Ontology Development for Flexible Content

Sami Jokela
Helsinki University of Technology
sami.jokela@hut.fi

Marko Turpeinen
Alma Media Corp.
marko.turpeinen@almamedia.fi

Reijo Sulonen
Helsinki University of Technology
reijo.sulonen@hut.fi

Abstract

Converging media industry calls for tighter integration of creativity, business processes, and technologies. Media companies need flexible methods to manage electronic content production and delivery, and metadata is a key enabler in making this goal a reality. However, metadata is useful only if its nature is understood clearly and its structure and usage are well-defined. For this purpose, an ontology, consisting of conceptual models that map the content domain into a limited set of meaningful concepts, is needed. This paper introduces an ontology development framework rooted at the core business processes of electronic publishing that can be used to define semantic metadata structures for electronic content. The framework underlines the different nature of ontology development and metadata publishing, and how these two processes influence each other. This paper discusses also the application of the ontology development framework in practice. The framework has been created in the SmartPush project, where media companies explore new business opportunities for electronic publishing and delivery.

1 Introduction

Content providers are facing new opportunities and challenges as new models of electronic publishing are emerging. Media companies have to re-use and personalize their content for multiple media platforms and content products. This calls for better understanding and management of both the production and delivery of media content.

In the SmartPush project [13], new methods and tools for content production and delivery are developed in co-operation with Finnish media companies. This work concentrates on personalized content delivery based on descriptions of the content, i.e. its semantic metadata. With the help of semantic metadata the SmartPush system tracks user's interests and adjusts user profiles based on customer feedback. Our work has shown that high quality metadata is essential in building customized news services. Moreover, even if the content provider does not produce personalized content, already the management of existing content production and delivery requires substantial amount of metadata. If metadata is not available or used, content providers cannot manage the content and typically end up wasting resources and money in reproducing the same content over and over again.

Integrating metadata with the publishing process calls for co-existence of artistic creativity and systematically managed content production. One of the most important issues related to metadata is how their structures, ontologies, are defined, and how the changes in the world are reflected to those structures. This paper introduces a framework that can be used to simplify and assist in defining ontologies for electronic publishing. The paper begins with a definition and description of the key concepts in ontology development. After that, electronic publishing process and ontology development framework are discussed in detail. Then we describe how the framework can be used in practice. We show also how these ideas can be applied to the development of new services, like information filtering on the SmartPush project, and news augmentation in the domain of business information.

1.1 Related work

The idea of describing and using semantic information in a formal and systematic way is not new. This interest

has extended from philosophical foundations over the discussion on representation and meaning to the more practical issues related to acquiring and using the semantics (see e.g. [11]). Linguistics and especially artificial intelligence have been among the most active fields of research. Internet with new types of multimedia content (see e.g. [12], [8] for further discussion) and new web-standards have been latest motivators for the work on ontology and semantic metadata issues.

Artificial intelligence community has used a considerable amount of resources to define common methods and tools for developing ontologies. One of the most notable efforts has been the work done at the Knowledge Systems Laboratory (KSL) at the Stanford University. Ontolingua project (see e.g. [4]) has developed a distributed collaborative environment to modify and use ontologies over the web. The work at KSL has been a starting point for a number of other projects. For example, the KA2 initiative [6] uses Ontolingua collaboration environment and aims at building an ontology for annotating WWW documents of the knowledge acquisition community in order to enable intelligent access to these documents. Ontologies and metadata have generated interest also outside the academic community. Reuters [10] among others have developed their own proprietary structures and methods for describing content, but full-scale standards for ontologies and content semantics do not yet exist although some work in the field has been done (see e.g. Dublin Core [18]).

The business community has lately put a lot of expectations on the Extensible Markup Language, XML [14] and its descendants (see e.g. RDF [17]) to solve interoperability issues between companies. Most of these issues are, however, not solved with existing XML standards as they provide only a transportation mechanism, but do not take stand in defining the supporting ontology. The situation is similar with publishing industry specific standards like Information Content Exchange [15], NITF/XML News [9], and BizTalk [1], although their latest versions have developed to the right direction. A number of attempts to conceptualize the domain and build a suitable ontology have been conducted in other domains such as Mathematics [16] or Chemistry (see e.g. [5]), but these results are not directly applicable to other domains.

2 Ontologies and semantic metadata for news content

Both the terms metadata and ontology have variable interpretations depending upon circumstances in which they are used. Metadata means information about information and it can be used for different purposes such as to describe media characteristics, content processing, and actual content semantics [2].

With the definition of the word ontology we stay away from the more complicated definitions used for example in knowledge acquisition and representation [11]. In our paper ontology means a set of formally specified metadata structures consisting of commonly agreed concepts that bear a limited sense of meaning within them. Ontology describes the semantics and can cover multiple different aspects, dimensions, of the content. With these dimensions the ontology should be able to cover the semantic needs that are needed to produce and deliver the content to the customer. The following figure [

Figure 1] visualizes the relations between different aspects of our ontology.

