Data Annotations in Collaborative Research Environments

Michael Gertz
Department of Computer Science
University of California at Davis
One Shields Avenue, Davis, CA 95616-8562, USA

gertz@cs.ucdavis.edu
http://www.db.cs.ucdavis.edu

In several of today's scientific application domains, in particular the computational sciences, the transparent and integrated access to distributed and heterogeneous collections of scientific data is key to leveraging the knowledge and findings of researchers. Standard database integration approaches, however, are either not applicable or insufficient due to the lack of local and global (database-like) schema structures. In such domains, data integration often occurs "manually" in that remote data is copied into local repositories or "semantically indexed" through different forms of book-marking. Naturally, such techniques do not provide for rich data querying, sharing, and management techniques in such environments.

It is well accepted that the creation, management, and utilization of different forms of metadata play a major role in realizing information system infrastructures that satisfy the above requirements in collaborative research environments. *Content-independent metadata* (such as data location, data format, authorship etc.) are typically handled through globally accessible data registries, for example, image registries. Sophisticated frameworks for the management of *content descriptive metadata* (recording the interpretation of data by researchers or programs), however, are still in their infancy. This is despite the fact that there have been major advancements in the context of the "Semantic Web", which aims at developing frameworks to associate semantic rich (ontology-like) structures with all forms of Web accessible data.

We claim that there is still a major gap between the creation of such semantic rich structures and the usage of these structures to actually enrich various forms of data. While most of the work has been and still is focusing on building ontologies for different domains, their usage in particular in collaborative research environments still is a major research issue. In a research project conducted at the University of California at Davis in the context of the Human Brain Project, we are currently developing components of an information system infrastructure that aims at (1) providing researchers the means to associate well-defined metadata (called *data annotations*) with heterogeneous, remote scientific data at different levels of granularity, and (2) utilizing data annotations to realize uniform and transparent data retrieval mechanisms that provide an integrated access to the distributed scientific (here Neuroscience) data. The following figure illustrates the conceptual components of this infrastructure. Scientific data residing at different sites in the collaborative research environment is assumed to be Web accessible. Concepts and relationships among concepts represent components of an ontology-like structure, ranging from simple standard vocabularies to complex, semantic rich domain specific ontologies. In our framework, concepts can be understood as agreed upon, well-defined templates for metadata that can be associated with data residing at research sites. A concept consists at least of a definition, a list of terms to refer to the concept, and a set of attributes used to describe instances of the concept. Data annotations are instances of concepts. They link scientific data (e.g., region of an image, fragment of a text document, or result of a query against a database) to a concept through Uniform Resource Identifiers (URIs). A data annotation furthermore contains instantiations of concept attributes.

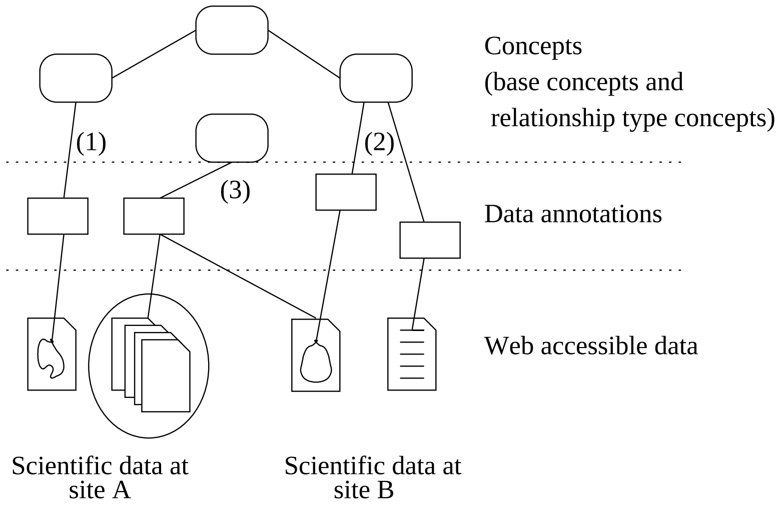
Concepts
(base concepts and
relationship type concepts)
(1)
(2)
(3)
Data annotations
Web accessible data
Scientific data at site A
Scientific data at site B

Figure 1: *Concept-Based Data Annotations: Concepts, relationships among concepts, data annotations, and data managed at different sites are represented in the annotation graph model, which allows for effective query formulation and processing through graph operations*

A concept can be as simple as the description of a class of objects of interest in the application domain ((1) and (2) in Figure 1) and as complex as a description of a process that is used to derive new data from existing data ((3) in Figure 1). Thus, assuming that processes and workflows can be described in form of concepts and concept relationships, data annotations can be used to specify the process through which some data has been obtained from existing data. Such types of data annotations can provide a means to record data provenance.

