtering of information.

Although this work emphasizes the importance of keeping the content essence and its representation separate for as long as possible during its production, it is important to understand how people behave with a certain media platform and what kind of requirements their behavior sets to the resulting products and content essence. For example, with the advent of Digital TVs users are expected to continue with behaviors learned with traditional TVs including the art of *channel surfing*. If we allow users to enter and exit the broadcast at any time, we cannot expect the applications running on Digital TVs to have previously broadcasted metadata available, leading to the periodic retransmission of required metadata [Chernock et al., 1999].

4.4.3 Intended use

In addition to available information feeds and user needs, the intended use of the content essence is the third factor affecting domain modeling. Intended use in this work covers not only the existing and planned end-user applications and services, but also all content-related activities in the content value chain. The difference between user needs and intended use is that user needs describe why users want the content, whereas intended use consists of the user context and methods how these needs are fulfilled. This work discusses the role of two activities, ***ontology visualization*** and the automated creation of metadata. More discussion on the aspects affecting intended use is available in the publications, especially in publication 4.

Ontology visualization may have a major impact on the ontology. If humans are not directly using the ontology, its structure can be complex. Anthropological linguists have come to the conclusion that

taxonomic hierarchies that exist in every language rarely have more than five or six levels, and even this number is so uncommon that they mostly occur in small fragments. However, these limitations do not necessary apply to expert, technical vocabularies [Cruse, 2000]. It is therefore important to understand, who the users of the ontology are, if the ontology will be visible to the administrator and/or the user, and furthermore if it is hidden and used only by computers. If the ontology is visualized, the usability and browsability of its structure must be considered. For example, structures in the *Yahoo* categorization are designed by keeping the amount of parallel choices at minimum. If the number of choices grows over a certain limit, the categorization will be pruned[40].

The level of automation is another aspect affecting the ontology. If the creation of metadata is automated, the process can be very detailed and lead to complex semantic metadata. However, fully automated extraction of metadata without human control is not recommended, as the process is often very complicated and may lead to inferior quality metadata. The routine work involved in creating semantic metadata should be as automated as possible, but human experts should still have control of the overall creation of metadata. Even though I did not personally concentrate on developing methods for measuring the quality of metadata, it is a fundamental aspect of content and should be addressed in future research.

[Rappaport, 1997] stated a similar principle regarding co-operation between humans and computers:

> *"It is this interdependence between the parts and the whole that makes cognition a complex system. This circularity presents a fundamental challenge in the formalization of artificial systems. One answer to this challenge is to build systems that are extensions to the human mind, that is, enhance the value of its functional properties and contribute to a positive change in the emerging intelligent behavior."*

4.4.4 Dynamics of the content domain

In addition to previously introduced qualities, the dynamics of the content domain – the frequency of change in content essence, user needs, and intended use – all impact the ontology. Even if the ontology could be defined to be absolutely ideal at a certain moment in time, it will gradually degrade and require maintenance. The dynamics of the content domain and their relation to domain modeling is illustrated in Figure 13. In that figure domain modeling produces the ontology needed to create metadata during electronic publishing. When the environment changes during electronic publishing, those changes are reflected back to the ontology.

[40] Interview with Srinija Srinivasan, editor-in-chief and vice president, Yahoo.com, 25-March-1999.

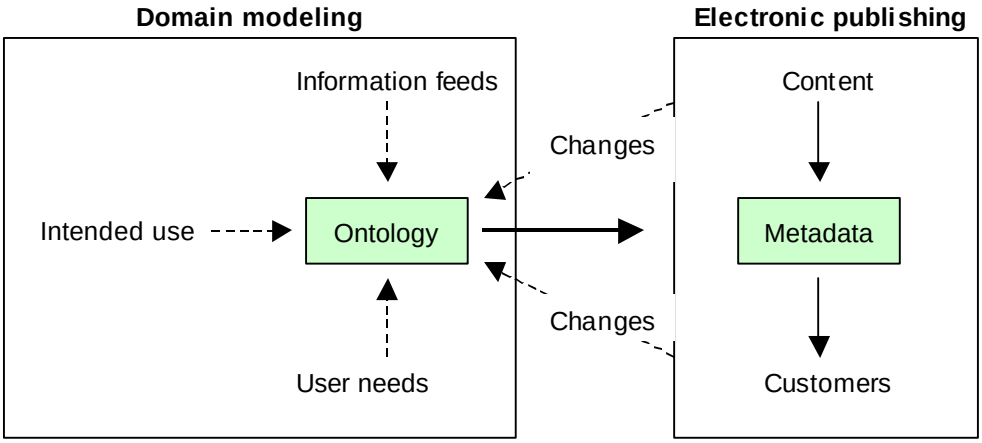

Figure 13. The impact of domain dynamics to domain modeling (adapted from publication 4).

Some content domains are more static than others. Content domains such as the news are often dynamic and lead to constant revisions of the ontology. The news content domain covers a wide range of topics, different users have different interpretations of the concepts, and new concepts surface frequently into the vocabulary. Media companies do not have a standardized vocabulary, and there are a minimal number of external influencers, such as legislation or standards bodies, that define and structure the content. Moreover, if media companies try to dictate the contents or vocabulary of (news) content essence, they might be accused of suppressing creativity and freedom of speech. More static content domains can be found in many established fields of science such as chemistry, where information is often based on a well-defined foundation such as the periodic table. In these fields, new content essence can often be described within the existing structures, or only require minimal changes in the ontology.

An example of how changes in the environment affect existing categorization was presented by one of the participating media companies in the SmartPush project. In the 1940's, the company saw an increase in news articles belonging to the category *odd phenomena*, but after the first manned space trip they had to create a new category for similar articles entitled *space travel*[41]. Similarly, outbreaks of diseases such as *AIDS* and *Bovine Spongiform Encephalopathy (BSE)*, also known as *Mad Cow Disease*, have introduced new vocabulary for news reporting.

4.5 Ontology mapping

The idea of having one common standard for semantic metadata sounds lucrative, but it is unfortunately extremely difficult to achieve given the plethora of media companies and other participants in the content value chain. We therefore need a method to make varied content and their semantic metadata compatible by defining rules with which one can homogenize metadata from multiple heterogeneous data sources. This process is called ***ontology mapping***.

Ontology mapping can take place at different levels of complexity. Ontology mapping may involve the translation of a used ontology into its semantically equivalent counterparts in other ontologies. For example,

[41] Interview with Markku Ylinen, Kustannus Oy Aamulehti, 1-November-2000.

a source ontology may use a concept named *car*, but a target ontology contains a concept called *automobile*. In this case the ontology mapping involves renaming *car* to its counterpart *automobile*. A more complex ontology mapping task occurs when concepts in source and target ontologies do not contain semantically equivalent concepts. For example, the source ontology contains a concept *competitor*, but the target ontology has only a concept *company*. In this case ontology mapping might require the use of multiple concepts, or even changes in the ontology, to express the semantics of the original concept.

Ontology mapping between semantically different concepts is a difficult problem that I have not researched in detail. [Rada and Carson, 1994] discusses some issues related to converting documents between different hypermedia systems and suggests that:

> *"Conversion success can be enhanced by limiting product use to only common features or by direct translation from one format to another without going through an intermediate standard format."*

These findings could be applied to mapping metadata between different ontologies as well, although common features based use is suitable only in limited cases and direct translation leads easily to a combinatory explosion of required translations. The more there are different end formats, the more advantageous it is to use reusable methods and a common ancestor format. For this method to work, all relevant parties must agree on a common ancestor ontology for semantic metadata.

In addition to different information feeds having their own ontologies, it is also quite normal that different stages in the content value chain have their own ontologies. An example of mapping different kinds of ontologies in the content value chain is given in the following Figure 14 and discussed more in the publication 4.

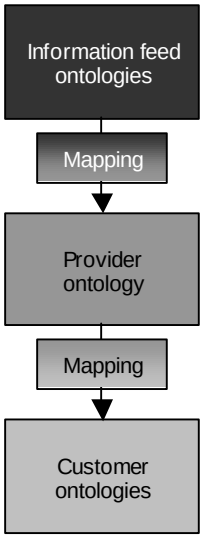

Figure 14. An example of the relations between different kinds of ontologies (adapted from publication 4).

4.6 Summary of findings

Semantic metadata describes the meaning of the content. By limiting the semantic metadata to a certain content domain, detail level, and usage, it is possible to define an ontology that captures the important semantics of the content domain.

Domain modeling is highly dependent on three elements: information feeds, user needs, and intended use. Information feeds represent the content essence for which the ontology will be used. User needs are connected to the user characteristics and the reasons why users acquire and use the content. Intended use describes how the content is, and will be, used.

The dynamic nature of the content domain must be taken into account when the ontology is developed and used. If any of the elements for domain modeling change, these changes must be reflected in the ontology.

Automation in domain modeling and the creation of semantic metadata is important, but to ensure the highest quality, human experts should still be in control of the creation of ontologies and semantic metadata. Only the fully routine creation of semantic metadata can be performed without human intervention.

Ontologies must be compatible if the media companies want to use semantic metadata originating from multiple sources. If the ontologies are not identical, participants in the content value chain must make the ontologies compatible by building methods for mapping individual elements from the source ontology, either directly or via some common intermediate ontology. This is, however, a difficult or even impossible task, especially if ontologies do not contain semantically equivalent concepts.

5 Electronic publishing process

I have previously discussed business aspects, semantic metadata, ontologies, and domain modeling. This chapter concentrates on creation and delivery of digital content and examines how the requirements of the content value chain can be taken into account in the electronic publishing process.

5.1 Background

Integrating metadata with publishing activities calls for co-existence and co-operation of artistic creativity and systematically managed electronic publishing. Electronic publishing covers activities in the content value chain, including authoring, content composition, and content delivery to the customer, whether it is the end consumer of the content, or simply the next possibly computerized step in the content value chain. These activities are discussed later in this chapter. Although research and development, which concentrates on developing advanced content-based products and ontologies, can be treated as its own process, it is closely interconnected to electronic publishing, and is discussed here as part of the electronic publishing process.

Individuals or communities are an important source for certain types of content essence, such as opinions, advice, and feedback. If and when media companies want to use content essence originating from individuals and communities, they must ensure that the content essence is accompanied with metadata that is required for further processing of the content. As sources outside the media company are not necessarily willing or capable of producing the required metadata, media companies must be prepared to augment incoming content essence with required metadata, as well as to monitor that the content is suitable to be used in the content value chain. In these cases, media companies operate as gatekeepers for content essence originating from external sources. However, extending working methods and control beyond the borders of media companies might be difficult. An example of the control issues is the recently halted web site annotation service, *Third Voice* [Maclachlan, 1999]. With *Third Voice*, users were able to add notes for others to see without asking permission from the content provider. This stirred a lot of debate on the control and justification of such services[42].

Electronic publishing, although having its roots in print publishing, has additional characteristics that differ from its traditional counterpart. The increasing demand and complexity of content-based products and services on multiple media platforms calls for semantic metadata, better content management, and tighter interaction between different participants in the content value chain. This in turn requires better understanding and standardization of the content and the ability to incorporate customer feedback and interaction into the electronic publishing process.

[Berghel, 1999] calls the new kind of publishing ***value-added publishing*** (VAP), which is seen as an extension of traditional publishing with the additional feature of linking and enhancing content with material from other publications, data sources, customers, and other networked media.

---

[42] See e.g. saynotothirdvoice.com as an example of these debates.

Content value chain and advanced content-based products require the electronic publishing to be seen more as a continuous production process than focusing on a single product or media platform. Furthermore, the separate, but interconnected roles of research and development and electronic publishing should be clear.

When media companies alter their publishing process to reuse content, it is important to understand the characteristics of different media platforms. A media platform is distinguished by various characteristics such as the sensory pathways through which they are perceived, the technology utilized on the media platforms, and the way the media platform changes over time [Rada and Carson, 1994]. The following Table 5, which is discussed more in detail in publication 3, contains a list of some typical media platforms and a description of their main characteristics. Although individual values in the table might be argued, the reason to include the table here is to highlight the unique combination of characteristics associated with each media platform.

Table 5. Typical end-user media platforms and their qualities (adapted from publication 3).

Characteristics	**A computer with Internet access**	**A computer with local mass media, e.g. CD-ROM, DVD**	**Mobile phone / wireless Internet devices**	**Digital TV with enhanced programming**	**Radio**	**Paper**
Network-dependence	Yes	No	Yes (wireless)	Yes	Yes	No
Interactivity	Yes	Yes	Yes	Limited	No	No
Personalization	Yes	Limited	Yes	Limited	No	No
Forms an interacting community	Yes	No	Yes	Limited	No	Limited
Real-time audio	Yes	Yes	Yes (voice)	Yes	Yes	No
Real-time video and animation	Limited	Yes	No	Yes	No	No
Updateable content	Yes	No	Yes	Yes	Yes	No
Location-awareness	No	No	Yes	No	Limited	Limited
Costs related to consumption	Free to medium	Free	Medium-high	Free to medium (return channel)	Free	Free
Portability	No/yes	Yes	Yes	No	Yes	Yes
Typical session	Interactive pull/push	Interactive pull	Interactive pull	Mostly one-way push	One-way push	One-way push
Local storage	Yes	Yes	Limited	Limited	No	Limited

Specific characteristics of each supported media platform impact product design and production. For example, Digital TV implementations have at least initially only moderate support for two-way interaction due to their limited capabilities for return channel. However, in the short term limited interaction is acceptable, if the goal of digital TV is to replace existing TV-tradition with enhanced TV experience. This means that people use TV mostly for consuming passive, broadcasted content essence aimed at a large

group of people [Chernock et al., 1999]. Likewise, ***portability*** restricts the size and capability of the media platform, but on the other hand makes certain kinds of characteristics, such as ***location-awareness***, very useful [Oinas-Kukkonen, 1999].

