SEMANTIC WEB TECHNOLOGIES FOR CONTENT REUTILIZATION STRATEGIES IN PUBLISHING COMPANIES

Andreas Andreakis[1], Adrian Paschke[1], Alexander Benlian[2], Martin Bichler[1], Thomas Hess[2]
[1] *Technische Universität München, Boltzmannstr. 3, 85748 Garching, Germany*

[2] *Ludwig Maximilians University Munich, Ludwigstr. 28, 80539 Munich, Germany*

Keywords: Content reutilization, Semantic Web, multi-channel strategies, semantic annotation.

Abstract: In recent times content reutilization in different media channels (such as in Cross Media Publishing or Windowing) is a much discussed concept in the media industry. It promises decreasing production and coordination costs by exploiting and leveraging synergy effects. First theoretical investigations emphasized the importance of efficient metadata-enriched linking and modularity techniques for a successful implementation of this multi-usage concept. However, nearly all recent approaches stop at the level of theoretical suggestions and do not include novel technological potentials provided by Semantic Web languages and tools. On the basis of previous works, this paper attempts to fill these research gaps by presenting a "proof of concept" implementation and by highlighting the possibilities and the potential of Semantic Web technologies in the context of media content reutilization.

1 INTRODUCTION

The progress in product and process digitization as well as the ever-growing bulk of media content are prompting print publishing firms to organize their most valuable resource in a way that it can be allocated efficiently to production and bundling processes. Effective and efficient content reutilization practices do not only promise a decrease of production and transaction costs by reducing search and coordination costs of editors (Schulze, 2005), but also an enrichment of media content portfolios due to higher exploitation rate of marketable content products (Anding , 2004).

Publishing companies (i.e., book, newspaper and magazine publishers) are traditionally organised in loosely-coupled, topic-specific departments with a strong trend to decentralized structures in order to cover a broad range of different issues (e.g. international, national news, business, technology, science, sports, etc.) (Schumann, 2002). This product orientation has also been reflected in the data or content architectures of publishers. Media content has most often been located in separate repositories spread over separate editorial units without any logical, semantic linkages (Stamer, 2002). However, such semantic relationships become more and more important in particular in the context of emerging multi-channel content reutilization strategies in publishing companies. Existing Content Management Systems (CMS) such as Cofax, OpenCMS, CoreMedia, already support editorial offices with XML-based content workflows, which can automatically serve different media channels (Stamer, 2002). Only in some cases, semantic annotations of the contents were applied in form of metadata (e.g. Dublin Core, IPTC) and predominantly in the form of simple XML-based "name-value" pairs. Large-scale semantic networks with lots of cross-references between content modules can hardly be found (Schek, 2005). As a consequence, editors are often forced to "reinvent the wheel" by producing and bundling pictures, graphics, and texts several times, again and again. This leads to a lot of redundant work and potential coordination conflicts between editorial units.

In this paper, we address this research gap by examining the question *whether Semantic Web technologies can be used to implement a distributed multi-media content management system capable of efficient metadata-based search and sophisticated*

Andreakis A., Paschke A., Benlian A., Bichler M. and Hess T. (2006).
SEMANTIC WEB TECHNOLOGIES FOR CONTENT REUTILIZATION STRATEGIES IN PUBLISHING COMPANIES.
In *Proceedings of WEBIST 2006 - Second International Conference on Web Information Systems and Technologies - Internet Technology / Web Interface and Applications*, pages 491-494
DOI: 10.5220/0001252604910494

content reutilization in editorial production and bundling processes. The objective of this study is to present a "proof of concept", which should spur further research efforts. The paper is structured as follows. Section 2 discusses related work. Section 3 describes the basic concept of a RDFS-bade media content management systems and discusses further extensions with respect to distributed ontology management and query optimization. Section 4 evaluates the presented approach and section 5 gives a summary of our contribution.