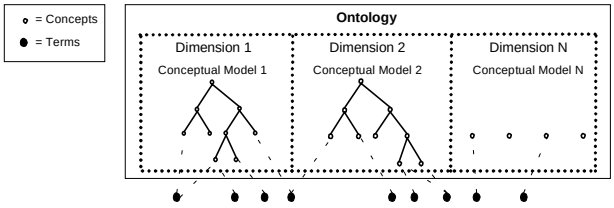

Figure 1. Different aspects of domain ontology

Ontology comprises a set of concepts and concept relationships representative to the domain. Concepts and their relations define *conceptual models* for classifying (or "tagging") information objects under different *dimensions*. Dimensions are typically independent from each other and they have their own conceptual models. We have preferred hierarchies in the conceptual models due to their computational and representation advantages, but this is not a mandatory requirement. Ontology can be more than a taxonomy or classification, and can include multiple types of relationships between concepts. An example of dimensions is Location, Subject, or Author of a document. Subject and Location -dimensions have well-defined hierarchical conceptual models, whereas Author typically contains only the author name without deeper conceptualization. A c*oncept* is an abstract term generalized from particular instances – the thing, entity, or idea a particular word refers to. Typical concepts for Location dimension in our example would be different country names. *Terms*, by contrast, are the actual words that refer to concepts. These would be for example different presidents and national events that can be linked to the respective countries. As can be easily seen from the example, ontology is more than an agreed vocabulary, because it provides a set of well-founded constructs that can be used to build meaningful higher level knowledge.

Metadata can be categorized in multiple ways (see e.g. [2]). We use a process-oriented approach by classifying metadata into three categories: semantic, structural, and control metadata. Even though all these three aspects are

important, this paper concentrates mostly on discussing different aspects of semantic metadata, i.e. the meaning of the content.

The ontology for the content must be able to describe both the incoming information feeds and the needs of the customers. If either is ignored, the ontology is seriously impaired. In addition, the ontology has to be able to cope with the dynamic nature of the information feeds and customer interests. This means, that there must exist methods for reflecting changes back to the ontology.

2.1 Justification for semantic metadata

Metadata production and utilization require considerable amount of resources and effort. There are, however, a number of reasons why it is advisable to produce and use semantic metadata:

- *Size*. Semantic metadata is a condensed representation of the content. It captures the essential semantics of the source. With textual content the savings depend on the detail level of the produced semantic metadata. With other source formats such as video or audio the savings can be, however, much more substantial. Instead of using megabytes of original content, many tasks related to the content production and distribution can be performed with a much smaller semantic metadata representation.
- *Support for multiple formats.* Metadata can support many different media formats - such as text, images, video, or audio - with same representation format. If we consider the difficulties borne with the management of multiple different formats, we can easily understand the advantages of having content semantics represented in a single uniform way.
- *Expressing hidden/author's views.* This can be considered to be either an advantage or a disadvantage. On one hand semantic metadata forces the author to express the message explicitly in the semantic metadata. This helps to define the key facts making the content clearer. On the other hand semantic metadata may reveal the hidden message author wanted to express without stating it explicitly.
- *Common view on machine-usable level.* Because key information is stored explicitly in the metadata, there is less room for different interpretations what the content is about.
- *Saves computation*. Content analysis, especially with video and audio, requires extensive computing. If we have analyzed the content during its creation, we do not have to perform the analysis later on the fly.
- *Higher information quality.* High quality metadata increases the possibilities to produce and deliver accurate information and new kinds of services to the customers.

On the other side of the coin are at least the following issues:

- *Ontology creation*. One should not underestimate the difficulties related to building a good ontology. The creation of an initial ontology requires a special set of skills as well as expertise on both the domain and the customers' needs. Moreover, ontology may require multiple iterations before it is usable.
- *Expressiveness of the ontology*. It is naïve to claim that textual metadata is able to express all possible aspects of the content. Emotions and subjectivity are two examples of the broad range of difficult challenges in this field.
- *Effort required in the metadata creation.* Metadata creation can be very expensive in terms of human resource consumption. It might also be difficult to pinpoint the exact value the metadata is producing.
- *Dynamic nature of metadata structures*. Metadata structures change over time, which raises a number of issues related to managing already existing metadata.
- *Degradation of the information.* Not only the structures change, but the overall correctness and value of information changes over time. We must be able to produce and manage multiple versions of both the metadata and the content.
- *Different media qualities and content types*. If we try to apply the same metadata model for different media and/or content types, we have to understand and produce metadata to cover all the unique aspects of the sources. For example, image metadata does not have to take into account temporal relations, whereas for metadata describing a video clip this information is essential.