5.2 Digital content versus print publishing and broadcasting

Although this work has many links to ***traditional publishing*** (see e.g. [Enlund, 1994], [Enlund and Maeght, 1995] for examples and issues in traditional publishing), certain aspects of traditional publishing are not necessarily valid with digital content. The absence of physical product restrictions and reduced importance of deadlines are two examples of these differences. Costs and time incurred in producing an additional copy of digital content are minimal compared to producing a new copy of a physical product such as a printed newspaper. As the actual setup, output, and delivery of digital content is fast, it does not necessarily have physical restrictions, and can be automated to a great degree, these processes can be reproduced partially or fully without major effort and costs. This in turn allows the content to be constantly revised and reduces the importance of deadlines.

The nature of electronic publishing is also different from broadcasting via television or radio. Broadcasting is based on a one-size-fits-all approach, where the possibilities to interact are typically limited to switching channels. With the electronic publishing possibilities to interact and collaborate are much broader. With digital content the user typically controls what to see or do next and when this change is going to take place.

It seems that one of the closest relatives to electronic publishing is textual information that broadcasters have been using as an addition to TV broadcasting. ***Text-TV*** is used mostly in European television systems and is based on textual pages that are transmitted within the standard broadcasting signal [Schneider, 1994]. Pages in Text-TV are refreshed constantly and the user decides which page to see next by entering its page number via remote controller or similar device. Although the Text-TV standard has many limitations such as slow speed, very limited bandwidth, poor interactivity, and the absence of multimedia content essence, it is still a good example for the future production of content and will reincarnate and improve in the future with the digital-TV standard [Rohde, 1999]. Text-TV has been successfully used especially for reporting breaking news and then revising the news story as the time goes by and more details become available. This method has been used also by some TV news stations, which have a continuous coverage of breaking news stories. However, because TV channels have their regular programming schedules to follow, their reaction ability is limited and they are still dependent on deadlines.

Regardless of the success of digital content, it is obvious that paper-based publications will exist for a long time to come. [Dodd, 1990] discusses the advantages and disadvantages of paper-based and electronic journals and suggests a hybrid that would combine the best qualities of both worlds. A paper version would be regarded as the definitive, easily accessible, basic version of the content essence, and the electronic counterpart would be used during authoring and for searching, cross-referencing, augmentation, collaboration, and discussions. In a longer time period, however, the advantages of paper are likely to decrease, enabling digital content to overtake more and more of the roles of paper-based content.

5.3 Reusing content on multiple media products and media platforms

One of the main advantages of metadata is the possibility to reuse content. A number of factors affect the possibilities and rationale to reuse content. These factors include the characteristics of content, user needs, the number and publishing frequency of derived products, costs, revenues, resources, rights, available skills needed for the reusability, as well as the existing publishing process and the capabilities to alter that process (see e.g. [Sabelström Möller, 2001] for further discussion on characterization of content essence and integrating the publishing of printed and electronic newspapers).

In the ideal situation, content and its presentation are kept separate as long as possible in the content value chain with methods such as ***style sheets*** [Bos, 2000]. The further the production of content proceeds without designating it for a certain media product or media platform, the less work is normally needed to produce different media products based on the same content essence. Although I have emphasized the separation of content and presentation in this work, media products and media platforms have different characteristics that should be reflected in the development of media products and the production of content, and that both these characteristics affect content reusability. For example, computer games often have high-level performance requirements as well as a highly interactive and specialized user interface, which might require content to be tightly integrated with the media platform and its presentation. In these cases it might be difficult or even impossible to reuse the content on other media platforms and media products. [Rada and Carson, 1994] describe a similar observation:

> *"As long as platforms continue to be distinguished by differing abilities to process various media, there will continue to be good technical reasons for having native media formats that are closely matched to platform-dependent interfaces."*

Despite these challenges, separation of content and presentation is possible for many kinds of media products and should be considered as a goal when designing different aspects of the electronic publishing process.

5.4 Four layers of electronic publishing

Publication 3 introduces a four-layer framework, the component model, for understanding and managing different elements of the electronic publishing. These four layers help to express the complexity of reusing the same content essence in multiple media products and media platforms. Figure 15 contains a simple illustration of the framework and how it can be used to describe the relations between different layers of the framework.

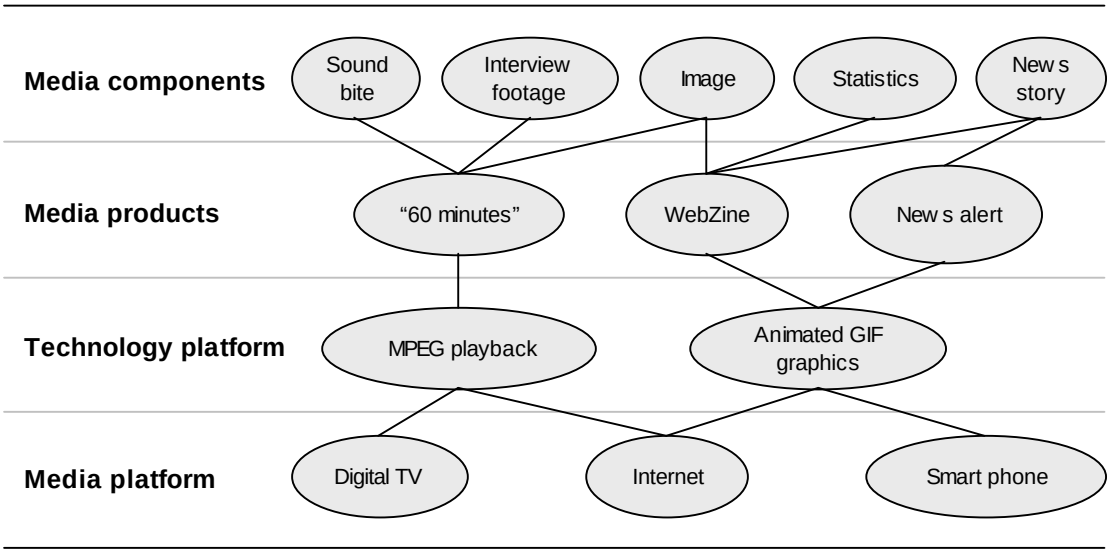

Figure 15. A possible example using four-layer framework for electronic publishing.

The first layer is the media platform, on which the content is delivered and accessed by the user. Typical examples of media platforms include end-user terminals and other possible platforms such as personal digital assistants, PDAs. The second layer is technology platform. It consists of the technology that enables and implements the functionality provided on the media platform, such as different alternatives to view pictures. In some cases the technology platform contains alternative technologies to perform the required functionality. The technology platform may also be dynamic so that the available functionality can be expanded or modified by downloading new software or adding new hardware to the end-user terminal. According to my experiences, media companies often treat the technology platform as an inseparable and invisible part of the media platform, which should be changed to better enable the reuse of content on other media platforms and media products.

The third level in the framework is media products that define how the content essence is selected and integrated into a deliverable media product. The last layer is media components consisting of both the content essence and its metadata. Media components form the atomic pieces of content that are used in media products. The relation between media components and media products is often vague. A single media component may be integrated with others to form a more complex media product, but it may also be delivered individually, in which case a media component may also treated as a media product per se.

5.5 Publishing process steps

The following discussion describes the process model. The process model explains the major steps of producing flexible content, i.e. content essence that can be reused for different purposes. Even though media companies do not necessarily model their day-to-day processes according to the presented process model, it assists in understanding the different steps and their relationship to the electronic publishing process.

I have divided the publishing of digital content into two processes: research and development and electronic publishing. Electronic publishing is further divided into three process steps, authoring, content composition, and content delivery. Different publishing activities are introduced later in this chapter and discussed in detail in the publications.

The following Figure 16 illustrates the main processes and steps for publishing digital content as well as the interactivity of publishing. In the figure squares represent the outcome of research and development and ovals represent the process steps of electronic publishing.

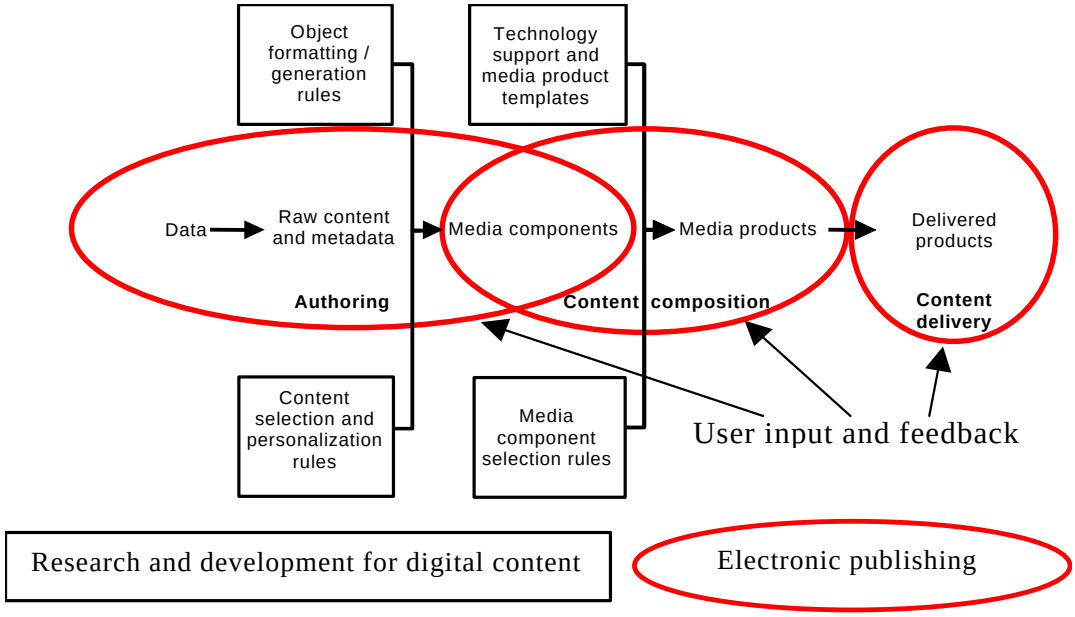

Figure 16. Key processes and process steps for publishing digital content (adapted from publication 3).

Unfortunately the processes used by the media companies today do not typically resemble the described situation, especially the creation and use of metadata as well as reuse of content are insufficiently supported. Figure 17 describes a top-level process flow that one of our partners in the SmartPush project used to produce electronic versions of a daily paper-based newspaper. This process could be optimized in a number of ways, such as integrating the authoring of the paper and electronic versions or by moving the indexing to be part of the authoring. This would have reduced costs and allowed the production of a much wider variety of advanced content-based products.

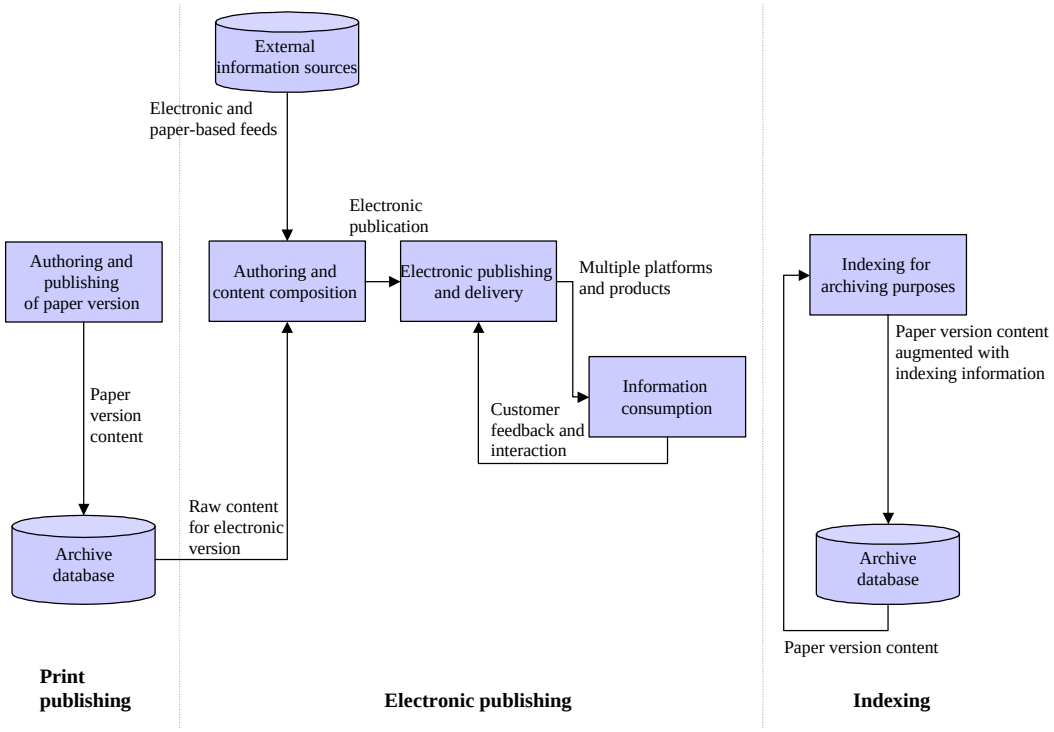

Figure 17. Exemplary electronic publishing process.

