Context Interchange: Sharing the Meaning of Data

Michael Siegel
Sloan School of Management, E53-323
Massachusetts Institute of Technology
Cambridge, MA 02139
msiegel@sloan.mit.edu

Stuart E. Madnick
Sloan School of Management, E53-321
Massachusetts Institute of Technology
Cambridge, MA 02139
smadnick@eagle.mit.edu

1 Introduction

Previously, users and application developers, through manuals and experience, understood the context of the few systems that served their portion of the enterprise. But the number, types, and increased scope of data and systems that are now being integrated make it impossible to expect users to understand and remain current on the context or meaning of all information. Our research has the goal of representing, moving, and processing the context along with the information it describes. This requires both representations, models, manipulation languages and reconciliation algorithms for context knowledge.

In this position paper we present a number of significant research areas which we believe must be resolved so that context knowledge can be used to simplify the integration of multiple disparate database systems. As shown in Figure 1, the *export context* defines the meaning of the data provided by a data source while the *import context* defines the context requirements for the data receiver. Providing this context knowledge requires an understanding of issues in context representation, context models, common metadata vocabularies, comparisons of contexts including transformations, and system's architectures and operations.

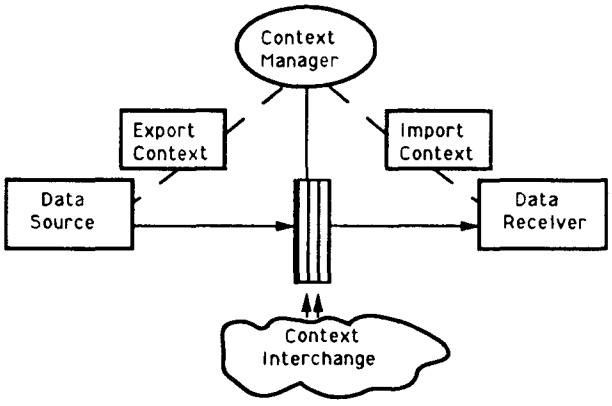

Figure 1: Architecture using Context Knowledge

2 Defining Context Using Metadata

Several researchers have provided different representations for data context through the use of metadata [CSR91,HBRY91,Law88,McC84,SM91, SG89]. However, context is not simply a schema-based concept. Rather a hierarchy of contexts may exist. There may be a context for an enterprise, a database, a relation, an attribute, a data value, or any other aggregation of data objects. For example, [WM90] defines the need for data value context to tag data with source information. Context may appear at different levels for different sources (e.g., the *currency* of a value may be an attribute in one relation but be a schema-based definition in another). Because of these complexities in context modeling there is a critical need for a well-defined metaschema model for context representation.

The most basic requirement for the use of context for multidatabase systems is the existence of a common metadata vocabulary.[1] In [SM91] we define a rule-based language for metadata representation that depends on a common vocabulary composed of a set of primitives whose meaning are identical in the context of all systems that share data. [CHS91] uses a global ontology to provide the common vocabulary and all component systems must provide semantic mappings to that global ontology.

In addition to a common vocabulary there needs to be a common understanding of the data topology [Tri90], that is the structure of the objects (e.g., different object structures for the same real world object) and their semantic components. In order to compare data semantics between systems, it may be necessary to place restrictions on the metadata defined for these systems.

Like standards, primitive concepts and topology may be agreed upon by committee, but in a multidatabase environment such standards will be very difficult to reach and maintain. However, standards as they apply to the common vocabulary are non-intrusive on the underlying systems because they require only agreement on primitive concepts used to describe the meaning of the data. Any system in the enterprise can use the common vocabulary to develop rules (i.e., context knowledge) describing data semantics. Terminology outside of this common language must translate to the common vocabulary otherwise comparison of data semantics will not be possible. An evolving approach to the development of a vocabulary might allow for negotiation among human experts [Tri90] leading to agreement on common terms

[1] Otherwise the need for a meta-metadata and so on.

and topology.