2 RELATED WORK

Use cases for the classification of media content are manifold in the research field of the Semantic Web. Examples for the organisation of media content in the Internet are RSS (Rich Site Summary) for exchanging news articles; the MPEG classification and Adobe's XMP (Extensible Metadata Platform) framework, which is integrated into the "network publishing" approach.
Furthermore, various research projects deal with the employment of Semantic Web technologies in the media industry. The OntoMedia Ontology (Lawrence, 2005) describes a specification for the visualization of semantically annotated media content in heterogeneous media environments based on OWL and is therefore comparable to current standards such as PRISM. Schreiber et al. (Schreiber, 2001) describes an approach for ontology based annotation of photos using RDFS and a Prolog-based inference machine.

Compared to these approaches the work described in this paper addresses practical aspects regarding the efficient search and multiple use of media contents in a Semantic Web-enabled multi-user environment

3 RDFS-BASED MEDIA CONTENT MANAGEMENT

Semantic Web Ontology languages such as RDFS and OWL are a promising approach to develop expressive meta data languages with comprehensive cross-linkages and taxonomic sub- and super-class relationships. They enable a compact representation of hierarchical meta data concepts and provide rich inference features for effective and precise query answering. Under real conditions with myriads of taxonomic cross-references between ever-growing bulks of media content modules and topic hierarchies RDF Schema (RDFS) which provides basic constructs for describing RDF vocabularies and taxonomic structures is more suitable as a representation language than OWL. According to our studies and theoretical complexity results actual Semantic Web (description logic) interference machines such as Jena (OWL Lite), Racer, Pellet (OWL DL) lead to a relatively insufficient performance and scalability for large OWL models with low performance for query answering and high memory consumption. Using RDFS adequate meta data languages for media content annotation and categorization within large taxonomies comprising a multitude of categories and topics can be represented and efficiently queried.

In our implementation (based on RDFS and Jena) sub-class relationships (*subClassOf*) organise individual topic categories (*classes)* into hierarchical structures (T-box model ~ *taxonomy*), in which the meta data *properties* of a parent class are inherited via RDFS inferences to the subclasses. The media contents are assigned as *instances* via type relations (*type*) to one or more topic categories and are annotated with the inherited properties from all superordinated categories reaching from *domain independent meta data* such as "*author", "creation date"* to *domain dependent meta dat*a, i.e. properties of particular topic categories. Via multi-typing a particular instance can be assigned to the several topic categories within the taxonomies as well as to arbitrary other RDFS-based categorization schemes.

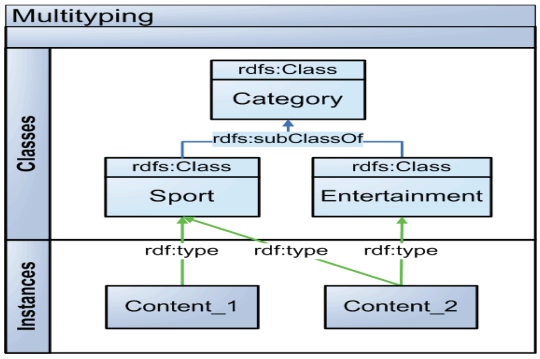

Figure 1: Multiple typing of media contents (instances) with topics (classes).

Beside the topic-related taxonomy an additional classification into data types (*text, image, video and audio*) has been implemented to enable categorization of media content according to their type, e.g. „*text*". Standard Semantic Web inference engines such as Jena capable of RDFS reasoning can be used to search for media contents whereas the

defined categories and meta data properties are used to refine and constrain the queries, e.g.:

"Get all contents (instances) of **category** "*Sport*" and **category** "*Entertainment*" of **data type** "*text*" not older than **date** "*31.1.2006*" and with the **author** „*John Doe*"?"