[Maybury, 2000] has identified similar process steps in the news customization that he calls *personalcasts*. He divides the news production into capturing, processing, and presenting, but does not identify the role of research and development as part of publishing activities. In the capturing phase, the content essence is digitized and stored. The processing phase consists of segmentation, transcription, translation, indexing, correlating, clustering, fusing, extraction, and summarization. The last phase, presentation, includes selection, organization, co-ordination, tailoring, and visualizing the content essence. Due to the amount of skills and steps required for these operations, Maybury sees news production as necessarily multidisciplinary, typically requiring organizationally and geographically dispersed teams.

5.5.1 Research and development

The research and development process, which precedes the actual production of content during electronic publishing, contains activities that are required to define and maintain different elements used in electronic publishing. These elements include the ontology, templates for media components, templates for media products, and the technology platform that is required to produce, deliver, and use content.

If the media product is ***unique*** – a substantial part of it is created for a particular purpose – there is no need to formalize the structure of the media product. If the media product is published multiple times with some variation in its structure, or if the same content is used in multiple media products, it is often advisable to formalize required characteristics into reusable ***media component templates*** and ***media product templates***. These allow better reuse and automation during electronic publishing. Media component templates define rules for producing media components from content essence. Media product templates define the properties

of media products, such as their selection and personalization criteria, presentation, supported functionality, and pricing information.

Some of the activities performed during research and development closely resemble the process steps of electronic publishing, but their skill and resource requirements are different. For example, the creation and modification of media product templates and media product templates usually takes place periodically and requires project-like effort with technology platform and media platform -related skills, whereas electronic publishing is an on-going effort relying more on domain expertise, creativity, and journalistic skills. Although these differences exist, successful electronic publishing requires close interaction between research and development and electronic publishing. In this way we can ensure that the quality of resulting media products remains high.

A major input for the research and development process should come from the customers. Their needs should be reflected in the design of media products. As substantial portions of content are usually subject to change, the content should be continually exposed and verified against changing user needs. This same fact has been identified in regards to knowledge by [Fahey and Prusak, 1998].

5.5.2 Authoring

Authoring is the first phase of the electronic publishing activities. It consists of steps that involve creating, acquiring and editing of the content essence together with its metadata. Once the authoring is completed, content essence should be structured and augmented with metadata so that computers are able to process the content even without human intervention, for example by selecting suitable content to be published in an online newspaper.

According to my experiences, it is important to integrate the creation of semantic metadata with the authoring, because the author is the only person, who truly knows what the content essence is about and what aspects should be emphasized explicitly in the metadata. Without close interaction between the person creating metadata and the source of content essence as well as without experience, clearly stated instructions and tool support, the interpretation of the content essence and quality of metadata may vary greatly between different persons creating metadata [Hirschman et al., 1999]. Likewise, if the creation of metadata is postponed until later in the process or performed automatically without human control, the quality of metadata, especially semantic metadata, is likely to suffer.

Authoring may take place in real-time or be based on some kind of ***publishing schedule***, which defines the times and deadlines when media products are published. Authoring produces media components that are more or less ***platform-independent***. The separation of media components from their presentation and use enables their reusability on multiple media products and media platforms.

5.5.3 Content composition

During content composition, suitable media components are selected and integrated into media products. After content composition, media products are ready for delivery to the customer. Content composition consists of the selection of content, layout formatting, personalization, and publishing. In addition to media

components, content composition requires structural information on media products usually in the form of media product templates.

The outcome of content composition, published media products, can be more or less independent of the media platform. Personalizing the presentation of a media product typically restricts the use of media products on multiple media platforms. For example, a user may request news headlines to be presented in a temporal presentation without user interaction, which would lead to tight coupling between content, presentation, and functionality.

Content composition requires a varying amount of work. In the simplest case, content composition consists of possibly automated steps of selecting, modifying and combining media components using an existing media product template. For example, a daily online newspaper may use the same media product template from day to day with the individual stories, media components, changing for each issue, which are selected based on their semantic metadata.

A more complex situation occurs when the existing media product template is not sufficient for the task and must be modified to contain, for example, a new section on domestic politics. This change requires both modification in the media product template and the availability of media components needed for the new section.

The most complex situation arises from the need to change and modify the technology platform. For example, an online newspaper is modified to have interactivity by enabling users' voice annotations to individual articles. To support the new feature, the media company has to build both the production system that supports interactive newspaper and ensures that the customers' terminals are able to support the new functionality.

5.5.4 Content delivery

The last step in the electronic publishing is content delivery, where the user gains access to the published media product. The customer may not be the end-user, but simply a next step in the content value chain such as a computer performing additional processing on media products. The content delivery process step covers both push- and pull types of delivery. In ***push delivery*** a media company determines when the media product is delivered, whereas in ***pull delivery*** the customer initiates the content delivery, for example, by retrieving the media product directly from the media company.

Content delivery can also be personalized by defining the preferences of the content delivery, such as the category of media products to be delivered and the time when the content delivery should take place.

5.6 Summary of findings

Some characteristics of electronic publishing are derived from print publishing and broadcasting; some of its qualities are unique. For example, interactivity, collaboration, customization, and the role of metadata as an implicit part of the authoring do not exist in print publishing.

Two models, a component model and a process model, assist in understanding electronic publishing of multiple media products on multiple media platforms. Even though these models are relatively abstract and

as such not necessarily directly applicable to day-to-day operations in media companies, they help to understand and analyze publishing processes and companies participating in the content value chain.

Media products and media platforms have both common and unique characteristics that must be supported in the content value chain. The component model helps to understand the elements and their dependencies in electronic publishing. The component model divides electronic publishing into four layers: media platform, technology platform, media product, and media component. The process model decomposes activities into two processes: research and development and electronic publishing. Electronic publishing is in my thesis further divided into process steps consisting of authoring, content composition, and content delivery. Process model describes the dependencies between the steps and allows for the analysis and optimization of the given electronic publishing process.

These two models allow easier understanding of relevant issues such as reusability and flexibility with their practical implications. Furthermore, by understanding the similarities, differences, and implementation alternatives of electronic publishing process, media companies are much better prepared to optimize and modify their management of content.

The key recommendations to efficiently publish content on multiple media products and media platforms are:

- Use a process approach for electronic publishing;
- Minimize manual work while retaining the control on human experts during electronic publishing;
- Use identified and separate roles for research and development and electronic publishing; and
- Keep content reusable by formalizing media components and media products into reusable media component templates and media product templates.

6 Advanced content-based products

This chapter outlines some generic characteristics, possibilities, and limitations of applications and services that make use of semantic metadata, flexible content, and the electronic publishing process. The main emphasis is on the SmartPush project, in which the author has been actively involved as a researcher and a project manager. This chapter aims to give the reader an idea of the possibilities that are enabled by semantic metadata, ontologies, and the electronic publishing process.

6.1 Overview of the advanced content-based products

In this work, advanced content-based products consist of applications and services that process and refine content using computers and automation. These products typically require semantic metadata before they are able to perform the desired functionality. Advanced content-based products often offer more functionality to the end-user than mere presentation, such as the reading, listening, or viewing of content essence. Examples of advanced functionality include personalization based on user preferences, dynamic formatting based on metadata, conditional actions based on content essence, and interaction with the user. However, processing and refinement of content might also take place earlier in the content value chain, so that the end products do not contain any advanced functionality. For example, many media companies today process content in structural formats such as XML, but when they produce media products, they quite often are not willing to release the XML-formatted content and offer the customers only the content essence without its metadata.

Many of the popular paper-based publications are based on the augmentation of existing content essence. Travel guides, television-, music-, and movie reviews all augment existing information and produce value to the user by providing extra information and reformatting existing material, as well as by enforcing and validating existing opinions. All of this is also possible and available with digital content. However, advanced content-based products are able to do more than just presenting a static set of additional information. The applications and services can, for example, filter and recommend content to manage information overload, or reformat content based on customer interaction and current context. They can also link, augment, and integrate content with other content, services, and even knowledge related processes in the organizations.

Advanced content-based products are still very much underutilized in media companies. [Palmer and Eriksen, 1999] studied 48 online newspapers around the globe. Although their research data was collected already in 1997, some of the identified trends are clearly valid with electronic publishing today. For example, advertising was the main source of revenue for most of the newspapers and only eight of them had a payment scheme in place for subscribers. These newspapers were either very specialized in their content domain or had a strong geographical focus. Eight out of 48 papers generated additional revenue through customized media products, and only eight of the 37 newspapers having an archive charged for its use. Newspapers offering fee-based searches had substantial research departments and huge paper-based archives. In most cases, newspapers had simply ported their existing paper versions to the World Wide Web

and they seemed to compete with each other through technological advances, not through the quality of their content or services.

6.1.1 Information filtering and information retrieval

Many of the advanced content-based products use some form of ***information retrieval*** (IR) or ***information filtering*** (IF) (see e.g. [Foltz and Dumais, 1992], [Salton et al., 1994], [Chen et al., 1994], [Kendall and Kendall, 1999]). IR and IF methods compare content against some other aspect, such as other content, query, or ***user profile*** enabling for example personalized information feeds. The following Table 6 compares some common characteristics between information filtering and information retrieval tasks.

Table 6. Issues related to information filtering and information retrieval (partially adapted from [Belkin and Croft, 1992]).

Issue	**Information Filtering**	**Information Retrieval**
Content type	Novel, dynamic	Existing, static
Timeliness of content	Important	Less important
Interest focus	New information	All available information ranked by given criteria
Interest is typically expressed with	User profiles	Queries
Type of information need	Long-term, regular, or periodic	Ad-hoc, immediate
Type of interaction	Passive, often automated	User is active, initiates operation
Selection mechanism	Adaptation to implicit or explicit feedback	Query refinement
Initiator	Sender	User
Target	Manages information overflow by removing non-relevant information, rank results *Show only the essentials*	Finds new information by expanding the information base, rank results *Show what is available*
The need for end user actions and judgment	Low, higher with explicit feedback	High
Human intelligence essential in	Authoring	Both authoring and selection
Prerequisites	Available user profile	Ability to express information needs explicitly
User motivation	Varies	High

Based on my experience, the main challenges in information filtering are related to representing the user with an accurate user profile and adapting the profile to the changing user needs as well as matching and ranking the content based on user profiles. With information retrieval, the main challenges are in interpreting the user query, query refinement, in query execution performance, and ranking the results. Additionally, information retrieval may be highly dependent on human intelligence. For simple and routine operations, such as reacting to a *no results* response with a query refinement using synonyms of used search terms, computers perform even better than humans. However, if the results require analysis and interpretation, humans outperform computers. An example of the latter kind of tasks is the decision, whether a response to a query satisfied diverse needs or if the query should be reformulated or redirected to other sources of information.

An alternative to using humans in managing information is to use ***information extraction*** (IE) [Cowie and Lehnert, 1996]. Information extraction is closely related to information retrieval and information filtering. Information extraction uses mostly natural language processing methods and aims at finding and linking relevant information while ignoring extraneous and irrelevant information. In addition to text finding and retrieval, information extraction distills information from the content essence by using a coherent network of relevant aspects of the content domain, i.e. an ontology. Although some academic and commercial applications exist such as FRUMP for news stories [DeJong, 1982], JASPER for extracting information from corporate earnings reports [Davies et al., 1996], and SCISOR prototype for analyzing articles on corporate mergers and acquisitions [Jacobs and Rau, 1990], a number of unsolved questions still remain. These include the quality of extracted information as well as the estimation and justification of costs and the effort needed to build and maintain such systems [Cowie and Lehnert, 1996]. As tools for creating metadata in the SmartPush project have shown, information extraction systems can be useful in extracting an initial semantic metadata for the advanced content-based products, but if the accuracy and high quality of results is important, fully automated information extraction might be insufficient as the sole method to produce semantic metadata.

The ontology used to describe the content essence does not necessarily have to be very complex. If user characteristics are difficult to capture, or if the production of metadata requires a lot of human effort that is not cost-justified, it might be better to use a simple ontology and leave further processing and analysis of content essence to the end-user and/or to information extraction methods.

[Belkin and Croft, 1992] have shown that even simple semantic metadata based on keywords, combined with appropriate methods for information retrieval, is surprisingly effective, efficient, and straightforward to implement. An extension of this process includes special-purpose processing for certain categories of content essence, such as names, organizations, and locations.