Most of these issues are discussed later in this paper as we introduce our ontology development framework.

Some of the metadata advantages appear during the authoring, some later during the delivery or consumption of the content. It seems, that the more complex the content production and manipulation is, the more advantageous it is to use metadata in the process. For example, content personalization benefits from semantic metadata that can be used in selecting actual content to be delivered to the customer’s site. This approach decreases network traffic and is more flexible than transferring and processing huge amounts of original content.

2.2 Nature of ontologies for news content

Librarians have worked for centuries to find usable ways of describing and categorizing information. Automatic information filtering systems, such as SmartPush [13] tries to route information by matching explicitly defined content metadata with customer profiles that describe user needs. The specification of these needs is based on a well-defined domain model.

The problem with news is that the domain is open and unbounded. Comprehensive computer-based representation of an ontology that covers the whole news domain is impossible to create. A typical news producer organization has access to information feeds, such as wire services from Reuters and AP. These news feeds have their own classifications of content that we refer to as information feed ontologies. These feeds are typically treated as raw information for more thorough news reporting, and even if the content from information feeds would be used as such, they typically require re-categorizing. Therefore a news ontology must be based on journalistic judgment by the content provider, which is called the provider ontology. The content provider tries to cover its content domain in a way that would be most useful to its customers. This provider ontology can be explicitly available for customers for making selections for their information needs, and typically serves as a basis for customization for different individuals and communities of news consumers. The customers may also have defined their own ontology, a customer ontology, which has to be linked to the provider ontology. These ontologies and their relations are described in [Figure 2].

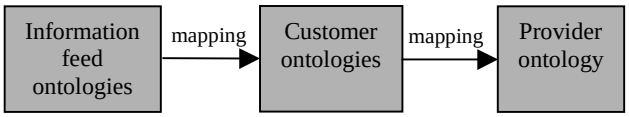

Figure 2. Mappings between ontologies

Ontology mapping means defining comparable and relating concepts to facilitate the usage of content over heterogeneous data sources. If mapping is not performed, the semantic metadata is not compatible with others.

3 Ontology development framework for electronic publishing

The following framework for ontology development has been developed in the SmartPush project. It reflects our experiences, according to which the ontology development and usage must be linked closely to the processes of content production. The main purpose of the framework is to separate ontology development from its usage and to explain, which factors affect these processes and how they are interrelated. Before we introduce the framework, we will start by introducing the key processes of electronic publishing [Figure 3].

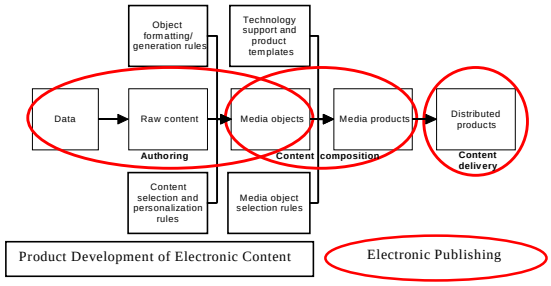

Figure 3. Key processes of electronic publishing

Electronic publishing and product development of electronic content are closely related, but distinct, processes. Whereas electronic publishing is an on-going activity performed by content experts, product development of electronic content is a project-like effort conducted by a team of technology, domain, and methodology experts. The ontology and metadata aspects are, however, inherent in both processes.

Our ontology development framework is divided onto two phases: ontology development and metadata publishing. The ontology development matches with the product development of electronic content. In this phase conceptual models and metadata structures are created and modified. The metadata publishing phase concentrates on the actual production of metadata, which takes place mostly during the content authoring, but is inherent also in other electronic publishing activities. Ontology development framework [Figure 4] illustrates the components involved in developing ontologies and using metadata in electronic publishing. Although these two phases use same kinds of inputs, they also contain fundamental differences, which we discuss after the introduction of the ontology development framework.

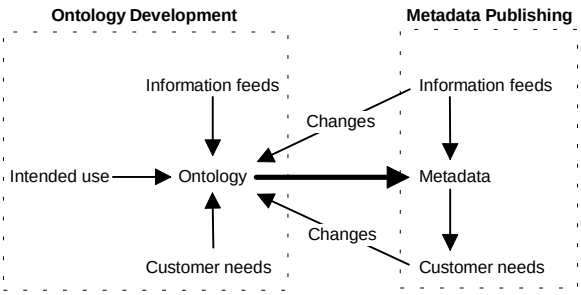

Figure 4: Ontology development framework

3.1 Ontology development

As in product development of electronic content, ontology development is a periodical effort and requires a different set of skills than the metadata publishing. Persons developing ontologies must understand the

content domain, and be able to concretize this knowledge into an ontology. When the ontology is developed, three interconnected aspects have a major impact on the structure of the ontology: intended use, information feeds, and customer needs.

