Report on the Eighth International Workshop on Knowledge Representation Meets Databases (KRDB), September 15, 2001

Maurizio Lenzerini
Daniele Nardi
Università di Roma
"La Sapienza", Roma, Italy
lenzerini@dis.uniroma1.it
nardi@dis.uniroma1.it

Werner Nutt
Heriot-Watt University
Edinburgh, UK
nutt@cee.hw.ac.uk

Dan Suciu
University of Washington
Seattle, WA, USA
suciu@cs.washington.edu

The Eighth International Workshop on Knowledge Representation Meets Databases (KRDB) was held at the Pontificia Università Urbaniana, in Rome, right after VLDB 2001. KRDB was initiated in 1994 to provide an opportunity for researchers and practitioners from the two areas to exchange ideas and results. This year's focus was on Modeling, Querying and Managing Semistructured Data. The one day program included ten research papers, one invited talk, and a panel. Eight of the accepted papers addressed various topics related to representation of information and reasoning in XML, one was on data integration and one on transaction processing. The invited talk, by Georg Gottlob, was on wrapper generation, also adopting XML to express the result of the wrapping process.

Georg Gottlob opened the program with the invited talk *Declarative Information Extraction, Web Crawling and Recursive Wrapping with Lixto*. Lixto is a system and method for the visual and interactive generation of wrappers for Web pages under the supervision of a human developer, for automatically extracting information from Web pages using such wrappers, and for translating the extracted content into XML. The talk addressed some advanced features of Lixto, such as disjunctive pattern definitions, specialization rules, and Lixto's capability of collecting and aggregating information from several linked Web pages. These features have been exploited in a demo developed in the commercial domain of a bookstore.

The four papers in the first group applied concepts and techniques from logic to XML. In *Containment and integrity constraints for XPath fragments*, A. Deutsch and V. Tannen consider a fragment of XPath expressions, with variable bindings, and study the expression containment problem. In addition, they further restrict these expressions by imposing constraints on the input XML document, such as key constraints, or referential integrity constraints. Depending on the XPath fragment (for expressions and constraints) the authors show the complexity of the containment problem to range from NP, to Π_2^p, to EXPTIME, to undecidable. A related problem is that of path constraints, studied in *Path Constraints from a Modal Logic Point of View*, by N. Alechina, S. Demri, and M. de Rijke. Here the authors consider full regular path expressions, and study the satisfiability and implication problems, showing them to range from PSPACE to EXPTIME. In the third paper, *Approximate reasoning in semistructured data*, G. Grahne and A. Thomo address a problem that has received surprisingly little attention in the past: how to retrieve results that "approximatively" match a regular path expression. Drawing upon the limitedness problem studied by Hashiguchi, the authors develop a framework and show that the approximate answering problem is decidable. Finally, in *Schema extraction from XML: A grammatical inference approach*, the author B. Chidlovskii establishes an interesting connection between two previously independent problems: schema extraction from XML data, and grammar inference.

The second group of regular papers was devoted to more practical aspects of XML. In *Semantic lossy compression of XML data*, by M. Cannataro, G. Carelli, A. Pugliese, and D. Sacca, the authors propose a scheme for lossy XML compression. This is especially useful in applications where data, usually accessed over a regular network, needs to be occasionally accessed over slow channels, like on a WAP phone. The proposed technique allows the user to negotiate the amount of compression, trading off speed for precision. The XML storage problem is addressed in *From XML to relational databases*, by M. Yan and A.W. Fu. They propose a four step method in which the DTD is first simplified, from which a schema prototype tree is derived and then a relational schema is generated. In the last two steps functional dependencies and candidate keys are discovered, then the relational schema is normalized. In *Capturing Data using XML Paragraph-centric Document*, the authors Y. Badr, M. Sayah, M. Laforest, and A. Flory address the problem of extracting data from textual information. The proposal is developed within a system working in a medical domain. The overall architecture is based on the representation of data in XML and a method for building an XML description of the data extracted from a medical prescription. In *A Framework for Generic Integration of XML Sources*, W. May introduces a Datalog-style extension of the XML query language XPath and illustrates how it can be used to define an integrated view on a collection of XML databases and to query this view.

The last group of papers included two contributions from data integration and transaction processing, respectively. In *Towards a comprehensive methodological framework for integration*, D. Calvanese, S. Castano, F. Guerra, D. Lembo, M. Melchiori, G. Terracina, D. Ursino, and M. Vincini survey three approaches to information integration, developed independently by the partners of an Italian national project, and present an architecture to put them into a coherent whole. In *Towards a general theory of advanced transaction models in the situation calculus*, I. Kiringa uses the framework of the situation calculus to specify transactions that only satisfy relaxed ACID properties and shows how to implement such specifications in the programming language GOLOG.

The workshop ended with a panel entitled *How much reasoning will be needed in applications of semistructured data?*, organized by Werner Nutt, with P. Atzeni, S. Ceri, and S. Chawathe as participants. Atzeni saw the main applications of semistructured data in tasks that involve the transformation of data into documents and, vice versa, of documents into data. He argued that, while techniques for the first direction seem to be straightforward, the converse direction raises serious difficulties. These, however, are more likely to be solvable by heuristic techniques rather than by reasoning. Ceri described XML as a data format whose primary application will be to support interoperability and advocated reactive rules as the appropriate computational mechanism. In the visions of the two panelists, a need for reasoning about queries arises because they see applications of semistructured data being based on declarative languages, which makes optimization mandatory. Finally, Chawathe described his view of semistructured data as "data that are not important enough to be structured thoroughly." Accordingly, he rephrased the question as *How much reasoning is good for XML?*, and answered that useful reasoning techniques are those that do not presuppose additional human effort.

The workshop generated considerable interest, in part because it traditionally attracts researchers from both databases and knowledge representation, and in part because of this year's focus on XML and semistructured data. There were 24 submissions (with 10 accepted papers) and there were 38 participants at the workshop. At the same time the workshop retained its traditional spirit of establishing strong bridges between logic and knowledge representation on the one hand, and databases and applications on the other hand. The proceedings are published electronically in the CEUR series and are available at `www.ceur-ws.org`. The next KRDB workshop will take place as an affiliate event to the Conference on Principles of Knowledge Representation and Reasoning (KR2002) in Toulouse in April 2002.