Available online at www.sciencedirect.com
ScienceDirect
Int. J. Human-Computer Studies 65 (2007) 624–639
International Journal of Human-Computer Studies
www.elsevier.com/locate/ijhcs

The semantic-document approach to combining documents and ontologies

Henrik Eriksson*

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

Available online 10 April 2007

Abstract

An ontology is a powerful way of representing knowledge for multiple purposes. There are several ontology languages for describing concepts, properties, objects, and relationships. However, ontologies in information systems are not primarily written for human reading and communication among humans. For many business, government, and scientific purposes, written documents are the primary description and communication media for human knowledge communication. Unfortunately, there is a significant gap between knowledge expressed as textual documents and knowledge represented as ontologies.

Semantic documents aim at combining documents and ontologies, and allowing users to access the knowledge in multiple ways. By adding annotations to electronic-document formats and including ontologies in electronic documents, it is possible to reconcile documents and ontologies, and to provide new services, such as ontology-based searches of large document databases. To accomplish this goal, semantic documents require tools that support both complex ontologies and advanced document formats. The Protégé ontology editor, together with a custom-tailored documentation-handling extension, enables developers to create semantic documents by linking preexisting documents to ontologies.

Keywords: Document; Ontology; Knowledge representation; Annotation; Metadata

1. Introduction

Ontologies provide a framework for conceptualization and knowledge modeling in a multitude of areas (Gruber, 1993). In the last decade, ontologies have emerged as one of the most popular modeling approaches for taxonomies, classifications, and other structures used in intelligent systems. Languages, such as resource description framework (RDF), RDF schema (RDFS), and ontology web language (OWL), are the foundation of semantic-web efforts to use ontologies for web services (World Wide Web Consortium, 2004a–c). Furthermore, development environments, such as Protégé (Gennari et al., 2003), provide several tools for building ontologies. Despite this progress, however, there are many areas in which ontologies have not yet reached their full potential in terms of utility and applications, such as integration with other types of personal and organizational information systems. A significant bottleneck is the lack of integration with other forms of knowledge expression. In particular, ontologies must coexist with written definitions and descriptions to ensure, for example, traceability, appropriate documentation, and justification of expressions in the ontology.

Historically, documents have been the key carrier of human knowledge and they continue to be an important medium in the age of electronic communication and the world-wide web. Like ontologies, a major role of documentation is to describe concepts, ideas, and phenomena and their relationships. Each day, millions of people use computers as enhanced typewriters to produce documents, and the number of authors well exceed the number of ontology developers.

It is easy to forget the significance of documents when developing ontologies. Unfortunately, there is a surprisingly large gap between the knowledge modeled in ontologies and the text documenting the same knowledge. In general, authors produce a document for certain

*Tel.: +46 13282673; fax: +46 13142231.
E-mail address: her@ida.liu.se.

1071-5819/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ijhcs.2007.03.008

purposes, such as communicating ideas and instructions to humans, whereas ontology developers define ontologies for other purposes, such as automated classification and reasoning (Uschold and Jasper, 1999). Current approaches make it difficult to use ontologies and documents in concert and as two views of the same knowledge. The tools available tend to support either ontology editing or document manipulation. For example, it is not difficult to relate classes and individuals[1] in an ontology to sections, regions, paragraphs, words, and so forth in a document.

It is possible to distinguish between two major areas where it is useful to integrate documents and ontologies. One possibility is to *annotate documents* with ontologies to add metalevel information and to provide ontological structures for document content. Here, ontologies describe entire documents and explain document parts, such as words and phrases. The semantic-web approach, for example, aims at supporting annotation of web pages with ontologies (Berners-Lee et al., 2001; Handschuh and Staab, 2003). Another possibility is to *document ontologies*; that is, to create documentation that describes different aspects of the ontology content and its development. As organizations develop more and more ontologies, the documentation of these ontologies becomes increasingly important. Many applications require a printed version of the knowledge content in a predetermined report format. Furthermore, it is sometimes difficult for domain experts to review ontologies using only computer-based tools for ontology editing and visualization. Printed versions of an ontology complement the alternative perspectives of the content provided by interactive ontology-visualization tools. Moreover, preexisting documents combined with ontologies can support knowledge management (Eriksson and Bång, 2006). For example, large document repositories can benefit from ontologies that facilitate search and retrieval. Thus, we need representation and communication formats for knowledge that adhere to both human reading and machine processing.

The semantic-document approach attempts to reconcile documents and ontologies by extending printable documents with annotations and additional knowledge bases. The ultimate goal of semantic documents is not merely to provide metadata for documents, such as keywords and Dublin core descriptions (Weibel et al., 1998), but to integrate documentation and knowledge representation to the point where they use a common structure, which provides both documentation and representation views. In our approach, Adobe's portable document format (PDF) (Adobe, 2004a) is the basis for semantic documents, which stores both a printable document and the related knowledge base as a single file. The OWL forms the basis for the ontology representation in this combined format. Alternatively, if description logic is not required, it is possible to use RDF/RDFS (which is a basis for OWL). The Protégé ontology editor, together with our plug-in extensions for supporting PDF, provides an annotation and preparation environment for semantic documents. We have applied this approach to the statistics and clinical-guideline domains and used the tools to annotate existing documents and to generate new semantic documents.

This paper is organized as follows. Section 2 provides the background in terms of ontology and document technologies. Section 3 introduces the semantic-document approach and the basic technology supporting it. Section 4 describes modeling with semantic documents. Section 5 presents applications of semantic documents. Section 6 discusses the pros and cons of the semantic-document approach and its relationship to the semantic web. Finally, Section 7 draws conclusions.

2. Background

Semantic web and knowledge management are two large areas related to the semantic-document approach. In addition, there is relevant previous work in terms of specific approaches to document-supported knowledge acquisition and support for active documents. Let us discuss these approaches before proceeding with semantic documents.

2.1. Semantic web

The *semantic web* aims at enabling communication between machines on the world-wide web (Berners-Lee et al., 2001). One of the major applications of the semantic web is to improve metalevel descriptions of the content of hypertext markup language (HTML) documents. The RDF and OWL languages allow web-site developers to define ontologies that add semantic content to HTML pages. Semantic-web search engines, such as Swoogle (Ding et al., 2004), can then use these metadata to provide ontology-based search services.

The task of annotating a significant proportion of the entire web is certainly formidable. Therefore, it is too difficult to achieve a critical mass of annotated web pages for semantic search engines to be effective. For smaller sets of documents and specific document categories, however, the annotation of a sufficient number of documents is much easier. For example, it is possible to annotate a corporate intranet (e.g., thousands of web pages) and to create a semantic search service for the internal sites. For mathematical documents, there are several extensible markup language (XML)-based markup languages, such as MathML (World Wide Web Consortium, 2003a), OpenMath (Buswell et al., 2004), and OMDoc (Kohlhase, 2006). Likewise, it is possible to annotate a specific category of pages published by multiple providers, such as product offerings and prices, if these providers can agree on a common annotation policy.

The semantic-document approach is similar to the semantic web in that it adds ontologies to textual documents. Just like in the semantic web, RDF/RDFS

[1]Following OWL terminology, the term *individuals* denotes instances of a class.