Intended use. Ontology can be used for many purposes, including both the content production and the usage of the finished content. Also, the nature of required metadata depends on the purpose it is used for. For example, content formatting requires mostly structural metadata whereas content selection and personalization is based on semantic metadata. The requirements of content production are relatively easy to reflect to the ontology, because most of the production normally follows well-defined production guidelines. A more challenging task is to estimate, how final content is used and how those operations affect the ontology. One way to approach this challenge is to examine all supported media as well as the ways and reasons customers consume the content on these media. From this analysis it is possible to conclude requirements for content products. These requirements can then be converted into metadata needs and required ontologies.

Information feeds. The contents and reusability of current and planned information feeds define the structure of the ontologies. If the same content can be used multiple times, the semantic metadata must fulfill all the requirements of resulting content products. It is also important to consider how much metadata the incoming feeds already contain and is this metadata usable. The key issue here is how much of the conversion from information feed ontology to the provider ontology can be automated. For example, if 95 percent of incoming information is produced by a single source, it might be advisable to develop the provider ontology so that manual conversion effort is minimized. Whatever the approach, the content provider must understand what kind of information it has access to and what are its key characteristics.

If the key competitive quality of information feed is its instantaneous nature, the time that is required to process the content must be minimized, in some cases even by sacrificing the depth and quality control of the metadata. An alternative to this approach is to produce metadata in two phases, where initial metadata is produced on the fly when the information is published, while a more detailed and higher quality version of metadata is published later. This resembles live broadcasts, where on-going events are reported immediately without analyzing them deeper.

Customer needs. Third part in the ontology development is customer needs, which ultimately define why the ontology exists. If existing or future users do not need a certain quality of metadata there is not point in producing it. Even though it is possible to produce vast amounts of metadata, it should be produced only if it is valuable to the production process or to the customers. The identification of these needs is a difficult but important task and requires a joint effort of different departments including management, marketing and editorial staff. Typical methods to identify customer needs include traditional business planning, customer segmentation, and marketing activities. The two-way nature of customized news services provides a new and efficient mechanism to discover and track customer needs.

3.2 Metadata publishing

Metadata publishing is part of the electronic publishing and is an on-going effort. It connects users to the information feeds by providing semantic metadata for the publishing and delivery of content.

During the metadata publishing authors analyze content from internal and external sources and create its semantic metadata description. Metadata publishing relies on the existing ontology and produces a description of the actual content in a machine-usable format. It is advisable to try to automate this process as much as possible while letting the author be in control.

Fred Brooks [3] proposed the formula IA > AI, which illustrates the symbiosis between mind and the machine. It means that intelligence amplification - or intelligence augmentation - is more important than artificial intelligence, which is a machine imitating the mind. In the context of ontology development and metadata creation for news content, this can be interpreted as a requirement for semi-automatic tools that assist human experts in these difficult tasks.

Information extraction tools can be used to select descriptive terms from the content. These tools can classify the extracted terms under logical types such as location, person, organization, industry, and subject area for categorization. This term extraction is then used to select best fitting categories according to different ontological dimensions. The end result is a list of proposed concepts belonging to the conceptual models. The list of concept candidates is shown to the author, who checks the quality and relevancy of the computer-based suggestions. This calls for tools that support the used ontology and generate a metadata suggestion that can be modified. When the semantic metadata is ready and accepted, it is linked to the actual content and sent further in the electronic publishing process. Metadata information is then used to select, customize, and deliver the content to the customer.

Metadata publishing requires domain and ontology knowledge as well as journalistic skills. The author must be familiar with the structure and contents of the ontology and understand how the metadata will be used in the electronic publishing process. The amount of work

required depends on the already available metadata in the information feeds and on the automation level of the supporting tools.

The metadata publishing process must also contain mechanisms to alter the ontology when needed. Changes in the information feeds, domain, or user needs must be reflected back to the ontology development. In addition, the impact of ontology modifications to the existing metadata must be analyzed. If the content provider wants to use already generated metadata, it must define the principles for converting existing metadata entries into revised ontology. For example, if the content provider has defined only country names in the ontology but constantly ends up having London in the information flows, the new entry can be added to the ontology. In this case, however, the content provider must decide, should they change all or some of existing metadata references for U.K. to point to London, or should they leave those references intact.

4 Using ontologies and semantic metadata in electronic publishing

This chapter links ontology development framework to electronic publishing and discusses important issues that are relevant when the framework and semantic metadata are used in practice. There are no right answers to these questions, but this analysis reflects some tactics and ideas we have discovered in our work.