6.1.2 Example: Recommendation systems

Recommendation systems are examples of systems that track user interests and recommend content based on given criteria. These systems can be divided into two categories, ***content-based recommendation systems***, such as ours in the SmartPush project or InfoFinder [Krulwich and Burkey, 1997] and ***collaborative recommendation systems***, such as FireFly or GroupLens [Konstan et al, 1997], [Good et al., 1999].

Content-based recommendation systems maintain a user profile – a representation of the qualities of the content that the user has liked or disliked in the past – and compare that user profile with features of different alternatives in order to determine which content to recommend. In collaborative recommendation systems, content is not analyzed. Instead, the recommendations are based on similar tastes, likes and dislikes, between users.

Although content-based recommendation systems can be invaluable assistants, their recommendations have a number of shortcomings. According to [Balabanovic and Shoham, 1997], these shortcomings include capturing the relevant qualities of content essence in explicit representations, the need to have an existing

user profile, and the dependency on user feedback in building user profiles. Content-based recommendation systems are also prone to ***over-specialization***, in which users see only content essence that they have seen before. Over-specialization is sometimes tackled by injecting a certain amount of ***serendipity***, randomness, into recommendations, which allows the user to discover new categories of content.

Collaborative recommendation systems are not dependent on the characteristics of the content, so they can recommend entities that are hard to describe, formalize, or have contradicting interpretations. Examples of challenging content domains are music, wines, and movies (see e.g. Movie Critic[43]). However, collaborative systems also have shortcomings, such as the need to have a critical amount of users and user information before the system is able to create proper recommendations [Konstan et al., 1997]. If the user has an unusual taste or the system has a vast amount of alternatives compared to the number of users, the recommendations are likely to be inaccurate.

Collaborative recommendation systems are also not able to merge and manage recommendations that span different variations of the same content essence. For a collaborative recommendation system, a weather page in a certain newspaper does not have any special relevance to weather in other sources. Another shortcoming is the inability to process new or controversial items. Before the system can recommend an item, it must be ***bootstrapped***, i.e. enough people with similar interests must have seen and liked or disliked the content. This complicates and slows down the recommendation of new documents, making the system less attractive to content domains in which the timeliness of content essence is important.

A solution for the problems impacting both content and collaborative recommendation systems could be to combine both of these approaches. [Balabanovic and Shoham, 1997] have developed a system to recommend web pages based on a combination of content-based and collaborative recommendations. Their system, *Fab*, seems to solve most of the problems of a single system, such as the need to have extensive feedback or related scalability problems, but the accuracy of its ontology is still a major concern that cannot be bypassed.

Although we discussed using collaborative methods for recommendations, our prototype implementation in the SmartPush project was a content-based recommendation system. We used a number of methods to tackle the four challenges presented by Balabanovic and Shoham, which are capturing relevant qualities of content essence, the need for an existing user profile, dependency on user feedback, and over-specialization.

For capturing relevant qualities, we used domain modeling, semi-automated information extraction tools, and process modifications. A user profile was created by using existing information on users, based on actual interaction with the content essence, as well as by using the same structure for user profiles and semantic metadata that describes the content essence. The collection of user feedback was built into the user interfaces. Overspecialization was avoided by simplifying ontologies with dimensions and by designing the user interface and service in such a way that end users were not restricted solely to our recommendations.

The SmartPush project is discussed extensively in the publications and later in this chapter. Publication 1 as well as [Savia, 1999] describe in detail the personalization mechanisms used in the SmartPush project.

---

[43] www.moviecritic.com

6.2 Issues affecting media companies and advanced content-based products

The following discussion introduces some issues that might surface when media companies begin developing advanced content-based products.

Researchers at the Center for the Study of Language and Information (CSLI) have conducted a number of studies in relation to how people treat computers, television, and new media [Reeves and Nass, 1996]. CSLI research has highlighted the need to understand the context of communication and to minimize and focus the information in the interaction between humans and computers. According to their research, all communication with computers should be guided with four basic principles: ***quality***, ***quantity***, ***relevance***, and ***clarity***.

Quality requires the computer to be truthful in the communication. If the computer is found to lie or if it makes a mistake, the content essence presented by the computer is considered less reliable. The second principle, quantity, requires each participant in the communication to provide only as much information as the situation demands. Relevance requires that all information should clearly relate to the purpose of the communication. Quantity and relevance are often violated by information overload, even with the advanced content-based products, where media companies dump an excessive amount of less important or irrelevant information to the user. An example mentioned in [Reeves and Nass, 1996] comes from the early days of news broadcasting, where media companies assumed that people watched news to be informed, not to be entertained or to be a socially active member of their community. The last point, clarity, calls for a common vocabulary, an ontology, as well as clear and precise communication without ambiguity.

[Porter et al., 2000] discusses the development of customized information products in regards to technology management. They have developed a suite of tools that mine textual information obtained from large databases and then package the resulting information in multiple formats to be used on multiple media platforms. Porter et al. have made a couple of observations that are relevant to this work. According to their experience, users need the right information in the right format and on a timely basis. Unfortunately, user requirements are often complex and come in many different forms that take time and resources to extract from the information sources. Users also have a difficult time accepting the presented information as credible. This distrust is due to the missing human expert during extraction and a lack of understanding, or trust, in the extraction mechanisms. Human participation is crucial for acceptance, but because the task of analyzing and processing content essence entails unfamiliar skills and significant time, people capable of extracting knowledge are rare.

The first issue, producing the right content essence in the right format and on a timely basis, calls for an understanding of the content domain, user needs, structured content, and improved processes that would computerize, automate and speed up the flow of content through the content value chain.

The second issue, credibility and distrust could be tackled by assuring the users of the high quality of content essence and semantic metadata as well as by branding. By integrating the creation of metadata with the authoring, the media company could improve timeliness and ensure that the expertise of human experts

is reflected to the semantic metadata. Constant exposure to the production of semantic metadata could also result in higher quality and more uniform usage of semantic metadata by different authors.

[Berghel, 1999] lists qualities that enhance content-based products and add value for each of the participants in the content value chain. Some of the interesting and potential areas for advanced content-based products include:

- **Confidence indicators**. Includes document status indicators and external recognition of the content essence, such as awards, reviews, peer reviews, and perceived quality indicators.
- **Recommendation systems**. Includes for example information brokers, community review systems, and helper agents.
- **Interactive and participative systems**. Consists of dynamic, real-time document categorization and visualization, data mining, data warehousing, as well as communication between users.
- **Document persistence**. Includes post-hoc data utilization, version control, validation and archiving, variable-link-strength technology, frequency of access and average visitor ratings for a site, and detection of the number of referring hyperlinks.
- **Information customization**. Client-side document extraction, non-prescriptive and non-linear document traversal, open distributed archiving, security enhancements, watermarking and digital steganography, push technology, citation tree construction, and agent-based citation locators.

6.3 The SmartPush project

This chapter introduces and summarizes the SmartPush project from the point of view of a media company operating with semantic metadata. More detailed information of the SmartPush project is available in the publications. Publication 1 discusses some of the early steps of the project, as well as the methods used in the personalization. Publication 2 discusses the technology and implementation of the SmartPush project, and publication 5 discusses the project activities related to metadata.

Most of my empirical work was conducted during the SmartPush project in close collaboration with a number of industrial partners. Most influential of the industrial partners regarding this work were media companies *Kauppalehti*, part of a Finnish media conglomerate, *Alma Media Corporation*, and *WSOY Uudet Mediat*, part of a Finnish media conglomerate, *SanomaWSOY Corporation*.

Kauppalehti is a daily newspaper concentrating on financial news in Finland and abroad. We collaborated mostly with the online group of *Kauppalehti*. This group is responsible for not only the online version of *Kauppalehti*, but also selling the same content essence in a variety of packages for large corporations and as raw material for other media products. *Kauppalehti Online* is an enhanced version of the printed newspaper containing, among other things, a section for breaking news. This section, which contains short news articles originating from internal and external sources, was the test bench for the creation of ontologies and production of semantic metadata during our final pilot described later in this chapter.

WSOY Uudet Mediat represents a different kind of electronic publishing. *WSOY Uudet Mediat* produces a variety of advanced media products, such as Encyclopedia CD-ROMs. Whereas most of the content essence

at *Kauppalehti Online* is created and published fresh on a daily basis, content essence at *WSOY Uudet Mediat* is created over a longer period of time, categorized, and stored in a database for later use. Subsets of the available content essence are then selected and published periodically as media products. The main challenges at *WSOY Uudet Mediat* were in efficient managing of content and reusing it for different purposes. *WSOY Uudet Mediat* provided us with categorized content essence at the early stages of the SmartPush project. We also developed different categorizations for their content essence later in the project.

6.3.1 Background of the project

The SmartPush project was a three-year long research project that ended in the spring 2000. The SmartPush project was conducted at the Helsinki University of Technology, TAI Research Centre, in close collaboration with a number of industrial partners representing media companies, infrastructure manufacturers, telecommunications service providers, and software companies.

The foundations of the SmartPush project were based on a preceding research project called ***OtaOnline***, which concentrated on experimenting and researching the possibilities and challenges of the net media and electronic publishing [Saarela et al., 1997]. When OtaOnline ended in 1996, the SmartPush project took over the research effort.

The project had a number of focus areas in the content value chain, including the creation and management of semantic metadata, ***adaptive*** user profiles, i.e. user profiles that are updated according to changes in user interests, and ***metadata matching***, i.e. comparison and selection of content essence based on user profiles and semantic metadata, as well as agent-based implementation of the personalization system. The following statement, which is derived from the presentation materials of the SmartPush project, describes the goals and nature of the project:

> *"SmartPush is a content distribution environment based on automated targeting via metadata and agent technology. SmartPush provides better information quality and value-added services to both the user and professional content providers."*

The SmartPush project concentrated especially on the semantic aspects of the interaction between the media companies and the customer, although different media platforms are undoubtedly needed for conducting this interaction. The focus area of the SmartPush project within a ***content triangle*** consisting of content, media platforms, and users, is illustrated with the circle in Figure 18.

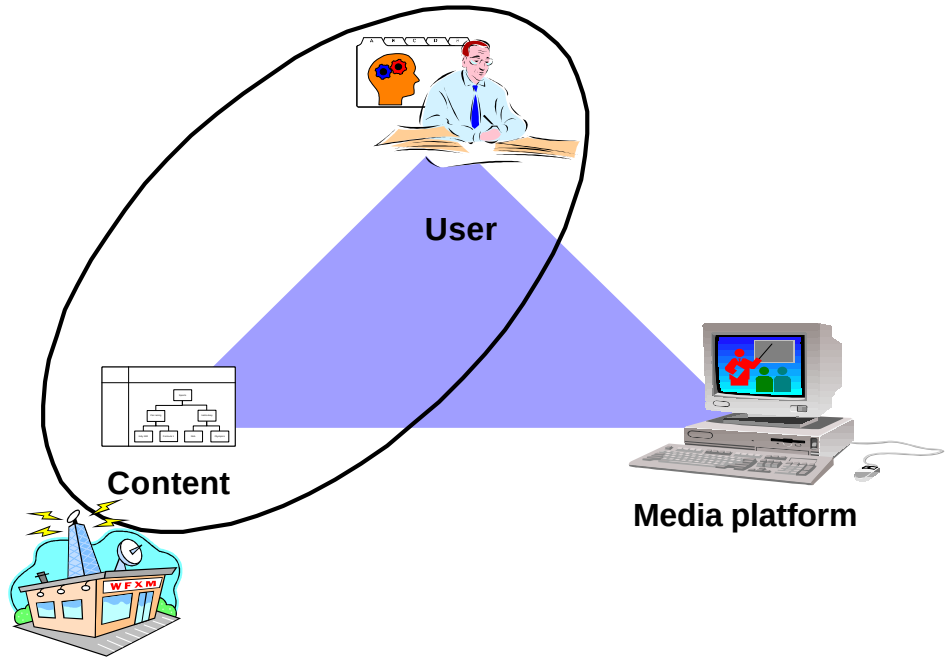

Figure 18. The content triangle and the focus area of the SmartPush project.

The goal of the SmartPush project was to be a proof of concept for personalized information feeds based on semantic metadata and user profiles. Alternatives identified in the SmartPush project for the personalized delivery of content are shown in Figure 19 and discussed more in publication 2.

