Methods for Porting Resources to the Semantic Web[1]

Bob Wielinga[a] Jan Wielemaker[a] Guus Schreiber[b] Mark van Assem[b]

[a]University of Amsterdam, Social Science Informatics (SWI)
{wielinga,wielemaker}@swi.psy.uva.nl
[b] Vrije Universiteit Amsterdam, Department of Computer Science
{schreiber,mark}@cs.vu.nl

1 Thesauri and The Semantic Web

Semantic Web applications require multiple ontologies or vocabularies for indexing and querying [1]. For example, the submissions to the SW Challenge at ISWC'04 often used simple thesauri for annotation purposes[2]. However, existing resources such as the MeSH (Medical Subject Headings) thesaurus and the WordNet lexical database are often not available in RDF(S) or OWL.

In this paper, we describe a method for converting existing source material from their native representation to RDF(S) and OWL. The problem that is addressed in this method is how to convert these representations without altering the original material, and at the same time assign semantics to these representations that is (presumed to be) compatible with the intended semantics of the source.

2 Method

The method can be divided into a syntactic and a semantic stage (stages 1 and 2), which are described below. In each step decisions have to be taken with respect to the syntax or semantics of the resulting representation. A few of the guidelines supporting these decisions are mentioned in the description of the steps. An important goal of this method is to separate between "as-is" conversion (steps 1a through 2a), and specific interpretations of the thesaurus (step 2b).

The first step (step 1a) in the conversion process is a structure-preserving syntactic conversion from the source format to RDF(S). We assume that a data model of the source (e.g., XML DTD, text format, UML) is available. From the data model an RDF(S) schema is derived, where classes with properties are defined. When the source is represented in XML some elements do not have to be represented as classes when they only serve as containers for other elements. For example, the element <TermList> used in the MeSH thesaurus is only used

[1]In: C. Bussler, J. Davies, D. Fensel and R. Studer (eds.), *Proceedings of the First European Semantic Web Symposium* (ESWS2004), pp. 299–311.

[2]http://www-agki.tzi.de/swc/swc2003submissions.html

to group <Term> elements. Therefore, <TermList> can be directly mapped to the property `[DescriptorRecord] hasTerm [Term]`.

When an RDF(S) schema is established, the data elements from the source can be converted to instances of the schema. In this structural conversion step, ideally no information is lost or added.

The next step (step 1b) in the conversion process focuses on the *explication* of information that is implicit in the original data format but that is intended by the conceptual model. For example, thesauri with an origin in the bibliographic sciences are often structured as a set of records, with fields for hierarchical relations. Such a thesaurus is MeSH, which has <DescriptorRecord>s with a <TreeNumber> stated in each. These TreeNumbers can be used to create (and are intended to signify) a hierarchy, e.g. by adding a `subTreeOf` property between instances of the class Descriptor.

The second stage deals with semantic conversion. In step 2a, the RDF(S) instances generated in the syntactic stage are augmented according to the intended semantics of the source conceptual model. For example, if the properties `broaderTerm` and `narrowerTerm` are used to represent the hierarchical relation between Terms, they can be defined in OWL as inverse property of each other and as transitive properties.

In step 2b, RDFS or OWL semantics are added which may or may not be intended by the original authors (i.e., reinterpretation). One metamodeling technique for reinterpretation is defining hierarchical relations such as `subtreeOf` in MeSH to be a subproperty of `rdfs:subClassOf`. This creates an interpretation of the source as a proper subclass hierarchy. This technique is not without a price, as the resulting schema is in OWL Full. To process such a schema, an RDF toolkit is needed that is able to interpret subproperties (e.g., Triple20 [2]).

3 Case Studies

Three case studies have been performed using the AAT, MeSH and WordNet thesauri. Based on these case studies, the method has been extended and applied to MeSH and WordNet. Full conversions of MeSH and WordNet are available at `http://thesauri.cs.vu.nl/` for use in Semantic Web applications.

References

[1] A. Th. Schreiber, B. Dubbeldam, J. Wielemaker, and B. J. Wielinga. Ontology-based photo annotation. *IEEE Intelligent Systems*, 16(3):66–74, 2001.

[2] J. Wielemaker, A. Th. Schreiber, and B. J. Wielinga. Prolog-based infrastructure for RDF: performance and scalability. In D. Fensel, K. Sycara, and J. Mylopoulos, editors, *The Semantic Web - Proceedings ISWC'03, Sanibel Island, Florida*, pages 644–658, 2003.