4.1 Ontology creation and modification

Ontology creation occurs usually when the content provider decides to extend its domain coverage. If the ontology for the domain is already defined, the development of new content products typically requires only ontology modifications, not a totally new ontology.

Ontology modifications, however, are likely to occur even if there is no need to modify the content products. This is the situation, when existing ontology does not match with the incoming information feeds or customer needs, but the content product remains the same.

Standardization. If the content provider wants to use metadata from external sources or if other partners want to use the produced metadata, the companies involved must have a common agreement on the ontologies as well as on their administration mechanisms. If the ontologies are not compatible via similar structures or mappings between them, the metadata must be reproduced.

Although there is a clear need for standardized ontologies, they are very difficult to develop. Numerous attempts to build standardized semantics for different domains have failed. The developers have not been able to conceptualize the domain, they have drowned in details, or they have not been able to create a shared understanding of the domain. We believe, however, that it is possible to create a shared standard, if one understands the domain, is able to define the ontology, and has enough power to expand the ontology to other organizations. Reuters [10] is a good example of a company that has been able to create a shared newsfeed standard. They have managed to build a network of some 4000 suppliers all producing content and metadata according to Reuters' proprietary ontologies.

The internal structure of the ontology. Some of the most critical questions related to the ontology structures are:

- *What dimensions should be included in the ontology?*
- *What should be the domain coverage and detail level of the ontology?*

Dimensions, detail level, and domain coverage all affect the complexity of the ontology. If the concept model for a dimension is simple or if the semantic metadata can be created automatically, the ontology may contain multiple dimensions. Adding a dimension that requires a lot of manual work needs to be carefully considered. If the content provider wants to produce highly detailed semantic metadata, the amount of dimensions and the domain coverage must be limited, or the metadata publishing process cannot be managed. Likewise, if the content provider wants to cover a wider domain, the detail level suffers.

Dimensions, detail level, and domain coverage all have common characteristics. They should be defined based on how much value the customers put on the information and how the semantic information will be used. If the production of a piece of information is too expensive in relation to its perceived value, there is no point to include it in the ontology. This in turn is related to how much of the metadata publishing process can be automated and how much of the information can be derived from the metadata in the incoming feeds. If the customers value highly the promptness of the information, the ontology and tools must allow quick pass-through. Even if fast processing times are not a necessity, the metadata entry tools must support easy browsing of the ontology without the need to memorize its internal structures.

Granularity. Granularity issues are related to the scope of semantic metadata. For some content types granularity is not relevant, as is the situation with short news stories. With larger publications, however, one has to divide the material into smaller pieces and to define, how these pieces and their metadata are combined on higher levels. For example, if the content provider wants to create semantic metadata for a book, the author has to divide the book into smaller parts such as chapters and then to define metadata for each part. The author has to define also how the parts are combined together and what the representative metadata on these higher levels will be. If

the author just adds it all together, the result is likely to be a huge pile of metadata without any information on which description is truly relevant at the book level. Temporal or geographical information may also pose granularity problems, because they may have complex interpretations and multiple values within a single piece of content.

Degradation. Even though the content provider is able to define an ontology for a certain domain, the ontology will change over time. It is thus important to understand the dynamic nature of the domain and how the degradation affects the conceptualizations. A good example is how, in the 1950's, a news story in a Finnish newspaper covering space travel and plans for a manned rocket launch was categorized under "Funny World". Today, the computer industry is a good example of a domain that is very fast-paced, and where terminology is changing rapidly.

There are many challenges involved with degrading and dynamically changing ontologies. It is important to understand what to do with existing categorized information, when the underlying conceptual models for categorization change.

Stabilizing the ontology structure. Developing a good ontology for a certain domain is extremely difficult. It is very likely that the first version of the ontology has to be modified thus calling for multiple rounds of improvement. Iterative development of the ontology is a good idea, but before altering the structure one has to consider the implications of change. If the structure is altered, we must convert the existing metadata to match the new structure, or otherwise the existing information becomes unusable.

The decision, whether the ontology is ready to be used, should be based on the status of different influencing factors of the ontology, i.e. information feeds, intended use, and user needs. When the ontology is capable of describing the incoming information feeds at such detail level that the company can use the content as intended and the user needs are met, the ontology is usable.

One way to measure this readiness is simply by testing the ontology in practice by producing semantic metadata with it. If customers experience no difference between a number of documents even though they all have different metadata descriptions, the ontology may be too detailed. On the other hand, if the customers feel that two documents should be differentiated although they both have identical semantic metadata, there might be a need to deepen the ontology.

4.2 Ontology usage

Metadata publishing covers production and usage of semantic metadata in the electronic publishing process. If content originates from the content provider, content and semantic metadata activities are often integrated. If the media company acts only as an integrator for the content from other sources, the company must incorporate metadata publishing to the integration activities. This can be achieved either by converting existing semantic metadata into company's own ontology, or by creating semantic metadata from scratch during the content publishing process.