The functionality described above has been realized in an initial prototype that allows researchers to associate content descriptive metadata (instances of concepts) with distributed and heterogeneous forms of image and text data [1, 2, 3]. Concepts and annotations are currently managed by a centralized database system. Users can browse and query concepts, and are pointed to data that has been annotated using the selected concepts. Concepts thus can be understood as a database schema and data annotations represent linkage information to and between scientific data at different levels of granularity. Furthermore, annotations that have been made for selected data can be displayed, and through "traversal" in the annotation graph, similar data (annotated using the same concept(s)) can be retrieved.

Although this framework provides researchers with an effective means to manage and query distributed scientific data, there are still major challenges that need to be addressed. These challenges arise from several requirements. First, data annotations should exhibit some kind of compatibility and consistency (since several users can annotate the same data). Second, conceptual structures underlying data annotations evolve over time, that is, new types of objects of interest are discovered and new processes are employed to analyze and transform the data. Finally, the distribution aspect is not well addressed; the creation, management, and usage of concepts and data annotations currently occur in a centralized fashion. These aspects lead to the following general challenges in the context of managing semantic-rich metadata.

- Given that researchers can associate (controlled) metadata with remote data at different levels of granularity, what models and techniques are needed to ensure a certain level of consistency and compatibility of data annotations? To what extend does the type of user making an annotation play a role (i.e., experts versus naive users)? To what extend does the notion of data annotation interact with the notion of data quality? Is it reasonable and manageable to allow users to annotate data annotations, e.g., to make (well-defined) statements about the quality of an annotation? How many levels of annotations do make sense?

- Classes of objects of interest in an application domain typically evolve over time. That is, new classes of objects are discovered, existing ones are refined, revised or deleted, and new processes are employed to analyze and transform (collections of) object instances. Assuming that concepts that have been used in data annotations are revised or deleted, how does one deal with existing data annotations? What versioning models are appropriate in the context of evolving concepts, data annotations, and scientific data?

- How expressive should data annotations and underlying concepts be? For example, if one annotates two pieces of data at two sites with a concept describing a data transformation process, how are such concepts specified and properly used? Here we suggest to first study one particular type of process that is frequently used in the context of scientific data management and analysis scenarios, namely data visualization (see also [4]).

- Scientific application domains typically span a variety of specialized interests. Assuming that several research sites (managing data generated at their site) exist, what data distribution aspects become important? What distribution principles for metadata (concepts and data annotations) can and should be employed? How does one deal with the dynamics of research foci of interest in the context of re-distribution models for metadata?

In summary, we conjecture that there are several open research issues in the context of metadata creation and management in collaborative research environments where metadata cannot always be extracted automatically from data but requires the expertise and insight of researchers operating in the application domain. We also claim that despite the advancements in the "Semantic Web" arena, there have been no major useful models that provide for using ontologies to actually enrich heterogeneous and distributed data. In particular, we claim that by using ontology like structures to annotate data, the way in which data has been annotated (e.g., type of user, region of interest in the data etc.) can actually be used to determine consistency and quality aspects of such structures.

References

[1] M. Gertz, K. Sattler: Integrating Scientific Data through External, Concept-based Annotations. In *Second International Workshop Data Integration over the Web.* Toronto, Canada, pp. 87–102, 2002.

[2] M. Gertz, K. Sattler, F. Gorin, M. Hogarth, J. Stone: Annotating Scientific Images: A Concept-based Approach. In 14th Int. Conference on Statistical and Scientific Databases, IEEE Computer Society, 2002.

[3] M. Gertz, K. Sattler: A Model and Architecture for Conceptualized Data Annotations. Technical Report, Department of Computer Science, University of California, Davis, 2001.

[4] T.J. Jankun-Kelly, K.-L. Ma, K.-L., M. Gertz: A Model for the Visualization Exploration Process. To appear in *Proceedings of IEEE Visualization 2002*, IEEE Computer Science Press.