Document Management using Protégé

Henrik Eriksson

Dept. of Computer and Information Science
Linköping University
SE-581 83 Linköping, Sweden

her@ida.liu.se

Introduction

Document management supported by organizational document repositories is an important tool for knowledge management. In addition, document annotation can support processing and retrieval of documents. However, metadata are often restricted to simple expressions, such as a list of keywords for each document, and it is difficult to describe complex terms and relationships. Unfortunately, there is little support for annotation of documents with ontologies and the use of ontologies as the basis for representing the content of document repositories.

It is possible to use Protégé [1] as a platform for document management. We believe that there is much to be gained from using OWL-based ontologies and Protégé as the basis for document-repository organization and management [2]. In this work, we use the annotation-handling facilities of a Protégé extension to develop a knowledge base that acts as a document repository. The documents in this repository are annotated using OWL individuals that model annotations of major document parts. The resulting knowledge base makes the annotations of all documents in the repository available as OWL individuals to the Protégé user. An important advantage of this approach is that it is possible to use other Protégé extensions to search and visualize the repository.

Background

There is a significant gap between knowledge expressed in textual documents and knowledge encoded in knowledge-representation formats, such as ontologies. The goal of *semantic documents* is to bridge this gap by combining printable documents with knowledge representation formats. This approach is similar to the semantic web in that the documents contain ontologies as metadata. However, the semantic-document approach emphasizes electronic documents that print well and document annotations that link well-defined text areas to classes and individuals in ontologies.

Currently, we are working with PDF files that contain OWL-based annotations. PDF is an open format in the sense that it is published [3]. Furthermore, there are many third-party tools for creating, managing, and viewing PDF documents. One of the advantages of PDF is that this format allows us to add metadata to the document structure. Normal PDF documents, which users can view and print using standard tools, can carry such metadata.

The document extension PDFTab allows Protégé users to annotate standard PDF documents with OWL-based mark up [4,5]. This extension runs Adobe Acrobat in a Protégé tab and provides functionality for highlighting text and other document areas and linking them to individuals in the Protégé ontology. Semantic documents and the PDFTab extension provide the foundation for ontology-based document management [2]. From an organizational perspective, however, it is important to link enterprise ontologies to document repositories. We have explored how it is possible to use Protégé as a document-management tool.

Method

We have collected on-line official publications by Statistics Sweden from 2006 and annotated them using an automated annotation tool. This custom-tailored annotation tool traverses the documents and identifies key headings and tables. The annotated documents were then loaded into PDFTab to create a knowledge base that represents the documents and all their annotations as OWL individuals. The document set consists of all statistics reports published in PDF in 2006 on the Statistics Sweden web site. In addition, we used the five volumes of the Statistical Yearbook (published annually from 2002 to 2006). The total number of automatically-annotated documents is 302. The combined number of annotations for these documents is 17,470.