Quality. Quality control of both content and its metadata is very important. Even though the media company does not produce the content itself, its reputation is a major factor determining how high the customers value the content. When metadata is added to the content and distributed, the provider must ensure that metadata meets the same quality requirements as the content itself. Poor quality metadata results in deteriorated end products.

The quality should also be measured, which is a challenging task. There are so many subjective players in the process that traditional relevancy/irrelevancy measures do not produce valid results. We suggest that the analysis should consider different aspects of the ontology development framework and be based on the subjective customer views as well as how well those needs are fulfilled.

Another question related to quality is whether we can maintain the metadata quality if the work is performed by different persons and over an extended period of time. Our initial findings state that this is possible, but requires proper training, constant exposure to the work, and proper tools to support the work.

Linking provider ontology to customer needs. Content producers consider themselves as experts in modeling their content domain, but they often forget that different customers and customer communities may have different views on the same subject matter. Customers have varying interests and expertise levels, their terminology differs and they interpret things differently. All these variations should be taken into account as much as possible when semantic metadata is produced. The goal of the producer should be to create semantic metadata that covers most of the needs of their customers.

Another method to improve the usability of the ontologies and semantic metadata would be to allow the customers access to the formal definition of the ontology and make modifications to it. Customers could combine the results with other ontologies that are either their own or from other information sources. This way the customer would have control on the scope and level of detail of the ontology.

4.3 Tools for ontology usage

Proper tools are essential in incorporating semantic metadata into the electronic publishing process. The tools should produce an automatic suggestion for the metadata that the author can then modify. The tool should also be

incorporated into the electronic publishing process so that additional sidesteps are not needed. These tools should allow the user to define, how the existing metadata is modified when ontologies are altered and ensure the consistency of the ontology. This is especially important when the ontology is under construction or when the domain changes noticeably over time. A desirable functionality for an ontology usage tool would be support for templates. With them the author could generate automatically semantic metadata for certain standardized information such as stock exchange quotations. The tool should also allow the user to teach term-concept associations instead of defining them manually. There should also be proper functionality to visualize both the associations and the ontology. If this functionality is not provided, it is difficult to understand how the system works and what kind of information can be expressed in the metadata.

The tools that use ontologies need to scale up to large conceptual models containing tens of thousands of concepts. Technical performance is important since the tool will be used in a process, where time is a critical element. A slow and cumbersome tool will be neglected or the authors will change their working methods to bypass the problems.

5 Cases based on ontology development framework

5.1 Personalized news filtering

SmartPush project has been going on since 1997 at the Helsinki University of Technology, TAI Research Centre. In SmartPush media companies produce semantic metadata for their news content. Semantic metadata is used in creating and delivering personalized news feeds on different media. Personalization is based on user profiles that have a similar structure as the semantic metadata. In order to adapt to the changing user needs the customer profiles are modified according to the user feedback. The delivery of the results is then determined based on customer preferences and the customer's delivery media [7].

Initial test material for SmartPush consisted of roughly 400 short news articles. Because suitable ontologies did not exist, an ontology for the domain was developed in the project. This work started by analyzing the articles and defining suitable metadata dimensions. Keywords were then collected and assigned to their relevant dimensions. After that the concepts were created based on the keywords and the initial ontology was built. The main emphasis was put on the subject dimension, although the keywords and location information was stored as well.

The initial ontology was used in testing for roughly a year. Although a considerable amount of effort was put into building the ontology, testing clearly indicated some problems in it. These problems were due to the inexperience in building ontologies and due to the lack of domain expertise. Proper tools for structure creation and metadata production were also missing, so the process was difficult to control and required a lot of manual effort.

Tool support was improved with a new application called Content Provider Tool, CPT, which assisted in generating semantic metadata. CPT produced keywords from a textual source using linguistic analysis, after which the author had the possibility to assign the keywords into relevant concepts. The goal was to assist the process but keep the author in control. At this phase the ontology administration was mostly manual.

Initial version of CPT showed room for improvement. Although the CPT could have speeded the process, metadata creation still required a lot of manual effort and the author had to know the structure and representation of the domain ontology. If the author did not know the ontology by heart, the quality of the resulting semantic metadata was poor. Additionally, constant changes in the ontology without proper tools and methods made the administration of existing metadata a difficult task.

