Chapter II
Introduction to Ontology Engineering

Paolo Ceravolo
University of Milan, Italy

Ernesto Damiani
University of Milan, Italy

ABSTRACT

This chapter provides an introduction to ontology engineering discussing the role of ontologies in informative systems, presenting a methodology for ontology design, and introducing ontology languages. The chapter starts by explaining why ontologies are needed in informative systems, then it introduces the reader to ontologies by leading him/her in a stepwise guide to ontology design. It concludes by introducing ontology languages and standards. This is a primer reading aimed at preparing novice readers of this book to understanding more complex dissertations; for this reason it can be avoided by expert readers.

CHAPTER SUMMARY

In this chapter, we introduce the fundamental notions of *ontology* and *ontology engineering* and explain why they turned out to be of paramount importance in the development of the new generation of Web-based systems that will compose the Semantic Web (Berners-Lee, 2001). Then, we focus on the emerging discipline of *ontology design*, which considers the design of ontologies for specific application *domains*. Our aim is to give some simple, yet general guidelines toward designing ontologies and managing their entire life cycle, including their design, validation, and deployment.

WHAT IS WRONG WITH INFORMATIVE SYSTEMS?

The ideas that we shall discuss in this chapter do require a bit of intellectual effort, so it makes sense to try to build some motivation before delving into them. In order to convince the reader that this effort is actually worth making, let us informally consider some unexpected consequences of the Internet-based information explosion.

The availability of a pervasive, global infrastructure for information distribution such as the World Wide Web, together with the dramatic decline in communication and storage costs per information unit, has greatly encouraged information production.

Government agencies, corporations, consortia, and a host of other organizations keep themselves busy producing all sorts of documents, databases, and spreadsheets, not to mention computer programs and their documentation.

Then, they publish their efforts on internal Intranets or on the global Net and wait for their employees, customers or someone else to take advantage of them.

Well-intentioned as these organizations may be when producing information, the sheer size of the data they publish every day may make those data practically useless.

Internet technology has been successful in boosting productivity; inasmuch it provided the means for information to reach its intended destinations quite easily, but processing such information (and deciding if it is relevant) is becoming more and more of a chore for the recipients. As a result, organizations around the world are discovering that using the Web or e-mail for knowledge distribution has become a rather unreliable and time-consuming process.

Few people nowadays bother to read unsolicited e-mail messages, even if the message's sender belongs to their organization and has a legitimate motive for sending it. Referring people to information posted on internal or public Web sites turned out to be also not very effective, as few users have enough time for checking the sites that might contain useful information regularly.

On the other hand, the very same people that eschew dealing with information directed at them may well later be found using search engines to look for information they need on the Web, often without much success.

The development of hypermedia has made Web search an intermediate situation between information seeking and content learning, where specific abilities are needed to obtain even modest results.

As expert users of Web-based search engines know very well, attempts at indexing and retrieving information using keyword-based indexing techniques have met only with partial success, and WWW content location is still more an art than a science.

In order to convince ourselves of the drawbacks of keyword-based indexing of network resources, we shall use an example.

Suppose that you are a masters student of modern literature searching the Web for information on the making of the Romantic novel in 19th-century Europe. For some reason, you ignored the lecture notes that were produced by your lecturer and sent to you weeks before via e-mail. Tomorrow, you are supposed to give a talk on this subject and the course Web site is down. So, conventional *direct* inquiry to someone else (either synchronously via a phone call, or asynchronously via an e-mail message) is out of the question; all you can do is to adopt a marketplace style of inquiry, and look for the information yourself.

Soon, you discover that feeding the keywords "romantic" and "novel" into a keyword-based WWW search engine will not solve your problem.

- First of all, while the search is likely to produce many results, not all of them will be relevant to your query. In other words, the *precision* of the search result, that is, the