LORE: A Lightweight Object REpository for Semistructured Data*

Dallan Quass, Jennifer Widom, Roy Goldman, Kevin Haas, Qingshan Luo, Jason McHugh, Svetlozar Nestorov, Anand Rajaraman, Hugo Rivero, Serge Abiteboul, Jeff Ullman, Janet Wiener

Stanford University Database Group, http://db.stanford.edu

The number of information sources accessible electronically is growing rapidly. Many of these sources store and export unstructured data in addition to or instead of structured data. In most cases, however, the unstructured data is not entirely devoid of structure, i.e., the data is *semistructured*. We consider data to be semistructured when there is no schema fixed or known in advance and when the data may be incomplete or irregular. For example, HTML files on the World-Wide Web usually contain some structure, but often the data is irregular or incomplete. In addition, data integrated from multiple, heterogeneous information sources often is semistructured.

Storing and querying semistructured data poses considerably different problems and requirements than those for traditional databases, where data storage and query processing are dependent upon structured data. Relational, nested-relational, and object-oriented database systems, for example, all depend upon the data having a known and regular schema. We have developed a system called *LORE* (for *Lightweight Object REpository*), and a query language called *LOREL* (for LORE Language), aimed specifically at handling semistructured data.

The data model used in Lore is a "lightweight" object model called *OEM* (for *Object Exchange Model*) [3]. OEM is a simple, self-describing model with object nesting and identity. Because we use Lore primarily for storing and querying data obtained from other information sources, Lore itself also is "lightweight," in the sense that it is a repository and a query engine but not a full-feature database management system. Currently, Lore does not provide transaction management, concurrency control, or recovery. Lorel, the query language supported by Lore, is a compatible extension to the OQL object-oriented query language [1], with new features designed specifically for querying semistructured data: partially specified path expressions, wildcards, automatic type coercion in comparisons, and a special semantics for disjunction. Unlike OQL, Lorel does not enforce strong typing, thus allowing similar objects to be compared and retrieved despite minor differences in their structure. Finally, Lorel allows querying and schema browsing when the object structure is unknown or only partially known. Details on an earlier version of Lore and Lorel can be found in [4].

In addition to its special features for querying over semistructured data, Lore includes enhancements for managing data in a heterogeneous environment. As well as the usual base types (integer, string, etc.), Lore supports "multimedia" types such as GIF images, URLs, Java applets, audio, and text; additional base types can be incorporated easily. Lore supports seamless access to "external objects"—objects fetched on demand from arbitrary information sources during query execution and cached for later use. Any object in Lore may be a placeholder for an external object, allowing Lore to serve both as a storage repository for semistructured data and as a query-driven integration engine.

Our demonstration shows how Lore can be used for warehousing, querying, and accessing information stored as HTML pages on the World-Wide Web. Some data is stored in Lore, while other data is fetched in response to user queries. Lore is accessed through an HTML interface, using the *MOBIE* system from the related *TSIMMIS* project for displaying and browsing query results [2].

*Research supported by the Air Force Wright Laboratory Aeronautical Systems Center under ARPA Contract F33615-93-1-1339, by the Air Force Rome Laboratories under ARPA Contract F30602-95-C-0119, and by equipment grants from Digital and IBM Corporations.

References

[1] R.G.G. Cattell, editor. *The Object Database Standard*. Morgan Kaufmann, San Francisco, CA, 1994.

[2] J. Hammer et al. Information translation, mediation, and Mosaic-based browsing in the TSIMMIS system. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, page 483, San Jose, California, May 1995. Demonstration.

[3] Y. Papakonstantinou, H. Garcia-Molina, and J. Widom. Object exchange across heterogeneous information sources. In *Proceedings of the Eleventh International Conference on Data Engineering*, pages 251–260, Taipei, Taiwan, March 1995.

[4] D. Quass, A. Rajaraman, J. Ullman, and J. Widom. Querying semistructured heterogeneous information. In *Proceedings of the Fourth International Conference on Deductive and Object-Oriented Databases*, pages 319–344, Singapore, December 1995.