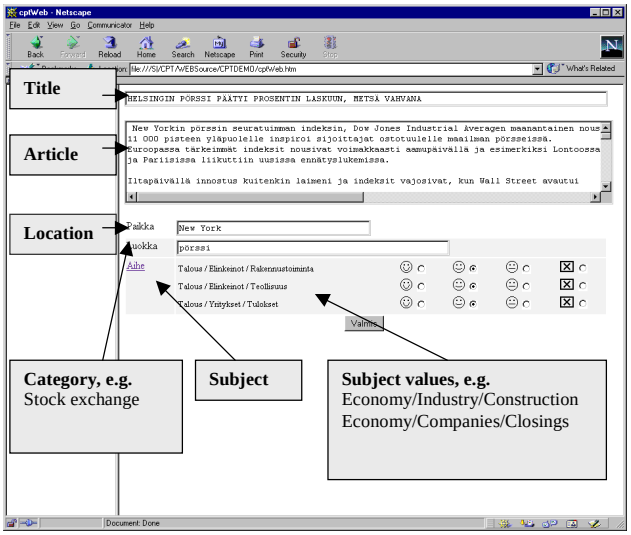

Figure 5. Web Version of the CPT

A new Web-based version of the Content Provider Tool has been developed [Figure 5]. The new version has a simpler and easier user interface with better support for browsing the ontology. It is also better integrated to the electronic publishing process. The main design goal for the new CPT version has been to reduce work in producing semantic metadata for the content. This is achieved by associating terms into the concepts of the ontology. If an extracted term clearly points out a certain concept in the ontology, that association has a strong weight. If the term is related to a number of concepts,

each association should have a smaller weight. This method generates a set of candidates for the semantic metadata, which are then scaled and presented to the author. Author can then modify the metadata before it is used further in the targeting and delivery process.

5.2 News augmentation

News augmentation is introduced here briefly as an example of a service that uses the ontology development framework when matching published content metadata with user and community models. Information Augmentation (IA) combines news streams with selected explanatory material from heterogeneous information sources. The augmentations can be customized to individuals and communities, based on customer models that consist of special interests, expertise level, previous activity, and community context. The structure of the user model mirrors the conceptual models along which the news material is categorized. An example of a user model that reflects the interest and expertise levels of a user in relation to the conceptual model is shown in [Figure 6].

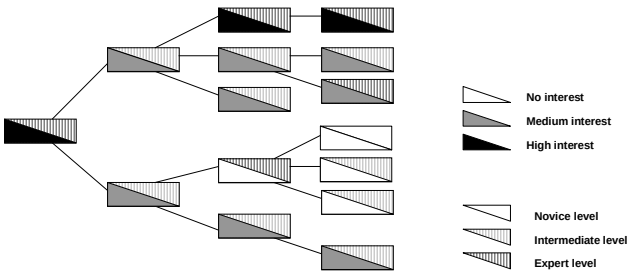

Figure 6. Levels of interest and expertise in a hierarchical conceptual model

These semantically rich customer models are then used to provide various kinds of augmentations to news content. Once a story is selected for reading, the reader can query a search engine for related material. An example of this approach is a news augmentation agent, *historical context visualizer*, which takes a dynamic approach to producing context using proactive and personalized conceptual information retrieval. It automatically creates a presentation of related articles and visualizes the relationships between the concepts in these articles. A historical context visualizer, which has been developed for the Web service of a Finnish financial newspaper called Kauppalehti, is shown in [Figure 7].

The main contents of this augmentation module are organizations, people, and content topic areas. Organizations are defined as a dimension in the ontology; people are stored as a free form metadata field. The content topic areas are also defined in the domain ontology as a separate dimension.

Figure 7. Historical background for an article

A dynamic text visualization tool shows the concepts, the relationships between the concepts, and the relevance of concepts to the user profile and community profile. Concepts are colored using specific presentation rules. Previous personal exposure to concepts is reflected as shades of gray, and concepts of community importance (with no indication of personal interest) are shown in shades of red. The user profile is given more weight than the community profile when visualizing the concepts. Timeline helps to see which concepts have been important at different moments in time. The headlines shown below the timeline slider change dynamically depending on the concepts and time period selected.

6 Conclusions and future work

Metadata is the key element in managing content production in the future. Metadata-based multimedia content management is dependent on high-quality domain ontologies, which are, however, very difficult to develop and manage. This paper has presented an ontology development framework to assist in defining a suitable ontology for electronic publishing. We have emphasized that ontology development and metadata publishing are dynamically interconnected processes. They are both tightly linked to incoming information feeds and the needs and ways how the users want to use the information. It is also important to understand the linkage between these two processes. Even though ontology development is a project-like effort, changes observed during metadata publishing must be reflected back to the ontologies.

There are a number of challenging issues in integrating semantic metadata into the electronic publishing. However, we strongly believe that effort is

recommendable. If semantic metadata is available, the content providers are able to produce new kinds of products as well as to reduce duplicated effort in producing the existing ones.

We will continue developing our test ontology with a pilot using multiple live information feeds and real customers. Content use, user needs, and available information sources all affect how the ontology will evolve in the future. Although some of the changes can be managed with altering the associations, the pilot is likely to cause a chain of iterative changes also in the domain ontology itself.