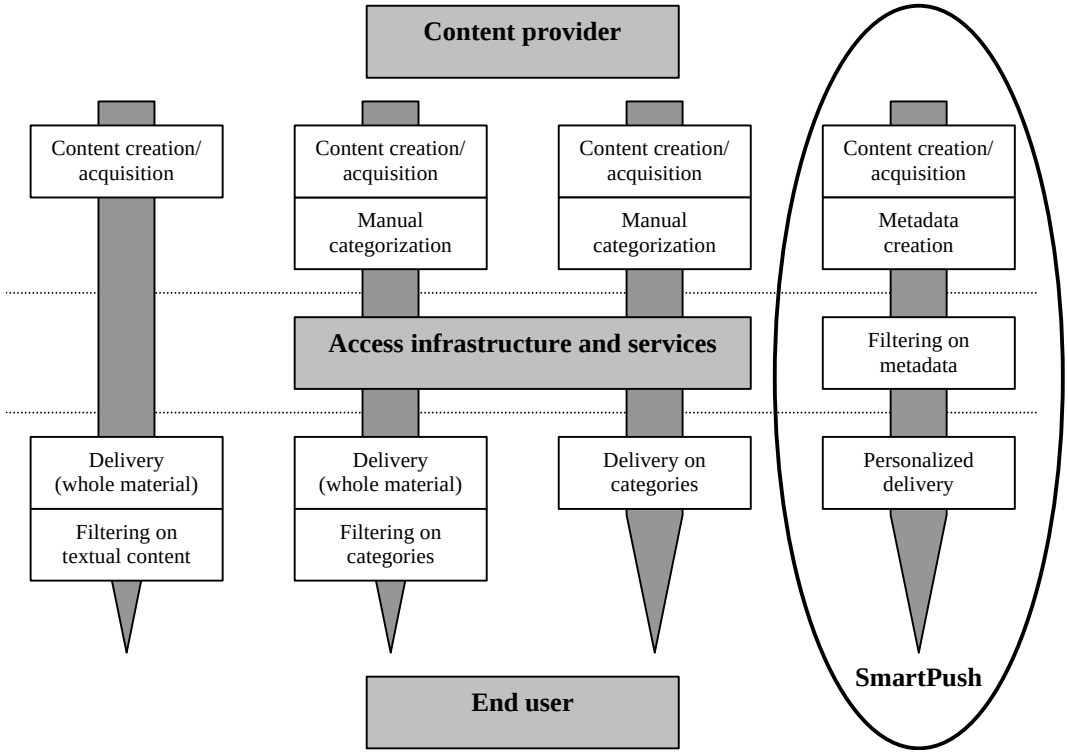

Figure 19. Different personalized content delivery alternatives (adapted from publication 2).

6.3.2 Ontology development in the SmartPush project

Personalization systems such as SmartPush rely on semantic metadata. If the content essence is not enhanced with semantic metadata, personalization of content must be achieved on an ad hoc basis with methods such as information extraction. As discussed before, methods for automatically extracting semantic metadata from content essence are extremely complex, especially for other than textual content essence. Even if automatic extraction is possible, the accuracy of the resulting semantic metadata is typically low. The SmartPush project therefore emphasized from the beginning the importance of integrating the creation of semantic metadata with the authoring of content essence. Also important is the need to have a human expert to accept and finalize the semantic metadata.

When the SmartPush project started in 1997, we assumed that suitable standardized ontologies and tools for describing semantics of content essence were readily available. However, further research in the field showed that such ontologies did not exist and that there were no commonly available tools for augmenting content essence with semantic metadata. The lack of suitable standards and methodology led us to devote substantial time to developing ontologies and tools for the content domains researched in the project.

We created a number of prototypes for the SmartPush project. We also needed a substantial amount of content essence and semantic metadata to test these prototypes, but we had to develop corresponding ontologies before the creation of semantic metadata was possible.

Most experiments during the SmartPush project were conducted with news content. Our first two prototypes covered generic news and the final one concentrated on the financial news. These content domains were selected mainly because the project partners had suitable expertise and content essence available for these content domains. The used news articles had a relatively short and simple textual format, they were typically objective, and they spanned a reasonably well-defined content domain. In most cases, the content essence was professionally authored so that the content essence would not suffer from misspellings, poorly formatted sentences, or varying capitalization.

News content is highly dependent on the timeliness of the content, which requires high throughput speed in the content value chain. Another challenge related to news content is its dynamic nature. New words are introduced constantly, and the meaning of existing words are altered or rendered obsolete impacting all operations that use semantic metadata, such as user profiles and recommendation systems [Balabanovic and Shoham, 1997]. Additionally, synonyms and meanings dependent on the content domain provide semantic challenges in the news content domain.

The first ontology for news was based on 319 Finnish newswire articles from 1995 and 1996 that had previously been categorized under 19 main categories. We selected samples from three main categories: *accidents*, *Finnish economy*, and *international economy*. To test the overall functionality of our prototype, this set was expanded with articles from other categories such as *nature*, *traffic*, and *sports*.

We created the first ontology by modifying an existing *Common Subject Categorization* (*Yleinen asiasanasto* in Finnish). The categorization was quite detailed, so we simplified the ontology by leaving out

some detailed concepts and modifying others to correspond with the semantics of the available content essence.

Although we revised the ontology a number of times, we were not completely satisfied with the results. We had to rebuild some parts of the ontology before it became acceptable. At least some of the initial problems were due to the inexperience in this content domain and in the overall process of creating ontologies and semantic metadata for the content essence.

We used the initial data set and its semantic metadata in our first two prototypes that were tested and used internally in the project. The second ontology was created for the financial news content domain and it was developed to describe the content essence originating from one of our partners, *Kauppalehti Online*, part of a Finnish media conglomerate, *Alma Media Corporation*. This ontology was developed jointly with content domain experts from *Kauppalehti*. First, we developed some rules and guidelines that the experts from *Kauppalehti* used to draft an early version of the ontology, which we then analyzed and revised. After an initial analysis, the ontology was imported into an information extraction tool, the ***Content Provider Tool*** (CPT), and a dedicated person from *Kauppalehti* tested the ontology by creating semantic metadata with the tool. The expert's comments led to modifications both in the tool and in the ontology. After a few iterations, we were able to begin producing semantic metadata for the incoming news articles.

We used the second ontology in a pilot in which we tested the functionality of SmartPush with live content essence and real users. The second ontology consisted of five dimensions: *Priority*, *Location*, *Industry*, *Company*, and *Subject*.

The roles of the five dimensions were:

- **Priority**. This dimension described the importance of the overall article, judged by the editor to establish which articles are most important. This dimension had initially seven alternative values. An example of a value in this dimension is *Hot*. However, we did not produce semantic metadata for this dimension, so it was excluded from the pilot.
- **Location**. *Location* was a hierarchical structure expressing the geographical coverage of the content essence. It was divided into *continents* and further into *substantial regions* and *countries*. As most of the financial news articles discussed *Finland*, *Finland* was further divided into *regions* and *cities/counties* within *Finland*. This dimension initially had a total of around 700 concepts, and the widest ***node*** in the hierarchy had around 20 sibling nodes. An example of a *location* is *Europe->Western Europe->Great Britain.*
- **Industry**. *Industry* contained 16 alternatives. The concepts were derived directly from an already existing categorization for news articles. An example of *industry* is *Communications and Electronics*.
- **Company**. *Company* contained mostly publicly traded companies in *Finland*. This dimension did not have any hierarchy, and the total amount of concepts in this dimension was 227. An example of a concept in this dimension is *Nokia.*

- **Subject**. *Subject* was a hierarchical structure describing the topics of the content and grouping them together. It consisted of roughly 95 leaf-level concepts covering mostly economic, financial, company, and society -related topics. These aspects formed a taxonomic hierarchy that was a maximum of three levels deep. At the top level the structure had eight different branches. The widest node in the hierarchy (*Companies->Production and trading*) had 17 child nodes. An example of a *subject* is path *Company news->Ownership->Merger or alliance.* The *Subject* dimension had ***weights*** that expressed the relative importance of different concepts in that dimension. Weights allowed the semantic metadata to have much more expressive power compared to simple binary membership relations. Weights and hierarchies provided the personalization system in the SmartPush project with a natural way to accumulate relevance of detailed concepts to higher-level concepts.

6.3.3 Metadata production and the Content Provider Tool (CPT)

The author put a considerable effort in developing a tool for ***semi-automatic metadata creation***. Semi-automatic metadata creation means that the tool creates a suggestion for relevant semantic metadata, after which the user is responsible for making required changes and accepting metadata before it is sent further in the content value chain. The resulting Content Provider Tool (CPT) allows a reporter to add semantic metadata to news articles. The following Figure 20 illustrates the role and functionality of CPT.

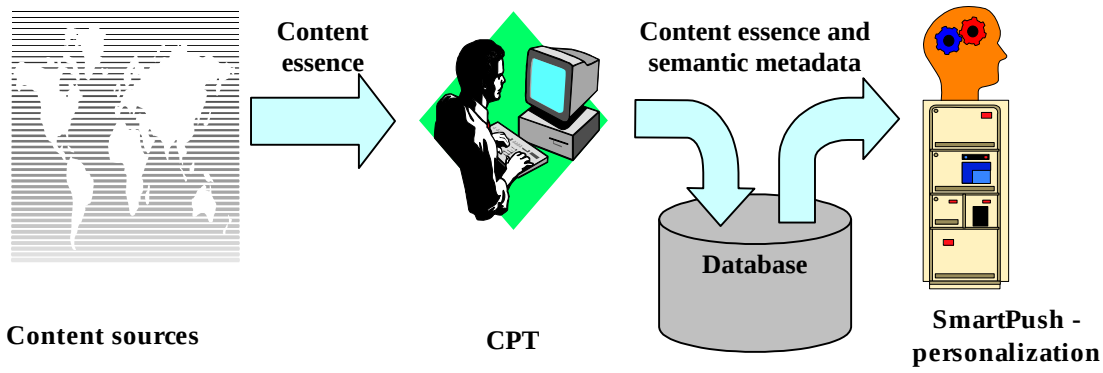

Figure 20. The role of CPT in the SmartPush project.

CPT Workflow

A reporter enters a news article into the CPT via a web page or imports it from an external source. After that, the tool's extraction module processes the textual body of an article by using a two-level morphophonological parser, TWOL from Lingsoft[44], which extracts nouns from the text. These nouns are compared with a set of manually predefined keywords that are mapped to concepts in the ontology. Each ***mapping*** has a weight attached to it that expresses the strength of the binding between the keyword and the corresponding concept. Distinctive keywords have a strong binding, while generic, frequently used words

[44] www.lingsoft.fi

typically have much weaker binding with concepts. Each keyword can have multiple mappings to different concepts and dimensions in the ontology. We can use the keyword *Wall Street* as an example of the mappings. Depending on the ontology, *Wall Street* would probably be mapped to *U.S.* and *New York* in respect to *location* dimension, as well as to relevant concepts in the *subject* dimension, such as *New York Stock Exchange* and *financial markets*.

Mappings between keywords and concepts can also be used to isolate the concepts and ontology from minor changes in the content domain. Even if the meaning of an individual keyword changes, those changes can be reflected in the mappings so that interpretation of concepts remains mostly intact. If we reflect every change in the content domain directly in the concepts in the ontology, we might need to update existing semantic metadata and structures that are built based on that ontology. For example, user profiles in SmartPush used the same ontology as the content, so when the ontology for content essence was changed, we had to reflect those changes in the existing user profiles. However, if we reflect minor changes in the mappings instead of modifying ontologies keeping the ontology and concepts relatively stable, we can have a ***semi-flexible ontology*** that is altered in larger batches, but less frequently.

After identifying all mappings, the CPT accumulates the weights from each mapping to the corresponding concepts. The resulting list of concepts is then scaled according to the accumulated weights, and the highest-ranking concepts in each dimension are presented as the suggested metadata for the document.

In the next step the user verifies the automatically produced metadata and weights and, if needed, modifies the entries through the CPT user interface.

During the SmartPush project we developed a number of tools for the creation of semantic metadata, each with a varying level of automated extraction of metadata. The final tool is illustrated in Figure 21.

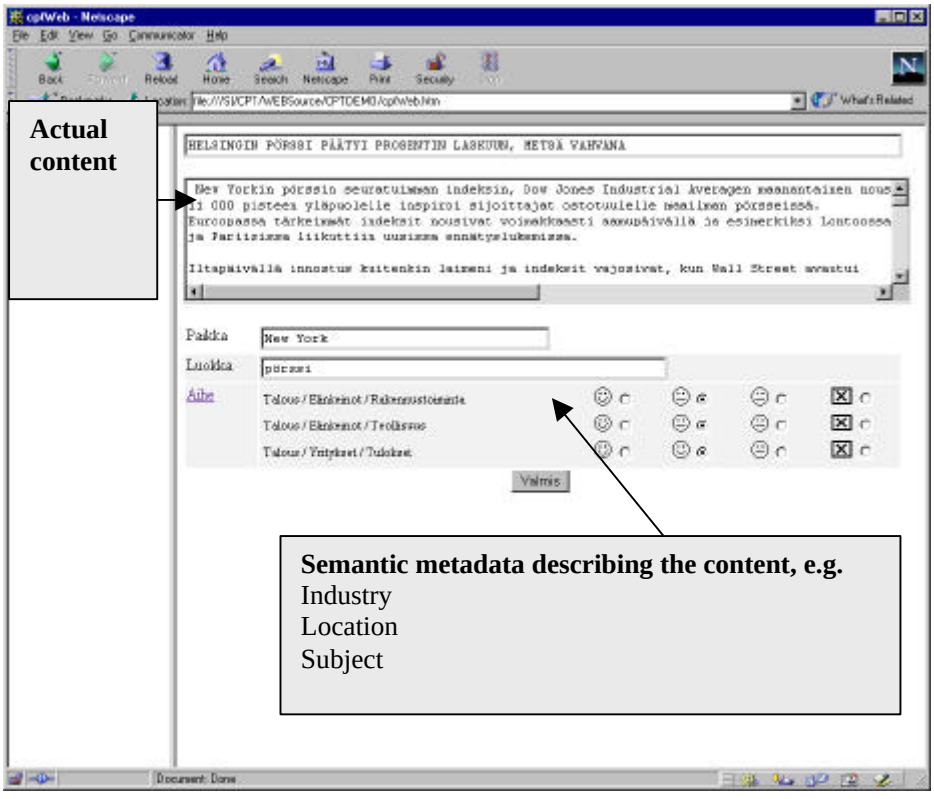

Figure 21. Metadata production with the Content Provider Tool.