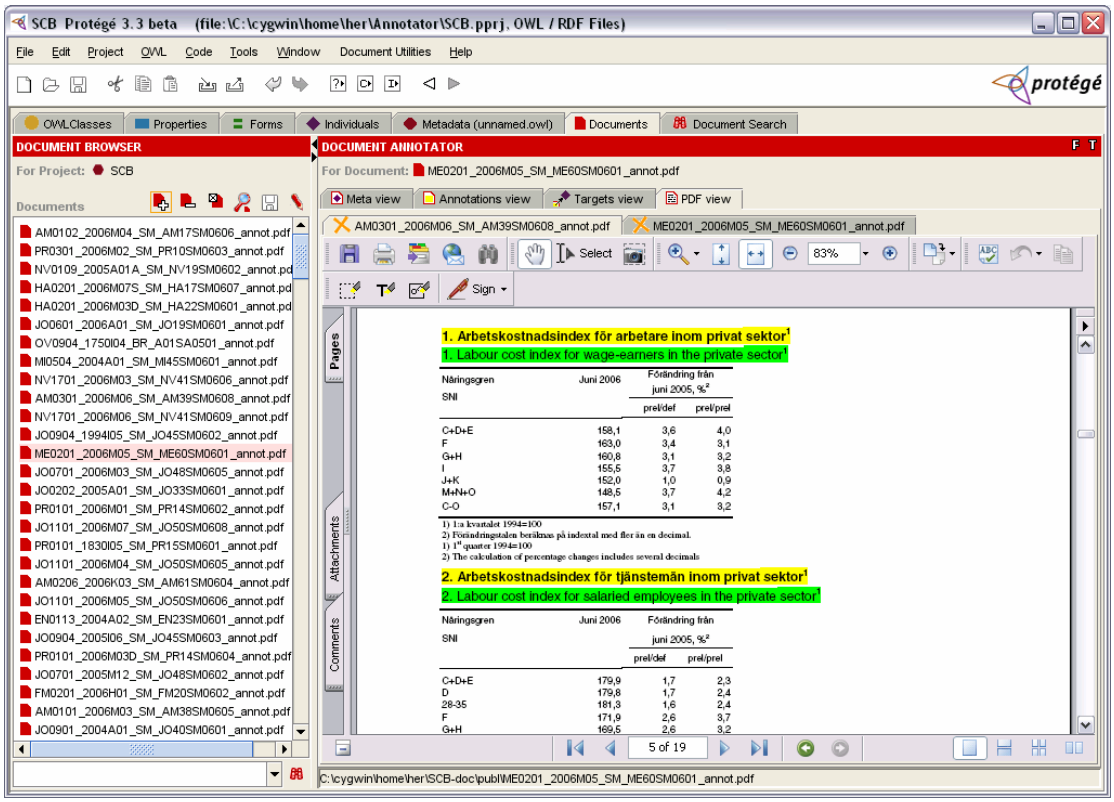

Figure 1. Protégé with the PDFTab extension. The left-hand side of the “Documents” tab shows a list of the loaded documents. The right-hand side of the tab is the PDF view, which shows the selected document with annotations.

Results

The resulting Protégé knowledge base consists of an *annotation ontology* that defines classes for the annotation types, such as PDFTextAnnotation, and a *document ontology* that defines key concepts for the publications, such as Table and Diagram. Furthermore, the knowledge base contains references to the annotated documents, in this case individuals of the DocumentReference class. When Protégé loads the knowledge base, it will first load the project ontology and then load the metadata for each of the referenced documents. The Protégé user will then have access to all the documents and their annotation individuals. Figure 1 shows the document view added by the PDFTab extension. In this tab, the user can browse the documents and examine the annotations in the PDF view. In addition, we have developed a tab extension for document search. This search function searches both the annotations and the full text of the documents.

Summary and Conclusions

In this approach, we use Protégé as a document-repository manager. This work illustrates that it is possible to use Protégé with the document extension PDFTab to manage collections of semantic documents. One of the advantages of this approach is that it enables us to develop semantic search services that access the ontology though the Protégé API. However, the performance of file-based OWL storage backend limits the number of annotated documents that can be effectively used. For example, it takes about 1.5 hours to load the complete repository of 302 annotated documents into Protégé on a standard PC. Nevertheless, we believe that it is possible to develop a database solution that acts as both a Protégé storage backend and a document repository, and that it is possible to scale this architecture to large document repositories.

Acknowledgements

This work was supported by Vinnova (grant no. 2003-01415), by The Swedish Research Council (grant no. 621-2003-2991), and by Statistics Sweden.

References

[1] John H. Gennari, et al. The evolution of Protégé: An environment for knowledge-based systems development. *International Journal of Human–Computer Studies*, 58(1):89–123, 2003.

[2] Henrik Eriksson and Magnus Bång. Towards document repositories based on semantic documents. In *Proceedings of the Sixth Conference on Knowledge Management*, I-KNOW 2006, pages 313–320, Graz, Austria, September 6–8, 2006.

[3] Adobe. PDF Reference Version 1.6. Adobe Press, Berkeley, CA, 5th edition, 2004.

[4] Henrik Eriksson. The semantic document approach to combining documents and ontologies. *International Journal of Human–Computer Studies*, in press.

[5] Henrik Eriksson. Support for semantic documents in Protégé. In *Proceedings of the Eight International Protégé Conference*, Madrid, Spain, July 18–21, 2005.