Contextual Metadata for Document Databases

Virpi Lyytikäinen  
*University of Jyväskylä, Finland*

Pasi Tiitinen  
*University of Jyväskylä, Finland*

Airi Salminen  
*University of Jyväskylä, Finland*

**INTRODUCTION**

Metadata has always been an important means to support accessibility of information in document collections. Metadata can be, for example, bibliographic data manually created for each document at the time of document storage. The indexes created by Web search engines serve as metadata about the content of Web documents. In the semantic Web solutions, ontologies are used to store semantic metadata (Berners-Lee et al., 2001). Attaching a common ontology to a set of heterogeneous document databases may be used to support data integration. Creation of the common ontology requires profound understanding of the concepts used in the databases. It is a demanding task, especially in cases where the content of the documents is written in various natural languages. In this chapter, we propose the use of contextual metadata as another means to add meaning to document collections, and as a way to support data integration. By contextual metadata, we refer to data about the context where documents are created (e.g., data about business processes, organizations involved, and document types). We will restrict our discussion to contextual metadata on the level of collections, leaving metadata about particular document instances out of the discussion. Thus, the contextual metadata can be created, like ontologies, independently of the creation of instances in the databases.

**BACKGROUND**

European legal databases offer an example of a new means for data integration. Due to the development towards European integration, the legal information needed in a particular situation often concerns not only regulations of the home country, but also European Union regulations and those in foreign countries. The information may be scattered in various European legal databases with varying retrieval techniques. The databases are organized in different ways, and their content is written in different languages. Differences in legal systems aggravate the retrieval problems. Similar problems and needs may be identified in other types of environments, as well. For example, the information needed during manufacturing processes may be created in a number of cooperating organizations and scattered in heterogeneous intranet and extranet repositories.

Where the creation of ontologies requires analysis and a description of concepts used on a domain, creation of contextual metadata requires analysis of the environment where documents are created. We will first describe methods for collecting contextual metadata, and then we will show how the metadata can be visualized to users in a graphical interface. The graphical models providing contextual metadata aid the users in understanding the context of documents and in locating information from correct sources.

**COLLECTING THE CONTEXTUAL METADATA**

A special methodology called Rakenteisten Asiakirja Standardien KEhittäminen (RASKE) meaning “Development of standards for structured documents,” has been developed for analyzing and describing document management environments. The RASKE method was tailored from more general information systems analysis methods and tested in a project where document management practices in the Finnish Parliament and ministries were redesigned (Salminen, 2003; Salminen et al., 2000; Salminen et al., 2001). The major practical result of the project is that the Finnish Parliament currently creates all Parliamentary documents in SGML (Standard Generalized Markup Language) format (Goldfarb, 1990). The RASKE methodology offers tools for gathering and representing contextual and structural metadata about documents.

The modeling techniques of RASKE are intended for describing documents, processes where documents are created and manipulated, and actors and their roles in the