Management of Semistructured Data

Dan Suciu
AT&T Labs — Research
suciu@research.att.com

A huge amount of data is available today on the Internet, or on the private Intranets of many companies. This data is structured in a multitude of ways. At an extreme we find data coming from traditional relational or object-oriented databases, with a completely known structure. At another extreme we have data which is fully unstructured, such as images, sounds, and raw text. But most of the data falls somewhere in between these two extremes, for a variety of reasons: the data may be structured, but the structure is not know to the user; the user may know the structure, but chooses to ignore it, for browsing purposes; the structure may be implicit, such as in formatted text, and is not as rigid and regular as in traditional databases; the data may be in non-traditional formats, such as the ASN.1 exchange format; the schema of the data is huge and changes often, so that we may prefer to ignore it. Several researchers have worked recently on problems related to data fitting this description, and have coined the term *semistructured data* for it. Two recent tutorials [Abi97, Bun97] contain an excellent introduction to semistructured data and a comprehensive bibliography on this new research topic.

Semistructured data is most naturally modeled as a collection of objects; each object may have any number of (possible repeating) attributes, whose values are other objects, or atomic data. All models proposed for semistructured data consists of some kind of labeled graph, in which nodes correspond to objects or values, and the edges correspond to attributes [PGMW95, QRS+95, BDHS96]. For example a record with fields `A, B, C` is represented in this model as a node with three outgoing edges labeled `A, B, C`; a set with n elements is represented as a node with n outgoing edges, all labeled `element`, each pointing to some element in the set. It is equally easy to represent data whose structure is less rigid than that of traditional databases. Data in the labeled graph model is self-describing and has no separate schema.

Techniques for processing traditional databases are often not well suited for this new data model. Semistructured data requires specific query languages, update methods, and techniques for query evaluation, object integration, structure discovery, query decomposition, etc. Research on semistructured data is relatively recent, and only a few of the topics have been addressed.

In May, 1997, some researchers in the area organized a *Workshop on Management of Semistructured Data*. The program committee received an unexpected high number of submissions (31 papers), which was perceived as an indication of the interest in the area. Thirteen papers were selected for presentation at the workshop, and are available from `http://www.research.att.com/~suciu/workshop-papers.html`. The workshop also included two panels: *How much structure is semi-structure ? Defining the topic*, moderated by Serge Abiteboul, and *Semi-structured data: useful or harmful ?*, moderated by Peter Buneman. The panelists and members of the audience discussed several examples of instances of semistructured data, and various approaches to defining this topic; some informal notes are available from `http://www.research.att.com/~suciu/workshop-panels.html`.

Five of the papers presented at the workshop were selected by the program committee to be included in this special section of the Sigmod Record. The papers offer the reader a sample of current research in semistructured data, and contain a number of additional references. We describe briefly the papers, then include a very short overview of previous work in semistructured data with references for further reading.

- *Wrapper Generation for Semi-structured Internet Sources*, by Naveen Ashish and Craig Knoblock. The paper describes how wrappers for structured text documents can be generated semi-automatically, with minimal user intervention. The system exploits the formatting information in the source, and generates an YACC specification for a parser for that source.

- *Semistructured and Structured Data in the Web: Going Back and Forth*, by Paolo Atzeni, Giansalvatore Mecca, and Paolo Merialdo. The paper describes a system, `ARANEUS`, which can apply the full power of a traditional relational databases system to web site management. The system includes tools for migrating a web site to a database, and for generating a web site from a database.

- *Extracting Entity Profiles from Semistructured Information Spaces*, by Robert Nado and Scott Huffman. The paper describes a system, NX, which integrates a large number of semistructured documents (over 300,000).

- *Inferring Structure in Semistructured Data*, by Svetlozar Nestorov, Serge Abiteboul, and Rajeev Motwani. This paper describes a novel approach for inferring structure in web sites. The authors propose an algorithm for classifying objects into classes, according to their set of attributes, and show how to identify relevant classes even when they are sparsely populated.

- *Integrating Dynamically-Fetched External Information into a DBMS for Semistructured Data*, by Jason McHugh and Jennifer Widom. This paper refers to LORE [QRS+95, AQM+96, AQM+97, MAG+97], a database management system specifically designed for semistructured data. Here, the authors describe the external data manager, whose purpose is to fetch and cache external data on a by-need bases.

A few areas in semistructured data have been explored previously. We list here a few of them, together with references: for a complete list of references we refer the reader to the tutorials [Abi97, Bun97].

Query Language Design Several new query languages for semistructured data have been proposed: LOREL [QRS+95, AQM+96, AQM+97, MAG+97], UnQL [BDS95, BDHS96], WebSQL [MMM96], WebOQL [AM98], StruQL [FFLS97]. Their common feature is the ability to traverse arbitrary long paths in the data, usually specified in the form of a regular path expression: thus these query languages are *recursive*. The languages proposed differ with respect to their target application, expressive power, and restructuring capabilities.