The CPT assists the reporter by automating routine tasks like extracting obvious semantic metadata or analyzing highly standardized content essence, such as *stock exchange tables*, but it allows the user to control decisions that require understanding and intelligence. This can be seen as an augmentation of human intelligence, as opposed to relying on full automation. [Lenat, 1995] suggests a similar kind of approach for the commercialization of the CYC, an inference engine for fundamental human knowledge, and an attempt to create an immense multi-contextual knowledge base. According to Lenat, the CYC could be used as a semantic backbone for linking semi-automatic, multiple and heterogeneous external information sources. Semiautomatic in Lenat's work means allowing the user to verify or disambiguate interpretations of a given relation, term, or other parameter.

6.3.4 Results of the SmartPush project

The SmartPush project developed methods for domain modeling, the creation of semantic metadata, and adapting user profiles to customer feedback, and then matching them with content essence that has been augmented with semantic metadata. My work in the project concentrated on the domain modeling and creation of semantic metadata. I, together with other researchers in the SmartPush project, identified and developed iteratively desirable characteristics for semantic metadata and conceptual models. I was also responsible for the development of a novel metadata creation tool and wrote most of its code. This tool was used in producing semantic metadata for the SmartPush project and helped in developing the principles related to the production of semantic metadata.

Kauppalehti and *WSOY Uudet Mediat* illustrated the point that all media companies are not alike, but despite of this they still have a lot in common. Regardless of their differences, they both had many similar issues regarding reuse of content and the need to have ontologies and semantic metadata supporting their operations.

The initial results indicating the usefulness of the SmartPush project and positive feedback from the participating media companies were encouraging, although further research with larger amounts of content would be necessary to validate these conclusions in detail. When compared against other existing systems for personalization, methods used in the SmartPush project functioned well, as was shown in the results of the project[45]. Due to my concentration on the domain modeling and the creation of semantic metadata, the performance of personalization in the SmartPush project is outside of the scope of this work, but is available separately as part of the results of the SmartPush project.

Despite the positive feedback, we did encounter challenges during the project, such as questions related to quality measurement, domain modeling, and ontologies. Some of those challenges are generic and well known, such as the problems with evaluating the accuracy of the information retrieval task (see e.g. [Belkin and Croft, 1992], [van Rijsbergen, 1981]). Some of the challenges were specific to content domains, resulting from the complex and dynamic nature of news. Some difficulties occurred simply due to the inexperience or lack of resources.

[45] SmartPush – Metadata Based Personalized Information Service, final report (internal), 30-June-2000.

I did not concentrate on methods for the creation and administration of keywords and their mappings for the CPT tool, so we defined keywords and mappings manually with the content domain experts. Although the person who created semantic metadata for the final pilot in SmartPush was relatively satisfied with the performance and quality of the automated semantic metadata suggestions in the CPT, more sophisticated and automated methods are needed for the creation and management of keywords and mappings, before tools like the CPT can be used extensively as part of the content value chain.

Although we were not able to perform full-scale testing of the personalization in SmartPush, we were able to convince our partners that semantic metadata and advanced content-based products, such as personalized information feeds, will be important in the future. As an indication of the success, the media companies participating in the project have told the author that they are incorporating the creation of ontologies and management of semantic metadata into their operations and developing new advanced content-based products such as CD-ROMs using the findings and recommendations of the SmartPush project[46, 47].

More information on the SmartPush project, its testing, and results is available in Publication 5.

6.4 Summary of findings

This chapter has introduced some examples and discussed how domain modeling, ontologies, semantic metadata and electronic publishing process can be combined to produce new kinds of advanced content-based products. These qualities may require additional effort and major modifications of the production. When they are fully implemented, however, new opportunities for advanced content-based products will emerge. These opportunities allow the content to be used for purposes beyond the mere presentation of content essence. For example, advanced content-based products can be used to manage information overload, produce recommendations, or to create interactive and personalized views of the content.

Media companies do not yet utilize the possibilities of structured content and produce advanced content-based products. The situation is changing, however, as more and more advanced content-based products are being developed. When companies develop new advanced content-based products, the development should consider the characteristics of the outcome and available content, including quality, relevance, and clarity of content, the amount of content provided, as well as the value of content. The content should also originate from credible sources, be suitable for further use, and be delivered on a timely basis.

[46] "SmartPush project has helped us to understand the importance of ontologies and semantic metadata and to identify what kind of content benefits most from semantic metadata. These findings are clearly reflected on our content production plans for the future." Interview with Mikko Laine, CEO of eWSOY, SanomaWSOY Corporation, 28-March-2001.

[47] "SmartPush project has opened our eyes regarding the need to integrate semantic metadata and ontologies as part of our production of news information. Some of the results of SmartPush project are currently being integrated into our content management." Interview with Marko Turpeinen, CTO of Alma Media Corporation, 1-April-2001.

Some of the new possibilities for advanced content-based products include information personalization services like SmartPush. The SmartPush project demonstrated a valid and viable approach, in which media companies augment content essence with semantic metadata and use that content in producing a personalized information service. Media companies that participated in the SmartPush project have acknowledged the impact of the SmartPush project in their plans and are currently implementing and altering their electronic publishing process. Their actions show the practical importance of semantic metadata, ontology, electronic publishing process and personalization research.

7 Conclusions

Convergence is changing electronic publishing into a horizontally connected content value chain. In the content value chain, the different stages of the content value chain operate in close co-operation. Business rationale for content is developed by understanding the media company's role in the content value chain, by analyzing the value of information, by understanding the advantages of semantic metadata to the product management, and by identifying the new business opportunities from advanced content-based products. In an ideal situation, the role for media companies in the content value chain is to act as the source, quality control, and brand for the content, calling for control of the content and creation of metadata. However, the roles and interfaces in the content value chain are still more or less unclear, and the current trend towards consolidation and horizontal acquisitions may change the roles of different players dramatically.

A key requirement for realizing the advantages of the content value chain is the existence of semantic metadata and content domain specific ontologies. Before semantic metadata can be used in the content value chain, ontologies must be defined and companies must standardize the ontologies and semantic metadata descriptions or make them compatible by mapping them directly or by using an intermediate ontology. Ontologies should have a clearly stated content domain and they should isolate different characteristics of content essence into independent dimensions. The lack of suitable ontologies and systems for content management have driven some media companies to develop their own methods and custom-made tools, but in the near future better alternatives, cost advantages, and co-operation requirements will likely lead media companies towards increasingly more standardized solutions for content management.

Electronic publishing should support reuse of content on multiple media products and media platforms. My work introduces a four-layer model for electronic publishing, which helps to identify and understand the role of different elements affecting advanced content-based products. Media companies should keep the content and its presentation separate as long as possible and try to use standardized formats of content. Separate treatment of content and presentation and standardized formats reduce costs and allow better content reusability. If the content and media products are used multiple times, the qualities of media products and media components should be captured into reusable templates.

The electronic publishing process is divided in this work into research and development and electronic publishing. Those activities must be treated as separate, but closely interrelated processes. This means that media companies should not develop new functionality as part of the production process, but should perform all modifications of technology platform as a separate activity, after which the results are integrated into production of content.

When semantic metadata and improved publishing processes are available, new advanced content-based products are possible. Interactivity, communities, user-originated content, multi-platform publishing, information management, information augmentation, and information valuation are some examples of the new possibilities. If organizations are willing to share information of their internal information needs, media companies may be able to enhance and augment those operations with external content in the fields such as competitor tracking. Although semantic metadata complicates the production process, the advantages

through content reuse and new advanced content-based products outweigh the negative aspects, such as more complex production or increased need to collaborate with other participants of the content value chain.

The SmartPush project has shown in practice that semantic metadata is useful in creating advanced content-based products, and that media companies are willing to alter their existing publishing processes.

Media companies should produce semantic metadata as part of the authoring of content. The goal is to automate the routine tasks during authoring as much as possible, while leaving the author in control of the creation of metadata. I have built a supporting application for this purpose, the Content Provider Tool, which has been successfully used in the production of content.

In addition to issues outside the scope of this work, such as organizational and legal questions, I have identified and touched on a number of important questions that remain open for future research. These questions include advanced methods for performing domain modeling, versioning and mapping ontologies, measuring the quality of the ontologies and semantic metadata, tool support for automated metadata creation, and proper tools for management and visualization of ontologies. Future work should also concentrate on validating and generalizing the presented results. Metadata enhanced content management in media companies is still in its infancy and lacks detailed research methods, so I relied heavily on iteration, constructive research methods, experimentation, and collaboration with the industrial partners of the SmartPush project. Although I gained quite a detailed understanding of the relevant activities at our partnering media companies, my results are still based on a very few examples and should be validated and generalized with methods such as longitudinal and comparative case studies (see e.g. [Myers, 2001]). Future work should also concentrate on improving the research methods and test individually the quality of ontologies, produced semantic metadata, and advanced content-based products. This is important, because the end-to-end performance of the content value chain, from the creation of semantic metadata to the use of advanced content-based products, is extremely difficult to measure and analyze as a whole.

Although a number of unanswered questions must be addressed before media companies can use semantic metadata as an established part of their publishing process, this research and the feedback from the media companies clearly indicate that media companies are positioning themselves in the content value chain and altering and expanding their operations towards advanced content-based products, reuse of content, and systematic electronic publishing.

List of figures

List of tables

References

All the links in the references have been tested for validity on August 15, 2001.

Balabanovic, M., Shoham, Y. (1997) *Fab: Content-Based, Collaborative Recommendation*, Communications of the ACM, March 1997, Vol. 40, No. 3.

Belkin, N. J., Croft, W. B. (1992) *Information Filtering and Information Retrieval: Two Sides of the Same Coin?*, Communications of the ACM, December 1992, Vol. 35, No. 12.

Berghel, H. (1999) *Value-Added Publishing*, Communications of the ACM, January 1999, Vol. 42, No. 1.

Boll, S., Klas, W., Sheth, A. (1998) *Overview on Using Metadata to Manage Multimedia Data*, in Sheth, A., Klas, W., editors (1998) *Multimedia Data Management, Using Metadata to Integrate and Apply Digital Media*, McGraw-Hill, New York, USA.

Bormans, J., Hill, K., editors (2001) *MPEG-21 Overview*, ISO/IEC JTC1/SC29/WG11/N4041, available at http://www.cselt.it/mpeg/standards/mpeg-21/mpeg-21.htm

Bos, B. (2000) *W3C Web Style Sheets home page*, available at http://www.w3.org/Style/

Bradley, S. P., Nolan, R. L., editors (1998) *Sense & Respond. Capturing Value in the Network Era*, Harvard Business School Press, Boston, Massachusetts, USA.

Bray, T., Paoli, J., Sperberg-McQueen, C. M. (1998) *Extensible Markup Language (XML) 1.0*, W3C recommendation, available at http://www.w3.org/TR/REC-xml

Brickley, D., Guha, R. V. (2000) *Resource Description Framework (RDF) Schema Specification 1.0*, W3C Candidate Recommendation 27 March 2000, available at http://www.w3.org/TR/rdf-schema/

Bruck, P. A. (1997) *The Content Challenge, Electronic Publishing and the New Content Industries*, European Commission DG XIII/E Report written by Techno-Z FH Forschung & Enwicklung GmbH, Salzburg, Austria, published by Information Engineering, Telematics Applications Programme, European Commission DGXIII/E.

Carrara, M., Guarino, N. (1999) *Formal Ontology and Conceptual Analysis: A Structured Bibliography*, version 2.5 - March 22,1999, available at http://www.ladseb.pd.cnr.it/infor/ontology/Papers/Ontobiblio/TOC.html

Chen, H., Hsu, P., Orwig, R., Hoopes, L., Nunamaker, J. F. (1994) *Automatic Concept Classification of Text From Electronic Meetings*, Communications of the ACM, October 1994, Vol. 37, No. 10.

Chernock, R., Dettori, P., Dong, X., Paraszczak, J., Schaffa, F., Seidman, D. (1999) *Approaches in Enabling Electronic Commerce Services over Digital Television*, Proceedings of the Second International Conference on Telecommunications and Electronic Commerce (ICTEC), Nashville, Tennessee, USA.

Chomsky, N. (1988) *Language and Problems of Knowledge*, The Managua Lectures, The MIT Press, Cambridge, Massachusetts, USA, Ninth printing, 1997.

