Metadata and the Semantic Web — and CREAM[0]

[1]Siegfried Handschuh, [1,2]Steffen Staab, [1,3]Alexander Maedche

[1]Institute AIFB, University of Karlsruhe, D-76128 Karlsruhe, Germany
http://www.aifb.uni-karlsruhe.de/WBS
{sha,sst,ama}@aifb.uni-karlsruhe.de

[2]Ontoprise GmbH, Haid-und-Neu Straße 7, 76131 Karlsruhe, Germany
http://www.ontoprise.de

[3]FZI Research Center for Information Technologies,
Haid-und-Neu Straße 10-14, 76131 Karlsruhe, Germany
http://www.fzi.de/wim

"The Web is about links;
the Semantic Web is about the relationships implicit in those links."
Dan Brickley

Abstract. Richly interlinked, machine-understandable data constitutes the basis for the Semantic Web. Annotating web documents is one of the major techniques for creating metadata on the Web. However, annotation tools so far are restricted in their capabilities of providing richly interlinked and truely machine-understandable data. They basically allow the user to annotate with plain text according to a template structure, such as Dublin Core. We here survey CREAM (Creating RElational, Annotation-based Metadata), a framework for an annotation environment that allows to construct *relational metadata*, i.e. metadata that comprises class instances and relationship instances. These instances are not based on a fix structure, but on a domain ontology.

1 Introduction

The Semantic Web is about scaling the syntactic Web of HTML pages that link to each other to the Semantic Web of object identifiers and object descriptions that semantically link to each other.

Typical metadata like Dublin Core mostly consist of descriptions attributed to resources like books or journal articles (cf. Table 1, which lists the 15 elements of Dublin Core). In fact, however, some few exceptions exist, viz. Identifier, Source, and Relation. In contrast to the target Semantic Web the number of Dublin Core relationship types is limited.

[0]The full paper will appear as [2].

Table 1: Dublin Core Elements

Element	Short Explanation
Title	
Creator	
Subject	
Description	
Publisher	
Contributor	
Date	
Type	
Format	
Identifier	Example formal identification systems include the Uniform Resource Identifier (URI)
Source	Reference to a resource
Language	
Relation	Reference to a related resource
Coverage	
Rights	

The objective of our approach is to combine the power of metadata and the power of the Semantic Web and provide a method and a tool in order to construct *relational metadata*.

This paper is about a framework for facing this challenge, called CREAM[1], and about its implementation, Ont-O-Mat.

The origin of our work facing this challenge dates back to the start of the seminal KA2 intiative [1], *i.e.* the initiative for providing semantic markup on HTML pages for the knowledge acquisition community. The basic idea then was that manual knowledge markup on web pages was too error-prone and should therefore be replaced by a *simple* tool that should help to avoid syntactic mistakes.

Developing our CREAM framework, however, we had to recognize that this knowledge capturing task exhibited some intrinsic difficulties that could not be solved by a *simple* tool. We here mention only some challenges that immediately came up in the KA2 setting:

- **Consistency**: Semantic structures should adhere to a given ontology in order to allow for better sharing of knowledge. For example, it should be avoided that people confuse complex instances with attribute types.

- **Proper Reference**: Identifiers of instances, *e.g.* of persons, institutes or companies, should be unique. For instance, in KA2 metadata there existed three different identifiers of our colleague Dieter Fensel. Thus, knowledge about him could not be grasped with a straightforward query.[2]

- **Avoid Redundancy**: Decentralized knowledge provisioning should be possible. However, when annotators collaborate, it should be possible for them to identify (parts of) sources that have already been annotated and to reuse previously captured knowledge in order to avoid laborious redundant annotations.

[1] CREAM: Creating RElational, Annotation-based Metadata.

[2] The reader may see similar effects in bibliography databases. E.g., query for James (Jim) Hendler at the — otherwise excellent — DBLP: `http://www.informatik.uni-trier.de/~ley/db/`.

- **Relational Metadata**: Like HTML information, which is spread on the Web, but related by HTML links, knowledge markup may be distributed, but it should be semantically related. Current annotation tools tend to generate template-like metadata, which is hardly connected, if at all. For example, annotation environments often support Dublin Core [4], providing means to state, e.g., the name of authors, but not their IDs[3].

- **Maintenance**: Knowledge markup needs to be maintained. An annotation tool should support the maintenance task.

- **Ease of use**: It is obvious for an annotation environments to be useful. However, it is not trivial, because it involves intricate navigation of semantic structures.

- **Efficiency**: The effort for the production of metadata is a large restraining threshold. The more efficiently a tool support the annotation, the more metadata will produce a user. These requirement stand in relationship with the ease of use. It depends also on the automation of the annotation process, e.g. on the pre-processing of the document.