The multiple-categorization and the extensive meta-data annotation of media content facilitates precise content search and leads to high recall which enables efficient content reutilization due to a meaningful reuse of existing media content modules simultaneously or subsequently in different output channels. Further extensions are needed in order to support collaborations of loosely-coupled, topic-specific departments which have their own heterogeneous ontologies and in order to optimize query answering in view of large taxonomies and a huge number of media contents. In the next two sections we will further elaborate these issues.

3.1 Distributed Ontology Management

Departments or companies can interact and reuse media contents with each other based on a shared taxonomy which unifies their heterogeneous decentralized ontologies. In order to achieve this, a common basis, i.e. a centralized ontology, can be used which might be dynamically extended at runtime. Therefore, we have defined a *base ontology* with one predefined root class "Category":

<rdfs:Class rdf:about="http://ns #Category" />

Via the "owl:imports" construct (e.g. Jena supports this import statement) other ontologies which define further category classes, category properties and associated media instances are linked into the base ontology at runtime whereas they extend the "Category" root class. Figure 2 illustrated the integration process of two ontologies into the base ontology.

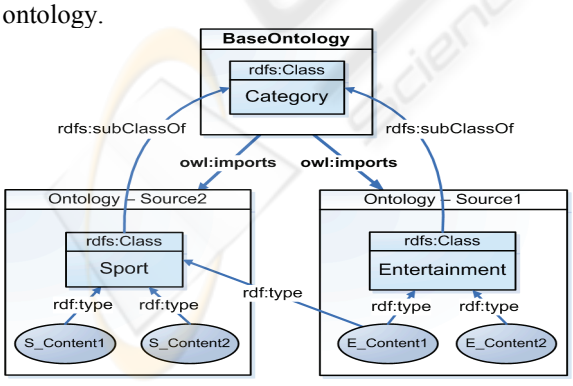

Figure 2: Integration of distributed ontologies.

Media contents might be assigned to multiple categories even between locally distributed sub-ontologies. The base ontology appears as a centralized virtual repository. All updates and extension take place in the local sub-ontolgies and are merged within the virtual base ontology. In the example the media content "*E_Content1*" is of type "*Sport*" and also of type "*Entertainment*".

From a user perspective, i.e. an editor, this multiple classification process of media contents takes place as follows:

1. The editor classifies any new produced media content, e.g. a new article, within its local ontology and associates it to one or more categories within the local taxonomy.
2. Then the editor searches the base ontology for further fitting categories from other ontologies and additionally assigns the content to these external categories.
3. In case that there is no adequate category neither in the local ontology nor in the unified base ontology or in case that a better suited category is needed, the editor adds this new category to her local ontology which is automatically imported by the base ontology.

3.2 Optimization of Content Search

Beside the expressiveness of a meta data language which influences precision and recall of queries, the efficiency of the content search is crucial. The high number of media content and the possibly large "base ontology" consisting of many local sub-ontologies necessitates additional query optimizations. We have implemented two basic approaches, namely search space reduction via *ontology splitting* and *caching.*

The idea behind the ontology splitting is to reduce the search space of queries via splitting an ontology into smaller models which can be consequently inferenced much faster. For example, to query and traverse the topic taxonomy only the T-box model is needed and the A-box model can be neglected. Accordingly a meaningful split might be to break down the ontology into a T-box model with the taxonomy data and an A-box model with all instance data. This can be further extended, e.g. by additionally separating all properties which describe domain-dependent meta data. As a result an ontology consists of three models an T-box category model, a property meta data model and an A-box content model. Specialized queries such as plain "subsumption" or "classification" queries without further meta data restrictions can be answered using one of these smaller models, whereas more complex queries are solved based on the virtual unification (implemented via "owl:imports") of these partial models, i.e. based on the complete (virtual) model.

A client-sided cache can be used to further optimize the content search. The cache is temporarily populated with indexes on frequently used topic categories. These references are reused to directly access the local ontologies and not the large base ontology for recurrently queries. The caching approach is related to the splitting approach, but is implemented on the client-side, where-as the splitting is done on the server side.

