Chapter VI
Semantic Annotation and Ontology Population

Florence Amardeilh
Mondeca, France and Université Paris 10, France

ABSTRACT

This chapter deals with issues related to semantic annotation and ontology population within the framework defined by the Semantic Web (SW). The vision of the Semantic Web, initiated in 1998 by Sir Tim Berners-Lee, aims to structure the information available on the Web. To achieve that goal, the resources, textual or multimedia, must be semantically tagged by metadata so that software agents can utilize them. The idea developed in this chapter is to combine the information extraction (IE) tools with knowledge representation tools from the SW for the achievement of the 2 parallel tasks of semantic annotation and ontology population. The goal is to extract relevant information from the resources based on an ontology, then to populate that ontology with new instances according to the extracted information, and finally to use those instances to semantically annotate the resource. Despite all integration efforts, there is currently a gap between the representation formats of the linguistic tools used to extract information and those of the knowledge representation tools used to model the ontology and store the instances or the semantic annotations. The stake consists in proposing a methodological reflexion on the interoperability of these technologies as well as designing operational solutions for companies and, on a broader scale, for the Web.

INTRODUCTION

This chapter deals with semantic annotation and ontology population in the context of the Semantic Web (SW). The aim of the Semantic Web is to structure information available on the Web. To achieve that goal, resources, both textual and multimedia, must be semantically enriched with metadata to allow software agents to use them. Explicit representation of the contents of
Web-based resources is enabled by ontologies. Ontologies play a major part in semantic annotation, since they define the concepts, attributes and relations used to annotate the document content. The ontology constrains the vocabularies and knowledge instances allowed as metadata for any given application. For example, an article may be annotated by a metadata “author” which the value should be an instance of the concept “Person”. But if it is essential for a Semantic Web application to rely on an ontology for the realization of this semantic annotation task, it is also important that the knowledge base, associated with this ontology, contains the instances to be used for semantic annotation. For example, the value of the metadata “author” might be “Florence Amardeilh”, an instance of the concept “Person” in a particular knowledge base. This is why the purpose of the ontology population task is to enrich (semi-)automatically the knowledge base with new instances of the concepts, attributes and relations defined by the ontology model.

First the chapter proposes definitions of semantic annotation and ontology population (section 2). It then describes each dimension of semantic annotation, a dimension in turn being composed of a set of characteristics. Those dimensions must be carefully taken into account when implementing a new solution for semantic annotation and ontology population. The differences between these two tasks are highlighted as well as their mutual interaction to generate new content for applications.

Second, the chapter presents the current state-of-the-art of semantic annotation and ontology population solutions, from both research and business perspectives (section 3). The capabilities of these solutions and their limitations are explained in order to indicate the future challenges to be solved in these fields during the next few years.

Third, the chapter presents a framework composed of modular software components, allowing maximum flexibility with respect to the needs and objectives for a new application (section 4). The chapter highlights the problems resolved by these components, in particular those from the annotation and instances consolidation task and from the maintenance of the different terminologies.

Finally, the chapter illustrates how the proposed framework can be used in a complete project involving information extraction, terminological and ontological resource enrichment, semantic annotation and terminology update (section 5). It describes a methodology in five stages, based on the software engineering recommendations. It aims to provide simple and effective instructions for the realization of a concrete semantic annotation or ontology population application within a company. It eases the burden of creating such an application by defining the roles of the different actors, the objectives of each step, the actions, the means and the set of deliverables needed to assess the success of the target system. It has already been successfully put into practice in various semantic annotation and ontology population projects.

The main objectives of the chapter are:

- To provide a clear description of Semantic Annotation and its close relationship to Ontology Population;
- To illustrate the components required to create a complete semantic annotation application;
- To provide a simple but efficient methodology to assist users in implementing such an application in an industrial context.

**SEMANTIC ANNOTATION VERSUS ONTOLOGY POPULATION**

The word *annotation* is a derivative of the Latin term ‘*annotare*’, meaning “to note; to annotate”, i.e. “to accompany a text with annotations, remarks, comments”. Therefore, an annotation corresponds to “the action of annotating, with the result of this action as being a critical or explana-