CREAM faces these principal problems by combining advanced mechanisms for inferencing, fact crawling, document management and — in the future — information extraction. Ont-O-Mat, the implementation of CREAM, is a component-based plug-in architecture that tackles this broad set of requirements.[4]

In the following, we sketch two usage scenarios (Section 2) and point towards some further work.

2 Scenarios for CREAM

We here only summarize two scenarios, two knowledge portals, for annotation that have been elaborated in [5]:

The first scenario extends the objectives of the seminal KA2 initiative. The KA2 portal provides a view onto knowledge of the knowledge acquisition community. Besides of semantic retrieval as provided by the original KA2 initiative, it allows comprehensive means for navigating and querying the knowledge base and also includes guidelines for building such a knowledge portal. The potential users provide knowledge, e.g. by annotating their web pages in a decentralized manner. The knowledge is collected at the portal by crawling and presented in a variety of ways.

The second scenario is a knowledge portal for business analysts that is currently constructed at Ontoprise GmbH. The principal idea is that business analyst review news tickers, business plans and business reports. A considerable part of their work requires the comparison and aggregation of similar or related data, which may be done by semantic queries like“Which companies provide B2B solutions?”, when the knowledge is semantically available. At the Time2Research portal they will handle different types of documents, annotate them and, thus, feed back into the portal to which they may ask questions.

[3] In the web context one typically uses the term ‘URI’ (uniform resource identifier) to speak of ‘unique identifier’.

[4] The core Ont-O-Mat can be downloaded from:
`http://ontobroker.semanticweb.org/annotation`.

3 Conclusion and Future Plans

CREAM is a comprehensive framework for creating annotations, relational metadata in particular — the foundation of the future Semantic Web. The framework comprises inference services, crawler, document management system, ontology guidance, and document viewers.

Ont-O-Mat is the reference implementation of CREAM framework. The implementation supports so far the user with the task of creating and maintaining ontology-based DAML+OIL markups, i.e. creating of class, attribute and relationship instances. Ont-O-Mat include an ontology browser for the exploration of the ontology and instances and a HTML browser that will display the annotated parts of the text. Ont-O-Mat is Java-based and provides a plugin interface for extensions for further advancement.

Our goal is a constant advancement of Ont-O-Mat and the CREAM framework in order to answer basic problems that come with semantic annotation.

We are already dealing with many different issues and through our practical experiences we could identify problems that are most relevant in our scenario/settings, KA2 and Time2Research. Nevertheless our analysis of the general problem is far from being complete. Some further important issues we want to mention here are:

- **Information Extraction**: We have done some first steps to incorporate information extraction. However, our future experiences will have to show how and how well information extraction integrates with semantic annotation.
- **Multimedia Annotation**: This requires considerations about time, space and synchronization.
- **Changing Ontologies**: Ontologies on the web have characteristics that influence the annotation process. Heflin & Hendler [3] have elaborated on changes that affect annotation. Future annotation tools will have to incorporate solutions for the difficulties they consider.
- **Active Ontology Evolvement**: Annotation should feed back into the actual ontologies, because annotators may find that they should consider new knowledge, but need revised ontologies for this purpose. Thus, annotation affects ontology engineering and ontology learning.

Our general conclusion is that providing semantic annotation, relational metadata in particular, is an important complex task that needs comprehensive support. Our framework CREAM and our tool Ont-O-Mat have already proved very successful in leveraging the annotation process. They still need further refinement, but they are unique in their design and implementation.

4 Acknowledgements.

The research presented in this paper would not have been possible without our colleagues and students at the Institute AIFB, University of Karlsruhe, and Ontoprise GmbH. We thank Kalvis Apsitis (now: RITI Riga Information Technology Institute), Stefan Decker (now: Stanford University), Michael Erdmann, Mika Maier-Collin, Leo Meyer and Tanja Sollazzo. Research for this paper was partially financed by US Air Force in the DARPA DAML project “OntoAgents” (01IN901C0).

References

[1] R. Benjamins, D. Fensel, and S. Decker. KA2: Building Ontologies for the Internet: A Midterm Report. *International Journal of Human Computer Studies*, 51(3):687, 1999.

[2] S. Handschuh, S. Staab, and A. Maedche. Cream — creating relational metadata with a component-based, ontology-driven annotation framework. In *Proc. of the 1st Int. Conference on Knowledge Capture*. ACM Press, 2001.

[3] J. Heflin, J. Hendler, and S. Luke. Applying Ontology to the Web: A Case Study. In *Proceedings of the International Work-Conference on Artificial and Natural Neural Networks, IWANN'99*, 1999.

[4] Dublin Core Metadata Initiative. http://purl.oclc.org/dc/, April 2001.

[5] S. Staab and A. Maedche. Knowledge portals — ontologies at work. *AI Magazine*, 21(2), Summer 2001.