OntoSelect: Towards the Integration of an Ontology Library, Ontology Selection and Knowledge Markup

Paul Buitelaar

DFKI Language Technology
Stuhlsatzenhausweg 3,
D-66123 Saarbrücken, Germany
paulb@dfki.de

A central task in the Semantic Web effort is the annotation of data and documents with appropriate semantic information (i.e. *knowledge markup* or *ontology population*) derived from one or more ontologies published on the Semantic Web. The added knowledge allows automatic procedures (agents, web services, etc.) to interpret the underlying data and/or documents in a unique, formally specified way, thereby enabling autonomous information processing.

Most of the current work in knowledge markup is concerned with annotation of concepts relative to a particular ontology that is typically developed specifically for the task at hand. Instead, a more realistic approach would be to access an ontology library and to select one or more appropriate ontologies. Although the large-scale development and publishing of ontologies is still only in a beginning phase, many are already available (see e.g. the DAML ontology library[1], OWL ontology library[2], or SchemaWeb[3]). To select the most appropriate ontology (or a combination of complementary ontologies) will therefore be an increasingly important subtask of knowledge markup.

Here we present an approach towards an integration of the collection and classification of ontologies in a dynamic web-based ontology library, methods for the selection of an ontology from this library and its use in knowledge markup. Building on the idea of the DAML and SchemaWeb ontology libraries, we aim to take this to its fullest consequence through the construction of a fully dynamic ontology library (OntoSelect) that will be updated continuously, organized in a meaningful way and with automatic support for ontology selection in knowledge markup.

The OntoSelect approach aims at providing an access point for ontologies on any possible topic or domain. However, unlike these libraries, OntoSelect is not based on a static registration of published ontologies, but instead includes a dynamic ontology crawling procedure that monitors the web for any newly published ontologies in the representation formats: RDF/S, DAML or OWL.

Collected ontologies are analyzed using the OWL API[4] that allows for the extraction of structure and content of any RDF/S, DAML or OWL ontology. There are cur-

[1] http://www.daml.org/ontologies/
[2] http://protege.stanford.edu/plugins/owl/ontologies.html
[3] http://www.schemaweb.info/
[4] http://owl.man.ac.uk/api.shtml

rently around 800 ontologies in the OntoSelect library, covering a wide range of topics and domains. Ontologies are stored in a database and are organized according to: format; ontology-, class- and property-names; class- and property-labels. The assignment of labels is unfortunately not so wide spread. However, specifically from the semantic annotation and knowledge markup perspective this is an important aspect, as automatic annotation or markup of documents crucially depends on the availability of terminology for classes and/or properties.

OntoSelect provides a dynamically updated library of ontologies that may be used in a knowledge markup process. However, as there is a rapidly increasing number of published ontologies available, it is becoming a more and more difficult task to select the most appropriate one(s). To provide semi-automatic support for this, OntoSelect includes a functionality for selecting ontologies for a given knowledge markup task, based on the following criteria that address ontology content and structure:

- *Coverage*: How many of the terms in the document collection of the particular knowledge markup task are covered by the classes and properties in the ontology?
- *Structure*: How detailed is the knowledge structure that the ontology represents?
- *Connectedness*: *Is the ontology connected to other ontologies and how well established are these?*

After selection of an appropriate ontology from the OntoSelect ontology library, a document collection under consideration will be marked up with the knowledge from this ontology. We are currently working towards an instance-based learning approach that considers knowledge markup as a classification task. Classifiers for the knowledge markup process will be generated by collecting occurrences (i.e. linguistic realizations of classes and properties: labels or class-/property-names with their linguistic contexts) from relevant text collections that are to be associated with each of the ontologies in the OntoSelect library.

A central problem to be addressed in this is the extraction of relevant terms in text and their appropriate classification by the constructed classifier. Additional problems that are to be addressed include multilinguality (e.g. the use of an English-based ontology in knowledge markup of German documents) and ambiguity (e.g. multiple definitions of the same concept in several ontologies or multiple use of the same label for different concepts within one ontology).

Acknowledgements

This research has been supported by research grants for the SmartWeb and VieWs projects. Thanks to Thomas Eigner and Srikanth Rmaka for their work on the OntoSelect framework.