Achieving Semantic Interoperability Through Controlled Annotations

Position Paper
Michael Gertz
Department of Computer Science
University of California, Davis, U.S.A.
gertz@cs.ucdavis.edu
www.db.cs.ucdavis.edu

We consider the main goal of a Digital Library as to provide individual users and collaborative research groups with a uniform and integrated access to a large collection of heterogeneous data. The major hindrance in creating a respective infrastructure is to integrate relevant data residing at different sites and encoded in different formats. Although recent developments in high-speed networks and data protocols have achieved a reasonably high degree of *physical connectivity*, that is, the fast exchange of bits and bytes among computer systems, the *logical connectivity*, i.e., the *meaningful* exchange and querying of data, is far behind what is needed in current information system infrastructures for Digital Libraries.

Describing data using metadata has proven to be useful not only for integrating data from different sources, but also for better information retrieval methods. Different types of metadata schemas have been proposed to enrich data by more semantics, which can then be exploited in diverse data processing tasks. As practice shows, however, metadata is rarely used. For example, less than 3% of the data available on the Web utilizes Dublin Core as a simple content description model for electronic resources. The role of metadata is even more crucial in collaborative research environments where research groups at different sites make their data accessible on the Web but also want to utilize data from other research groups. In such environments, the "right" employment of metadata creation and usage schemes is crucial for building Digital Libraries that provide effective access to heterogeneous types of domain specific data created at different sites.

In our current research we are developing methods, concepts and tools that try to alleviate the above problems by realizing a semantic rich layer on top of heterogeneous Web-based sources. For this, we employ domain specific ontologies, providing an effective means to conceptualize a domain of interest, and data annotation techniques, providing a more sophisticated means to associate *controlled* metadata with data. The core idea of the approach is that researchers build and use ontologies for conceptualizing domain specific concepts (knowledge) and relationships among concepts. A concept basically consists of three components: terms (synonyms, preferred terms etc.), definition(s), and properties. An ontology thus provides some kind of standard vocabulary, extended by semantic rich relationships among concepts and terms. Concepts represented in an ontology are then used to annotate (remote) data represented on the Web, e.g., in form of a domain specific Digital Library. Thus, an annotation provides a well-defined semantic carrying link between data and metadata (concepts). Annotations can be associated with Web-based data at different levels of granularity, ranging from complete Web sites to specific

portions of a Web page. Both, the modeling of an ontology and the annotation of data residing at Web sites is based on concepts provided by XML and associated technologies, e.g., XLink and XPointer. For querying and retrieving information of interest, a domain specific ontology is used as a uniform and semantic rich repository of links to and among diverse types of Web based information sources. In context of this workshop, the major challenge this approach faces is to build information system infrastructures that support the integration of ontologies developed for the same domain but in different languages. Assuming such an infrastructure can be built by carefully specifying relationships among concepts in the ontologies (i.e., on the metadata level), it would be possible to access cross-language data through cooperatively developed ontologies.

We currently employ the ontology creation and data annotation techniques described above in several research projects. With the Center for Neuroscience at UC Davis, we are working on methods to link images (and portions thereof) of the human brain with neuroscience concepts modeled in a domain specific ontology. With the Department of Classics and Religious Studies, we are working on data annotation schemes that help researchers to semantically enrich and link a large collection of Judeo-Spanish oral literature.

Regarding the methods proposed above, we hope to have discussions, suggestions, and comments on the following issues

- Is it reasonable to assume that research groups or Web communities with a common interest in a particular topic are willing to create ontology like structures? What are the necessary and required components of such an infrastructure? Does XML and related technologies (e.g., Unicode) provide a reasonable basis to create and maintain cross-language ontologies and annotation techniques?
- Are creators of information sources (as part of a Digital Library) willing to semantically enrich the information they create and maintain? What support do they expect?
- Are users of Digital Libraries willing to make use of annotations, and are they also willing to add they own annotations?
- What are typical behavior models users employ to retrieve information from a Digital Library?
- Does the so-called semantic Web provide an appropriate layer to operate on Digital Libraries? How can the semantic Web be extended in order to take different languages into account?

Biography

Dr. Michael Gertz obtained his Ph.D. from the University of Hannover, Germany, in 1996. Since 1997 he is an assistant professor at the Department of Computer Science at the University of California, Davis. He is a member of the Graduate Groups in Computer Science and Medical Informatics, and he is the principal investigator of the Database and Information Systems Group. His research interests include various aspects on data integration techniques, interoperability among heterogeneous (Web-based) information systems, and data quality and integrity. His current focus is on data integration techniques

that utilize XML technology for data annotation schemes to enrich data extracted from the Web by semantic carrying metadata. Particular aspects include HTML conversion techniques, schema discovery approaches, and querying XML data. Dr. Gertz's research interests also include database and information systems security with a particular focus on authentic data publication schemas, trusted query mediation, and security reengineering for information systems through user and application profiling.