In a nutshell, these optimizations significantly boost the query speed (see section 4) and demonstrate the feasibility and scalability of the Semantic Web based media content management approach in case of large ontologies and many contents.

4 EVALUATION OF PROTOTYPE

Within the scope of the presented proof-of-concept implementation, the quality of the suggested concept was evaluated with respect to the applicability and usability of current Semantic Web technologies for meta-data enhanced content reutilization. Although the implemented system just represents a prototype, data concerning the performance and scalability of the system have been collected. Furthermore, a qualitative benchmarking to search technologies in conventional content management systems that are based on key word search or simple name-value-metadata was carried out. The measurement results collected during these assessment cycles are very promising with high levels of precision and recall. Based on an ontology with a T-box model of 15 classes each having at least 6 properties and an A-box model with 3.000 instances the optimization approaches as described in section 3.2 have been benchmarked based on the number of meaningful results and query performance. Table 1 shows the results.

Table 1: Query optimization benchmark.

Number of results	Plain Jena	With Optimization
400	0,8 sec.	0.2 sec.
1500	1,2 sec.	0,6 sec.
2500	1,6 sec.	0,9 sec.

5 SUMMARY

The objective of the paper on hand was to show that current Semantic Web technologies can be efficiently applied in the context of content reutilization in print publishing companies. Semantic Web related ontology languages were evaluated with respect to their capability of supporting content management in print publishing firms which are challenged to handle masses of media content in heterogeneous category structures on a daily basis. Here, RDFS qualified as an adequate representation language and the subsequently evolved RDFS ontology-based proof-of-concept implementation verifies the technical feasibility not only in case of a stand-alone, but also of a distributed multi-user application system. Although this paper could only treat some aspects of content reutilization, crucial points have been addressed by focusing on the search of content modules in different contexts for the production and bundling of different media products. The findings about the application of ontologies in the media industry give enough reason to make a relevant contribution and simultaneously motivate further research in this field. In a nutshell, it can be recapitulated that semantic web technologies and tools can be fruitfully applied in media content management systems supporting the processes of searching and bundling media content in print publishing firms.

REFERENCES

Anding, M., Hess, T., 2004. Modularization, Individualization and the First-Copy-Cost-Effect – Shedding new light on the Production and Distribution of Media Content. Working Paper 1/2004, Institut für Wirtschaftsinformatik und Neue Medien der Ludwig-Maximilians-Universität, München.

Lawrence, K. F., Tuffield, M. M., Jewell, M. O., Prugel-Bennett, A., Millard, D. E., Nixon, M. S., Schraefel, M. C., Shadbolt, N. R., 2005. OntoMedia - Creating an Ontology for Marking Up the Contents of Heterogeneous Media. Proceedings of Multimedia Information Retrieval Workshop (in press), Brazil

Schek, M., 2005. Automatische Klassifizierung und Visualisierung im Archiv der Süddeutschen Zeitung. Medienwirtschaft, Vol. 2, 1 20-24

Schreiber, G., Dubbeldam, B., Wielemaker, J., Wielinga, B., 2001. Ontology-based photo annotation. IEEE Intelligent Systems, Vol. 16, 6 66-74

Schulze, B., 2005. Mehrfachnutzung von Medieninhalten: Entwicklung, Anwendung und Bewertung eines Managementkonzepts für die Medienindustrie, Josef Eul Verlag, Lohmar

Schumann, M., Hess, T., 2002. Grundfragen der Medienwirtschaft. Springer-Verlag, Berlin, Heidelberg, New York

Stamer, S., 2002. Technologie als Enabler für effizientes Cross-Media Publishing. In: Müller-Kalthoff, B. (Hrsg.): Cross-Media Management. Content-Strategien erfolgreich umsetzen. Springer-Verlag, Berlin, Heidelberg, New York 89-124.