There are many ways the ontology development framework could be further extended. One possibility is to examine what kinds of organizational aspects semantic metadata production is causing in the media industry. The applicability of this framework to other rapidly expanding domains of electronic business could also be evaluated.

7 Acknowledgements

This research is part of the SmartPush project sponsored by the Finnish Technology Development Centre TEKES and Alma Media, WSOY-Sanoma, Sonera, ICL, Nokia Research Centre, and TeamWARE Group.

8 About the authors

Sami Jokela has worked as a project manager and researcher at the Helsinki University of Technology (HUT)/ TAI Research Centre. His project team researches personalized electronic publishing. Their project, SmartPush, is conducted together with a number of industry partners and has a strong focus on using developed methods in practice. Jokela continues his research currently at the Andersen Consulting Center for Strategic Technology Research (CSTaR) in Palo Alto, CA.

Marko Turpeinen is a Ph.D. candidate in computer science at HUT. Since 1996 he has worked at Alma Media Corporation, a Finnish media company, after managing the Ota Online project at HUT in 1994-95. Currently he is on his second year as a researcher at the MIT Media Laboratory, in Cambridge, MA.

Reijo Sulonen is a Professor of Computer Science at the Helsinki University of Technology. His research interests include database systems, product data management, process modeling, software engineering, and electronic media.

9 References

[1] *BizTalk*, WWW-site for the BizTalk organization, http://www.biztalk.org/

[2] Boll S., W. Klas, and A. Sheth, "Overview on Using Metadata to Manage Multimedia Data", in *Multimedia Data Management. Using Metadata to Integrate and Apply Digital Media*, McGraw-Hill, New York, U.S.A, 1998, pp. 1-24.

[3] Brooks F. P., "The Computer Scientist as a Toolsmith II", *Communications of the ACM*, Vol. 39, No. 3, March 1996, pp. 61-68

[4] Fikes, R., A. Farquhar, and J. Rice, *Tools for Assembling Modular Ontologies in Ontolingua*, Knowledge Systems Laboratory, Stanford University, 1997, http://www.ksl.stanford.edu/

[5] Gómez-Pérez A., M. Fernández, and A. de Vicente, *Towards a Method to Conceptualize Domain Ontologies*, Workshop on Ontological Engineering, ECAI'96, Budapest, Hungary, 1996, pp. 41-52.

[6] *KA2*, WWW-site for the Knowledge Annotation Initiative of the Knowledge Acquisition Community, 1999, http://www.aifb.uni-karlsruhe.de/WBS/broker/KA2.html

[7] Kurki T., S. Jokela, M. Turpeinen, and R. Sulonen, *Agents in Delivering Personalized Content Based on Semantic Metadata*, Intelligent Agents in Cyberspace, Papers from the 1999 AAAI Spring Symposium, Technical Report SS-99-93, AAAI Press, Menlo Park, CA, 1999.

[8] Meersman R., Z. Tari, and S. Stevens, *Database Semantics. Semantic Issues in Multimedia Systems*, Kluwer Academics, Norwell, MA, 1999.

[9] *NITF/XML News*, WWW-site for the XMLNews organization, 1999, http://xmlnews.org/

[10] *Reuters*, WWW-site for the Reuters, Limited, London, United Kingdom, 1999, http://www.reuters.com/

[11] Shapiro S., "Hermeneutics, Knowledge Acquisition, Knowledge Representation", *Encyclopedia of Artificial Intelligence*, Volume 1 A-L, 2nd Edition, John Wiley & Sons, Inc., 1992, pp. 596-611, 719-742, 743-758.

[12] Sheth A. and W. Klas, *Multimedia Data Management. Using Metadata to Integrate and Apply Digital Media*, McGraw-Hill. New York, U.S.A., 1998.

[13] *SmartPush*, WWW-site for the SmartPush -project. Helsinki University of Technology, TAI Research Centre, 1999, http://smartpush.cs.hut.fi/

[14] W3C, *Extensible Markup Language (XML™)*, W3C Architecture Domain, 1999, http://www.w3.org/XML/

[15] W3C, *Information and Content Exchange (ICE) Protocol*, Submission to the World Wide Web Consortium, 1998, http://www.w3.org/TR/NOTE-ice

[16] W3C, Mathematical Markup Language (MathML), *W3C User Interface Domain*, 1999, http://www.w3.org/Math/

[17] W3C, *Resource Description Framework (RDF™)*, W3C Technology and Society Domain, 1999, http://www.w3.org/RDF/

[18] Weibel, S. and E. Miller, *Dublin Core Metadata Element Set WWW homepage*, 1999, http://purl.org/metadata/dublin_core