Collins, D. J., Bane, P. W., Bradley, S. P. (1997) *Winners and Losers. Industry Structure in the Converging World of Telecommunications, Computing, and Entertainment*, in Yoffie, D., editor (1997) *Competing in the Age of Digital Convergence*, Harvard Business School Press, Boston, Massachusetts, USA.

Cowie, J., Lehnert, W. (1996) *Information Extraction*, Communications of the ACM, January 1996, Vol. 39, No. 1.

Cruse, D. A. (2000) *Meaning in Language. An Introduction to Semantics and Pragmatics*, Oxford University Press, New York, USA.

Curtis, K., Foster, P. W., Stentiford, F. (1999) *Metadata - The Key to Content Management Services*, Proceedings of the Meta-Data '99, The third IEEE Meta-Data conference, April 6-7, 1999, Bethesa, Maryland, USA, available at http://computer.org/proceedings/meta/1999/papers/56/curtis.html

Davies, J., Weeks, R., Revett, M. (1996) *Jasper: Communicating Information Agents for WWW*, World Wide Web Journal, VOL. 1(1), winter 1996, available at http://www.w3j.com/1/davies.180/paper/180.html

DeJong, G. F. (1982) *An overview of the FRUMP system*, in Lehnert, W. G., Ringle, M. H., editors (1982) *Strategies for Natural Language Processing*, Erlbaum, Hillsdale, New Jersey, USA.

Denning, P. (1982) *Electronic Junk*, Communications of the ACM, March 1982, Vol. 25, No. 3.

Dodd, W. P. (1990) *Convergent publication, or the hybrid journal: paper plus telecommunications*, Electronic Publishing, Vol. 3(1), February 1990, available at http://cajun.cs.nott.ac.uk/compsci/epo/papers/volume3/issue1/ep027wd.pdf

Dowling, M., Lechner, C., Thielmann, B. (1998) *Convergence - Innovation and Change of Market Structures between Televisions and Online Services*, in Buchet, B., Schmid, B. F., Selz, D., Wittig, D. *EM - EC in the Insurance Industry / Converging Media*, EM - Electronic Markets, Vol. 8, No. 4, 12/98, available at http://www.businessmedia.org/netacademy/publications.nsf/all_pk/1124

Dublin Core. *Dublin Core Metadata Element Set, Version 1.1: Reference Description*, 1997-07-02, available at http://purl.org/dc/documents/rec-dces-19990702.htm

Dumbill, E. (2000) *Syndicating XML, XML in News Syndication*, XML.COM, July 17, 2000, available at http://www.xml.com/pub/2000/07/17/syndication/newsindustry.html

Enlund, N. (1994) *Production management for newspapers*, Seybold Newspaper Conference Proceedings 1994, London, April 1994, available at http://www.gt.kth.se/research/bigbrother/seybold.html

Enlund, N., Maeght, P. (1995) *A Recommendation for the interconnection of production tracking systems in newspaper production*, in Bristow, J.A., editor (1997) *Advances in Printing Science and Technology*, Volume 23, John Wiley & Sons, Chichester, 1997, available at http://www.gt.kth.se/Forskningsrapporter/1995/Enlund.et.al.A.recommendat.pdf

Ericsson, K. A., Charness, N. (1997) *Cognitive and Developmental Factors in Expert Performance*, in Feltovich, P. J., Ford, K. M., Hoffman, R. R., editors (1997) *Expertise in Context: Human and machine*, AAAI Press/The MIT Press, Menlo Park, California, USA, Massachusetts/London, United Kingdom.

Fahey, L., Prusak, L. (1998) *The Eleven Deadliest Sins of Knowledge Management*, California Management Review, Volume 40, Number 3, Spring 1998.

Feltovich, P. J., Spiro, R. J., Coulson, R. L. (1997) *Issues of Expert Flexibility in Contexts Characterized by Complexity and Chance*, in Feltovich, P. J., Ford, K. M., Hoffman, R. R., editors (1997) *Expertise in Context: Human and machine*, AAAI Press/The MIT Press, Menlo Park, California, USA, Massachusetts/London, United Kingdom.

Foltz, P. W., Dumais, S. T. (1992) *Personalized Information Delivery: An Analysis of Information Filtering Methods*, Communications of the ACM, December 1992, Vol. 35, No. 12.

Good, N., Schafer, J. B., Konstan, J. A., Borchers, A., Sarwar, B., Herlocker, J., Riedl, J. (1999) *Combining Collaborative Filtering with Personal Agents for Better Recommendations*, AAAI-99, Proceedings of the Sixteenth National Conference on Artificial Intelligence, AAAI Press/The MIT Press, Cambridge, Massachusetts, USA.

Gruber, T. (1997) *What is an Ontology?*, Stanford Knowledge Systems Laboratory, available at http://www-ksl.stanford.edu/kst/what-is-an-ontology.html

Hartley, J. (1982) *Understanding News*, Methuen, London, United Kingdom.

Heeman, F. C. (1992) *Granularity in structured documents*, Electronic Publishing, Vol. 5(3), September 1992, available at http://cajun.cs.nott.ac.uk/compsci/epo/papers/volume5/issue3/ep067fh.pdf

Hirschman, L., Brown, E., Chinchor, N., Douthat, A., Ferro, A., Grishman, R., Robinson, P., Sudheim, B. (1999) *Event 99: A Proposed Event Indexing Task for Broadcast News*, Proceedings of the DARPA Broadcast News Workshop, February-March 1999, Herndon, Virginia, available at http://www.itl.nist.gov/iaui/894.01/publications/darpa99/html/dir5/dir5.htm

ISO8879. *SGML, Standard Generalized Markup Language*, IS 8879.

Jacobs, P. F., Rau, L. F. (1990) *SCISOR: extracting information from on-line news*, Communications of the ACM, October 1990, Vol. 33, No. 10.

Kauffman, R. J., Riggins, F. J. (1998) *Information Systems and Economics*, Communications of the ACM, August 1998, Vol. 41, No. 8.

Kendall, J., Kendall, K. (1999) *INFORMATION DELIVERY SYSTEMS: An Exploration of Web Pull and Push Technologies*, Communications of Association for Information Systems, Volume 1, Paper 14, April 1999, available at http://cais.isworld.org/articles/1-14/article.htm

Kim, H. M. (2000) *Developing Ontologies to Enable Knowledge Management: Integrating Business Process and Data Driven Approaches*, Papers from the 2000 AAAI Spring Symposium, Technical Report SS-00-03, AAAI Press, Menlo Park, California, USA.

Konstan, J. A., Miller, B. N., Maltz, D., Herlocker, J. L., Gordon, L. R., Riedl, J. (1997) *GroupLens: Applying Collaborative Filtering to Usenet News*, Communications of the ACM, March 1997, Vol. 40, No. 3.

Krulwich, B., Burkey, C. (1997) *The InfoFinder agent: Learning user interests through heuristic phrase extraction*, IEEE Intelligent Systems Journal (Expert), 1997, Vol. 12, No. 5, available at http://www.computer.org/intelligent/ex1997/x5022abs.htm

Kubala, F., Colbath, S., Liu, D., Srivastava, A., Makhoul, J. (2000) *Integrated Technologies for Indexing Spoken Language*, Communications of the ACM, February 2000, Vol. 43, No. 2.

LaFrance, M. (1997) *Metaphors for Expertise: How Knowledge Engineers Picture Human Expertise*, in Feltovich, P. J., Ford, K. M., Hoffman, R. R., editors (1997) *Expertise in Context: Human and machine*, AAAI Press/The MIT Press, Menlo Park, California, USA, Massachusetts/London, United Kingdom.

Lassila, O., Swick, R. R. (1999) *Resource Description Framework (RDF) Model and Syntax Specification*, W3C Recommendation 22 February 1999, available at http://www.w3.org/TR/REC-rdf-syntax/

Lenat, D. B. (1995) *CYC: A Large-Scale Investment in Knowledge Infrastructure*, Communications of the ACM, November 1995, Vol. 38, No. 11.

Lenat, D. B. (1998) *From 2001 to 2001: Common Sense and the Mind of HAL*, in Stork, D. G., editor (1998) *HAL's Legacy: 2001's Computer as Dream and Reality*, MIT Press, available at http://www.cyc.com/halslegacy.html

Mack, G. (2000) *AOL/Time Warner marriage causes M&A shakeup*, Redherring.com, inside tech, January 12, 2000, available at http://www.redherring.com/insider/2000/0112/inv-aol.html

Maclachlan, M. (1999) *Third Voice Aims for More Interactive Web*, CMP Techweb News, May 17, 1999, available at http://www.techweb.com/wire/story/TWB19990517S0026

Malone, T. W., Grant, K. R., Turbak, F. A., Brobst, S. A., Cohen, M. D. (1987) *Intelligent information-sharing systems*, Communications of the ACM, May 1987, Vol. 30, No. 5.

Martínez, J. M., editor (2000) *Overview of the MPEG-7 Standard (version 4.0)*, ISO/IEC JTC1/SC29/WG11 N3752, September 2000, available at http://www.cselt.it/mpeg/standards/mpeg-7/mpeg-7.zip

Maybury, M. (2000) *News on Demand*, Communication of the ACM, February 2000, Vol. 43, No. 2.

Meyer, M. R., editor (1998) *Final Report of the EBU/SMPTE Task Force for Harmonized Standards for the Exchange of Programme Material as Bitstreams*, EBU Technical Review, Special Supplement 1998, European Broadcasting Union (EBU) and the Society of Motion Picture and Television Engineers, Inc, Geneva, Switzerland.

Minsky, M. (1986) *Negative Expertise*, in Feltovich, P. J., Ford, K. M., Hoffman, R. R., editors (1997) *Expertise in Context: Human and machine*, AAAI Press/The MIT Press, Menlo Park, California, USA, Massachusetts/London, United Kingdom.

Myers, M. (2000) *Qualitative Research in Information Systems*, available at http://www.auckland.ac.nz/msis/isworld

Myers, M. (2001) *References on Case Study Research*, available at http://www2.auckland.ac.nz/msis/isworld/case.htm

O'Reilly, T. (1996) *Publishing Models for Internet Commerce*, Communications of the ACM, June 1996, Vol. 39, No. 6.

Oinas-Kukkonen, H. (1999) *Mobile Electronic Commerce through the Web*, Proceedings of the Second International Conference on Telecommunications and Electronic Commerce (ICTEC), Nashville, Tennessee, USA.

Palmer, J. W., Eriksen, L. B. (1999) *Digital Newspapers Explore Marketing on the Internet*, Communications of the ACM, September 1999, Vol. 42, No. 9.

Pepper, S., Moore, G. (2001) *XML Topic Maps 1.0 TopicMaps.Org Specification*, available at http://www.topicmaps.org/xtm/1.0/

Porter, A. L., Newman, N. C., Watts, R. J., Zhu, D., Courseault, C. (2000) *Matching Information Products to Technology Management Processes*, Papers from the 2000 AAAI Spring Symposium, Technical Report SS-00-03, AAAI Press, Menlo Park, California, USA.

Rada, R., Carson, G. S. (1994) *The New Media*, Communications of the ACM, September 1994, Vol. 37, No. 9.

Rappaport, A. T. (1997) *Context, Cognition, and the Future of Intelligent Infostructures*, in Feltovich, P. J., Ford, K. M., Hoffman, R. R., editors (1997) *Expertise in Context: Human and machine*, AAAI Press/The MIT Press, Menlo Park, California, USA, Massachusetts/London, United Kingdom.

Reeves, B., Nass, C. (1996) *The media equation: how people treat computers, television, and new media like real people and places*, CSLI Publications, Cambridge University Press, Cambridge, United Kingdom.

Reuters (2000) *Reuters to pioneer new multimedia news delivery*, Press release for the news agency Reuters Ltd, July 27, 2000.

van Rijsbergen, C. J. (1981) *Information Retrieval*, second edition, Butterworth, available at http://www.dcs.gla.ac.uk/Keith/pdf/

Rohde, L. (1999) *Digital TV develops in the U.K.*, IDG News Service, November 30, 1999, available at http://www.idg.net/crd_idgsearch_93807.html?sc=48370658_20622

Rust, G. (1998) *Metadata: The Right Approach. An Integrated Model for Descriptive and Rights Metadata in E-commerce*, D-Lib Magazine, July/August 1998, available at http://www.dlib.org/dlib/july98/rust/07rust.html

Saarela, J. (1999) *The Role of Metadata in Electronic Publishing*, Ph.D. Thesis, Helsinki University of Technology, November 1999, Published in Acta Polytechnica Scandinavica, Mathematics and Computing Series No. 102, Finnish Academy of Technology, Espoo, Finland.

Saarela, J., Turpeinen, M., Korkea-aho, M., Puskala, T., Sulonen, R. (1997) *Logical Structure of a Hypermedia Newspaper*, Information Processing and Management, 1997, Vol. 33, No. 5, Elsevier Science Ltd.