Semistructured Database Systems LORE [QRS+95, AQM+96, AQM+97, MAG+97] is the only general-purpose database management system specifically designed for semi-structured data. A number of systems have been designed for querying the WWW or managing web sites [KS95, MMM96, AM98, FFLS97].

Data Integration Although an old research area, data integration has received new attention in the context of semistructured data, due to the need to integrate data in a variety of formats. Tsimmis [PGMW95, PAGM96] is a data integration system for semistructured data.

Optimizations Work on query optimiztion for semistructured data has barely started. Some results are presented in [BDHS96, AV97, FPS97, GW97, FS98].

Describing Structure Two proposals exists: graph schemas [BDFS97] and data guides [NUWC97, GW97].

We wish to thank the authors for their prompt submissions and cooperation in overcoming several problems which occurred in connection with electronic submissions, and the National Science Foundation, who sponsored the Workshop on Management of Semistructured Data through NSF grant number IRI93-18791.

References

[Abi97] Serge Abiteboul. Querying semi-structured data. In *ICDT*, 1997.

[AM98] G. Arocena and A. Mendelzon. WebOQL: Restructuring documents, databases, and webs. In *International Conference on Data Engineering*, 1998. To appear.

[AQM$^+$96] Serge Abiteboul, Dallan Quass, Jason McHugh, Jennifer Widom, and Janet Wiener. The Lorel query language for semistructured data, 1996. Manuscript available from `http://www-db.stanford.edu/lore/`.

[AQM$^+$97] S. Abiteboul, D. Quass, J. McHugh, J. Widom, and J. Wiener. The Lorel query language for semistructured data. *International Journal on Digital Libraries*, 1(1):68–88, April 1997.

[AV97] Serge Abiteboul and Victor Vianu. Regular path queries with constraints. In *Proceedings of ACM Symposium on Principles of Database Systems*, 1997.

[BDFS97] Peter Buneman, Susan Davidson, Mary Fernandez, and Dan Suciu. Adding structure to unstructured data. In *ICDT*, pages 336–350, Deplhi, Greece, 1997. Springer Verlag.

[BDHS96] Peter Buneman, Susan Davidson, Gerd Hillebrand, and Dan Suciu. A query language and optimization techniques for unstructured data. In *SIGMOD*, 1996.

[BDS95] Peter Buneman, Susan Davidson, and Dan Suciu. Programming constructs for unstructured data. In *Proceedings of DBPL'95*, Gubbio, Italy, September 1995.

[Bun97] Peter Buneman. Tutorial: Semistructured data. In *PODS*, 1997.

[FFLS97] Mary Fernandez, Daniela Florescu, Alon Levy, and Dan Suciu. A query language for a web-site management system. *SIGMOD Record*, 26(3):4–11, September 1997.

[FPS97] Mary Fernandez, Lucian Popa, and Dan Suciu. A structure based approach to querying semistructured data. In *Proceedings of the Workshop on Database Programming Languages*, 1997.

[FS98] Mary Fernandez and Dan Suciu. Optimizing regular path expressions using graph schemas. In *Proceedings of the International Conference on Data Engineering*, 1998. To appear.

[GW97] Roy Goldman and Jennifer Widom. DataGuides: enabling query formulation and optimization in semistructured databases. In *VLDB*, September 1997.

[KS95] David Konopnicki and Oded Shmueli. Draft of W3QS: a query system for the World-Wide Web. In *Proc. of VLDB*, 1995.

[MAG+97] J. McHugh, S. Abiteboul, R. Goldman, D. Quass, and J. Widom. Lore: A database management system for semistructured data. *SIGMOD Record*, 26(3):54–66, September 1997.

[MMM96] A. Mendelzon, G. Mihaila, and T. Milo. Querying the world wide web. In *Proceedings of the Fourth Conference on Parallel and Distributed Information Systems*, Miami, Florida, December 1996.

[NUWC97] S. Nestorov, J. Ullman, J. Wiener, and S. Chawathe. Representative objects: concise representation of semistructured, hierarchical data. In *ICDE*, 1997.

[PAGM96] Y. Papakonstantinou, S. Abiteboul, and H. Garcia-Molina. Object fusion in mediator systems. In *Proceedings of VLDB*, September 1996.

[PGMW95] Y. Papakonstantinou, H. Garcia-Molina, and J. Widom. Object exchange across heterogeneous information sources. In *IEEE International Conference on Data Engineering*, March 1995.

[QRS+95] D. Quass, A. Rajaraman, Y. Sagiv, J. Ullman, and J. Widom. Querying semistructure heterogeneous information. In *International Conference on Deductive and Object Oriented Databases*, 1995.