Research is needed into the understanding of context models that define the hierarchies of context knowledge. An understanding of common vocabulary requirements will be necessary for the development of acquisition methods and representations for context knowledge.

2.1 Comparing Context Among Systems

The semantic integration of multidatabase systems will depend on the development of algorithms for the identification and resolution of semantic conflicts (i.e., semantic reconciliation). These algorithms will use the context knowledge to mediate [Wie91] among systems to provide for meaningful data exchange.

In [SM91] we define algorithms for semantic reconciliation in a source-receiver environment. The architecture as shown in Figure 1 introduces the need for a *context manager* that mediates by comparing the data source's export context with the receiver's import context. The context manager must determine if the source can provide meaningful data to the receiver even as the meaning of the source data changes.

When data is being exchanged it must be determined when the semantics are found to be equivalent. Sometimes the semantic equivalence might be achieved through some *trivial mapping* (e.g., "yds" to "ft"). Mappings such as currency conversion are often *non-trivial* since they may require considerable additional context-dependent information such as time, place, quantity, and regulations. Finally, there are conflicts that are *unmappable*. For example, "average trade price" to "last trade price" or "age" to "birth date." The context manager must identify routines that can be used to transform the source semantics to those required by the application (e.g., currency conversion). Significant work is needed to understand the specification and evaluation of semantic equivalence between two context representations.

Finally, systems architectures and operational procedures need to be defined. For example, are the users of systems required to maintain their own import and export context knowledge or is it the responsibility of some specific group within the organization to maintain context and management algorithms for all systems?

2.2 Conclusion

We believe that the ability to represent and manipulate context will be an extremely important part of providing semantic integration in multidatabase systems. This capability will depend on the selection of an appropriate metadata representation, a means for establishing and maintaining a common vocabulary and algorithms for semantic reconciliation that include the use of semantic knowledge to resolve conflicts. Negotiation and coordination techniques and the development of standards will be important in the creation and evolution of a common vocabulary. Considerable research is needed to develop a better understanding of the types of metadata and the restrictions and operations that make metadata comparable among multiple disparate systems.

Acknowledgements: This work has been funded in part by the International Financial Research Services Center at MIT, National Science Foundation Grant #IRI902189, and Reuters.

References

[CHS91] C. Collet, M. Huhns, and W. Shen. *Resource Integration Using an Existing Large Knowledge Base.* Technical Report ACT-OODS-127-91, MCC, 1991.

[CSR91] S. Cammarata, D. Shane, and P. Ram. *IID: An Intelligent Information Dictionary for Managing Semantic Metadata.* Technical Report R-3856-DARPA, Defense Advanced Research Projects Agency, 1991.

[HBRY91] C. Hsu, M. Bouziane, L. Rattner, and L. Yee. Information resources management in heterogeneous, distributed environments. *IEEE Transactions on Software Engineering*, 17(6):604–625, June 1991.

[Law88] M. H. Law. *Guide to Information Resource Dictionary System Applications: General Concepts and Strategic Systems Planning. 500-152*, National Bureau of Standards, 1988.

[McC84] J. McCarthy. Scientific information = data + meta-data. In *Database Management: Proceedings of the Workshop November 1-2, U.S. Navy Postgraduate School, Monterey, California*, Department of Statistics Technical Report, Stanford University, 1984.

[SG89] A. Sheth and S. Gala. Attribute relationships: an impediment in automating schema integration. In *Position Papers: NSF Workshop on Heterogeneous Databases*, December 11-13, 1989.

[SM91] M. Siegel and S. Madnick. A metadata approach to resolving semantic conflicts. In *Proceeding of the 17th International Conference on Very Large Data Bases*, September 1991.

[Tri90] A. Trice. *Facilitating Consensus Knowledge Aquisition.* PhD thesis, Massachusetts Institute of Technology, 1990.

[Wie91] G. Wiederhold. Mediators in the architecture of future information systems. *Submitted to IEEE Computer*, 1991.

[WM90] R. Wang and S. Madnick. Data-source tagging. In *Proceeding from the Very Large Database Conference*, 1990.