Sabelström Möller, K. (2001) *Information categories and editorial processes in multiple channel publishing*, Ph.D. Thesis, Royal Institute of Technology, Department of NADA, Stockholm, Sweden, available at http://www.gt.kth.se/staff/~kristinas/home/phdthesis.html

Sacharow, A., Mooradian, M., Keane, P., McAteer, S. (1999) *Cross-Media Programming. Creating Promotion and Distribution Opportunities*, Jupiter Communications Vision Report, Content & Programming, May 1999.

Salton, G., Allan, J., Buckley, C. (1994) *Automatic Structuring and Retrieval of Large Text Files*, Communications of the ACM, February 1994, Vol. 37, No. 2.

Savia, E. (1999) *Mathematical Methods for a Personalized Information Service*, Master's Thesis, Helsinki University of Technology, Espoo, Finland.

Schenker, J. L. (2000) *Bertelsmann's Dilemma*, TIME Europe, March 31, 2000, available at http://www.time.com/time/europe/webonly/bertelsmann.html

Schneider, M. (1994) *What is Teletext?*, Philips Semiconductors datasheet, June 1994, available at http://www.semiconductors.philips.com/acrobat/8129.pdf

Shadbolt, N., O'Hara, K. (1997) *Model-Based Expert Systems and the Explanation of Expertise*, in Feltovich, P. J., Ford, K. M., Hoffman, R. R., editors (1997) *Expertise in Context: Human and machine*, AAAI Press/The MIT Press, Menlo Park, California, USA, Massachusetts/London, United Kingdom.

Shapiro, C., Varian, H. R. (1999) *Information Rules. A Strategic Guide to the Network Economy*, Harvard Business School Press, Boston, Massachusetts, USA.

Skyrme, J. (1999) *Knowledge Networking. Creating the Collaborative Enterprise*, Butterworth-Heinemann, Oxford, United Kingdom.

Smith, J., Jutla, D. (1999) *Dimensions in Emerging Business Models in eCommerce*, Proceedings of the Second International Conference on Telecommunications and Electronic Commerce (ICTEC), Nashville, Tennessee, USA.

Stadnyk, I., Kass, R. (1992) *Modeling Users' Interests in Information Filters*, Communications of the ACM, December 1992, Vol. 35, No. 12.

Teece, D. (1998) *Research Directions for Knowledge Management*, California Management Review, spring 1998, Vol. 40, No. 3.

Turpeinen, M. (2000) *Customizing News Content for Individuals and Communities*, Ph.D. Thesis, Helsinki University of Technology, February 2000, Published in Acta Polytechnica Scandinavica, Mathematics and Computing Series No. 103, Finnish Academy of Technology, Espoo, Finland.

Varian, H. (1999) *Economics and Search*, SIGIR Forum, Fall 1999, Vol. 33, No. 1, available at http://www.acm.org/sigir/forum/F99/Varian.pdf

Wiederhold, G. (1995) *Digital Libraries, Value, and Productivity*, Communications of the ACM, April 1995, Vol. 38, No. 4.

Wilson, N. (2000) *Net content companies announce layoffs, money shortages*, CNET News.com, June 7, 2000, available at http://news.cnet.com/news//0-1005-200-2035978.html?tag=st.cn.sr.ne.1

Wittgenstein, L. (1953) *Philosophical Investigations: The English Text of the Third Edition*, translated by Anscombe, G. E. M., Prentice Hall, 1999.

Yoffie, D., editor (1997) *Competing in the Age of Digital Convergence*, Harvard Business School Press, Boston, Massachusetts, USA.

Zeitz, C. M. (1997) *Some Concrete Advantages of Abstraction: How Experts' Representations Facilitate Reasoning*, in Feltovich, P. J., Ford, K. M., Hoffman, R. R., editors (1997) *Expertise in Context: Human and machine*, AAAI Press/The MIT Press, Menlo Park, California, USA, Massachusetts/London, United Kingdom.

Appendix A: Glossary of Terms

To help understanding and to aid in finding the key terms used in the dissertation, this appendix introduces them in an alphabetical order. A more detailed discussion on the key terms is available elsewhere in the work and in the publications.

Term definitions in this dissertation are marked with ***this style***. When a term is marked with this style, the term or its plural form is defined in the Glossary of Terms.

Advanced content-based products are products and services using advanced computerized methods and automation to process and use digital content. These products typically require metadata, especially semantic metadata, to perform the desired functionality. Examples of advanced content-based products are different systems for recommending content, such as FireFly or GroupLens [Konstan et al, 1997].

Authoring is one of the three process steps related to electronic publishing. It consists of steps that involve creating, acquiring and editing of the content essence together with its metadata.

Component model is a four-layer framework for understanding and managing different elements of the electronic publishing. These four layers help to express the complexity of reusing the same content essence in multiple media products and media platforms.

Concepts are basic building blocks of ontologies. Concepts bear a limited sense of meaning within them, which enables the ontology to be able to capture the semantics at such a level of detail that content essence can be produced and delivered to the customer and used in advanced content-based products.

Conceptual models describe the structure of dimensions in ontologies. They consist of agreed concepts and the relations between the concepts.

Content. A European Union special task force [Bruck, 1997] defines content as

> *"A wide range of information, entertainment, communication and transaction services that combine smart texts, intelligent graphics/simulations, motion in images and texts and are made available to an identifiable user group. The distinction does not matter whether these services are created to be sold or if they are available for free of end-user charges (e.g. corporate publishing and advertising)."*

The definition of content in this work is restricted to professionally created short factual and objective text, such as financial news articles. The applicability and extensibility of these restrictions is discussed elsewhere in the work. I consider content to consist of two components, content essence and its metadata. Although content has traditionally meant only the content essence, the co-existence of content essence and metadata is gaining popularity. For example, the British Broadcasting Corporation (BBC) recently defined content as consisting of both metadata and content essence [Meyer, 1998].

Content composition is one of the three process steps related to electronic publishing. During content composition, suitable media components are selected and integrated into media products.

Content delivery is one of the three process steps related to electronic publishing. In this process step the user gains access to the published media product.

Content domain defines a domain for which certain set of content essence characteristics are true. An example of content domain is *financial news*.

Content essence is the actual substance of the content that is processed, distributed, and used in the content value chain. Without content essence, content is useless. Content essence may be in a variety of forms such as audio, video, or text.

Content management consists of the production and management of content.

Content provider is an alternative term for a media company. Although individuals and other organizations are often a source of content, by associating content provider and media company I want to emphasize the need to have professional organizations responsible of production and control of content in the content value chain.

Content value chain consists of the different participants and their activities in which content is needed, starting from its creation and lasting until it is used for the last time. Value aspect emphasizes the connection between value creation and processing of content.

Convergence describes the ongoing activity, where communications, computing, and content industries are merging into a single, interconnected industry.

Digital content emphasizes content that is in digital format.

Dimension describes a semantic perspective of the content domain. With dimensions the ontology should be able to cover the semantics that are needed to produce and deliver content in the content value chain.

Domain model is an alternative term for ontology.

Domain modeling describes the process needed for building an ontology.

Electronic publishing consists of authoring, content composition, and content delivery to the customer. The customer is not necessarily the final customer of the content, but simply the next possibly computerized step in the content value chain. Although research and development can be considered as a separate process, it is closely interconnected to electronic publishing. Media companies are typically responsible for most or all of the listed activities, although third parties often provide access services for the delivery of content.

Electronic publishing process consists of two, but closely interrelated processes, research and development and electronic publishing.

Flexible content consists of content essence that can be reused for different purposes.

Media company represents companies that create, possess, and/or modify content that is used in the content value chain. My definition excludes companies that provide only generic communications services and products without identifying the kind of data they are distributing. The intermediate and deliverable content produced by media companies is normally in a digital format. The media companies discussed here are

typically large conglomerates that produce different kinds of content essence to be used in multiple media products and media platforms.

Media components consist of both the content essence and its metadata. They are used in media products as individual pieces of content.

Media platform is the medium used to consume the content essence. In most cases, the media platform is equivalent to the end-user terminal, but it may also be seen in a more abstract way such as covering all the methods used to access content essence online. In its simplest form, consuming means the presentation of content essence, which includes reading, listening, and watching the content essence. If the media platform and the content support more sophisticated functionality, consuming may also involve operations such as personalization or interaction. Examples of media platforms are online publications, mobile services, digital television, and digital radio.

Media products is an alternative term for advanced content-based products.

Metadata means information, data with relevance, about content essence. Metadata is needed for the processing, delivery, and usage of content essence, and it can be used for different purposes such as describing media characteristics, status, and semantics.

Metadata enhanced content management extends the definition of content management to the production and administration of metadata, including domain modeling activities. Additionally, metadata enhanced content management emphasizes the use of metadata, in this thesis especially semantic metadata, as an essential part of content management.

Metadata management is the production and administration of metadata and ontologies independent of the management of content essence.

Ontology. The term ontology in this work describes a set of formally specified conceptual models consisting of agreed concepts and the relations between these concepts. Ontology typically covers multiple semantic angles, dimensions, of the content domain.

Process model explains the major steps of producing flexible content, i.e. content essence that can be reused for different purposes.

Research and development concentrates on developing advanced content-based products and ontologies. It is closely interconnected to electronic publishing and is discussed in this thesis as part of the electronic publishing process.

Semantic metadata is a kind of metadata that describes the semantics of content essence. In this work semantic metadata adheres to its ontology and is in machine-usable format.

Semantic metadata management describes activities related to the production and management of semantic metadata.

SmartPush refers to the prototype system built during the SmartPush project for recommending content essence based on semantic metadata and user profiles.

SmartPush project. SmartPush project was a research project concentrating on personalized information feeds based on semantic metadata. The author was actively involved in the project as a researcher and project manager.

Structured content emphasizes the content to consist of both content essence and its metadata.

Technology platform consists of the technology that enables and implements the functionality provided on the media platform, such as different alternatives to view pictures.

Part II

Summary of publications

The following summary briefly introduces the publications in Part II of this thesis, as well as the author's role and contribution to each publication. In addition, the descriptions contain a short comment on the context where these publications were created. All publications are reprinted with the permission of their respective copyright owners.

Publication 1.

Savia, E., Kurki, T., and Jokela, S. (1998) *Metadata Based Matching of Documents and User Profiles*, in Human and Artificial Information Processing, Proceedings of the 8th Finnish Artificial Intelligence Conference, Finnish Artificial Intelligence Society, August 1998, Jyväskylä, Finland, pp. 61-70.

Publication 1 introduces the early ideas for the personalization methodology in the SmartPush project. The focus of this paper is on user profiles as well as on matching and distance measurement algorithms. The work also discusses the nature of metadata, hierarchical ontologies and asymmetric distance measures. The author was a project manager and researcher in the SmartPush project between 1997-1999. In this work the author was a co-author responsible for the metadata section of the paper and contributed to the overall discussion as well.

Publication 2.

Kurki, T., Jokela, S., Turpeinen, M., and Sulonen, R. (1999) *Agents in Delivering Personalized Content Based on Semantic Metadata*, Intelligent Agents in Cyberspace -track, Papers from the 1999 AAAI Spring Symposium, Technical Report SS-99-03, AAAI Press, March 1999, Menlo Park, California, USA, pp. 84-93.

Publication 2 discusses the SmartPush system with an emphasis on its software implementation. The SmartPush system consisted of a metadata creation tool, a number of software agents for personalization, and user interfaces for different types of users. The author was responsible for the metadata tool and wrote most of its code. In this work the author was the principal co-author focusing on the metadata and its generation during content development.

Publication 3.

Jokela, S. and Saarela, J. (1999) *A Reference Model for Flexible Content Development*, Proceedings of the Second International Conference on Telecommunications and Electronic Commerce (ICTEC), October 6-8, 1999, Nashville, Tennessee, USA, pp. 312-325.

Publication 3 discusses different aspects of the electronic publishing with the goal of reusing content on multiple media products and media platforms. The models introduced in the paper were developed based on a close co-operation with the media partners in the SmartPush project. The author was the main author of the paper and developed the process-related aspects of the work as well as participated closely in authoring other parts of the paper.

Publication 4.

Jokela, S., Turpeinen, M., and Sulonen, R. (2000) *Ontology Development for Flexible Content*, Proceedings of the HICSS-33, IEEE Computer Society, January 4-7, 2000, Maui, Hawaii, USA, *Best Paper Award of the Internet and Digital Commerce track.*

Publication 4 discusses the prerequisite for the electronic publishing and advanced content-based products, the existence of ontologies for semantic metadata. The paper is a result of research and discussions during the SmartPush project, as well as from the experiments with the media companies participating in the project. The author was the main author of the paper and created the model for describing the main components of the domain modeling activity. This paper received the best paper award of the *Internet and Digital Economy* track in the HICSS-33 conference.

Publication 5.

Jokela, S., Turpeinen, M., Kurki, T., Savia, E., and Sulonen, R. (2001) *The Role of Structured Content in a Personalized News Service*, Proceedings of the HICSS-34, IEEE Computer Society, January 3-6, 2001, Maui, Hawaii, USA.

Publication 5 discusses in detail the metadata related activities in the SmartPush project. It explains the steps taken to create the ontologies and what kind of experiences and lessons were learned during that project. The paper also describes the development of the Content Provider Tool, CPT, which was the author's software contribution to the project. The author was the main author of the paper.