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.

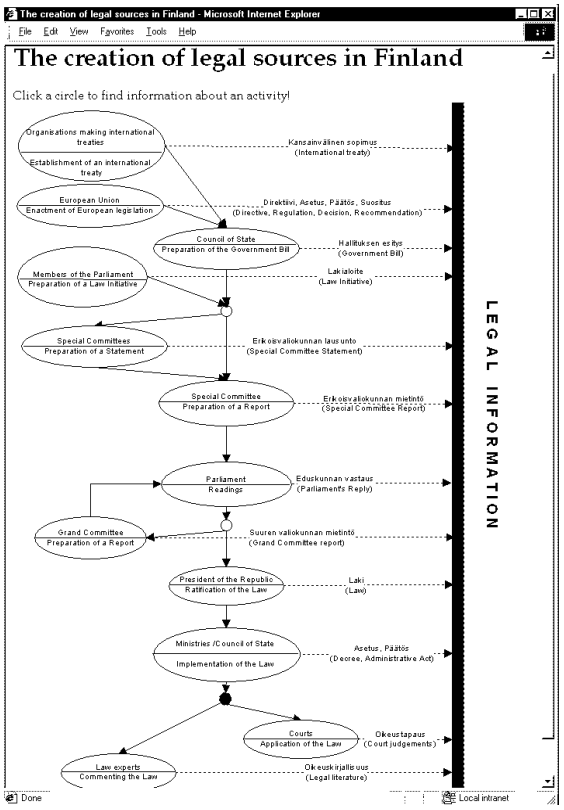
The modeling techniques of RASKE are intended for describing documents, processes where documents are created and manipulated, and actors and their roles in the

activities of the processes. The most important origins of the techniques are in the object-based modeling methodology of Shlaer and Mellor (1992), in Information Control Nets (Ellis, 1979), and in the document structure specification methods of SGML standard and elm graphs (Goldfarb, 1990; Maler & El Andaloussi, 1996). From the various models created in a RASKE analysis, three kinds of models provide important contextual information. The *organizational framework model* describes the organizations involved and their roles on the analysis domain. The activities of the domain are described by process models. The *document output model* is a process model showing the activities and actors of a process together with the document types produced in the activities. The relationships of different document types are described by a *document-relationship diagram.* In addition to a set of models, the RASKE methodology produces textual descriptions (e.g., about document types and actor roles).

VISUALIZATION OF THE CONTEXTUAL METADATA

To support users in retrieving information created on a complex domain, we suggest using the models created during the document analysis as metadata visualizations. The visualization of information itself is not a novel issue. Graphical models have been used, for example, in software engineering (e.g., UML) (Booch et al., 1999), business process redesign (Abeysinghe & Phalp, 1997), workflow management systems (Ellis, 1983), and computer supported cooperative work systems (Sarin et al., 1991). The visualizations we propose add, however, a new perspective for the uses of visual presentations. Our solution for visualizing contextual metadata by graphical models is intended to be used as a retrieval interface for distributed document repositories. The solution has been tested in the European User Views to Legislative Information in Structured Form (EULEGIS) project, whose main goal was to improve access to European legal information (Lyytikäinen et al., 2000). The EULEGIS prototype system demonstrated how a single-point access to various European legal databases could be implemented.

Figure 1. Process view describing the Finnish legal system



The three models discussed previously provide three different views to the information on the domain. The *actor view* describes the most significant actors on the domain; the *information source view* shows the different kinds of documents and their relationships; and the *process view* describes activities related to the information production process. In order to enable querying the document databases, links from each graphical element of the view to the query form of the relevant database can be defined.

The process view describing the Finnish legal system is shown in Figure 1. The graphical model originates from the document output model created during the RASKE analysis. The main activities of the process are depicted by circles. Each circle shows both the name of the activity and the actor(s) performing the activity. The order of activities is expressed by solid arrows between the activities. The information flow from an activity to the collective document repository is shown by a dashed arrow labelled with the type of the documents created in the activity. Additional textual information is linked to the graphical symbols. By clicking, for example, an activity symbol in the graph, the user can obtain more information about the activity. From this additional information, a link leads to a search form by which the user can search for the documents that originate from the selected activity. Similarily, the user can retrieve information about a particular actor and a search form for accessing documents created by the actor (see Figure 2).

The three views were implemented and tested in the EULEGIS prototype. The feedback from the users was encouraging. The views were seen as a good tool to become acquainted with a previously unfamiliar legal system and to limit the scope of a query statement.

For the dynamic generation of graphical views, the data pertaining to the models has to be formally defined.