Advanced XML Processing:

Guest Editor's Introduction

Karl Aberer

Distributed Information Systems Laboratory
EPFL-DSC, CH-1015 Lausanne, Switzerland
email: karl.aberer@epfl.ch

XML has established itself over the recent years as THE standard for representing data in scientific and business applications. Starting out as a standard data exchange format for the Web, it has become instrumental in the development of electronic commerce applications and online information services, and draws in its tailwind a multitude of standardization efforts for all kinds of applications. Documents are not only used for representing multimedia information content but also for many other purposes, like the representation of meta-information and the specification of component interfaces, protocols, and processes. As a consequence, the amount of XML data being stored and processed is large and will be increasing at an astonishing rate. This has caused XML data management to become a focus of research efforts in the database community. This was the motivation for putting the RIDE-DM'2001 [1], the eleventh workshop in a series of annual workshops on Research Issues in Data Engineering, held April 1-2, 2001, in Heidelberg, Germany, under the theme of "Document Management for Data Intensive Business and Scientific Applications". Not surprisingly, at this workshop a number of papers on XML data processing have been presented, which nicely reflect recent developments in the area of XML data processing. This was an excellent opportunity to set up this SIGMOD RECORD special issue to give a snapshot of the current state of the art in this area and indicate potential future directions of research.

When we look back to a former, related issue of this journal from 1997 on "Semi-structured data management", edited by Dan Suciu [2], we can observe a natural development. At that time the following research problems had been identified: query language design, semi-structured database systems, data integration, query optimisation, and schema languages. One group of papers in this issue demonstrates that substantial progress in these areas has been achieved in the interim, adopting XML as the canonical representation for semi-structured data. The systems and the standards for semi-structured data management based on XML have gained some maturity.

- *Query languages are now being standardized.* The database community has taken over a great responsibility in this process as illustrated by the fact that many members of it were instrumental in the W3C XQuery standardization effort. In Peter Fankhauser's article "XQuery Formal Semantics: State and Challenges" we obtain insight into the query algebra and its semantics that has been developed by this Working Group and that will underlie as a formal framework the future XML query processing systems. It is also interesting to see that ideas developed much earlier in data management research found again their place in today's setting.

- *Database systems for XML are becoming widely available.* Both relational and native storage systems have been developed. Relational database vendors, like IBM and Oracle, offer XML extensions for their relational database management systems. Researchers have been and are playing an instrumental role in investigating the inherent problems of mapping between the hierarchical data model of XML and the flat relational data model, which has incited numerous publications. One particularly nice result on this topic is shown in the article "A General Technique for Querying XML Documents Using a Relational Database System" by Jayavel Shanmugasundaram. Other vendors, like Software AG with Tamino and ObjectStore with Excelon provide native solutions for XML storage. It is of course of greatest interest to compare the specific characteristics of various XML storage systems and the approaches they take, in order to make informed decisions when selecting a XML storage system for a specific application setting. Therefore, the question of evaluating their performance and comparing them is becoming very critical and leads to research on developing comparative benchmarks for XML database systems. Despite the fact that this is an inherently difficult task and may lead to sensitivity on the side of the vendors, benchmarks will be crucial to set up measurable objectives and thus incite future performance improvements. In the article "Why And How To Benchmark XML Databases" by Albrecht Schmidt et al. we find an overview on an ongoing effort in that direction.

- *XML is playing a fundamental role in the integration of heterogeneous data*. Tools for *wrapping* and integrating heterogeneous data sources, in particular on the Web, using XML as the data model at the integration layer are developed and are becoming popular. The article "Wrapping Data into XML" by Wei Han et al. presents one interesting example of an XML-extended wrapper code generation system, XWRAP-ELITE. It wraps arbitrary dynamic HTML pages and turns them into XML documents for further processing.

Another group of papers in this issue shows that new advanced questions of data processing, resulting from requirements of new XML applications have emerged, and led to new research directions. In his article "On Database Theory and XML", based on the keynote speech given at RIDE-DM'2001, Dan Suciu discusses which new theoretical questions arise from advanced applications of XML. He looks at three particular problems: XML publishing, the problem of transforming relational data into XML, XML type-checking, the problem of checking whether the output of an XML generating program conforms to a given DTD, and XML storage.

An interesting and not necessarily expected result following the presentations at RIDE-DM'2001 was, that a new aspect of XML data processing is moving into the focus of research interest: the question of *changes* in XML databases. In fact, changes are occurring at various granularities of the time scale throughout the life cycle of an XML application.

- *Short-term at data level*: the contents of XML stores are not static but evolve over time. Since XML data processing has its roots in *document* management, in many cases one wants to maintain the history of the changes applied to data. This leads to the need for versioning mechanisms. Shu-Yao Chien et al. present in their article "XML Document Versioning" solutions that can be provided for that.

- *Medium-term at schema level*: Data schemas in XML applications - either DTDs, or more recently XML Schemas - are not static either; rather they evolve due to changing application requirements. This leads both to technical requirements for supporting schema changes [4] and to software engineering related questions of maintaining consistency and allowing schema adaptations [5].

- *Long-term archiving*: As more and more data will be stored in XML format, XML stores will become the final home of huge amounts of relevant data. Among the many problems in long-term archiving, such as durability of storage media or compatibility with future computing platforms, the problem of maintaining the necessary meta-information for enabling the access and proper interpretation of XML data by future generations of users also needs to be addressed. This problem is taken up in the article "Preservation of Digital Data with Self-Validating, Self-Instantiating Knowledge-Based Archives" by Bertram Ludäscher et al..

We have covered some of the important issues in XML data processing, such as XML databases and querying, XML transformations and the management of change in XML databases. There are more questions on XML processing, which are not discussed in this issue. Some of the big themes we can expect to generate research over the coming years are

- *Semantics*: The ongoing discussion on the Semantic Web shows that there is strong desire to better support information search and integration by making semantic information on XML data more explicitly available, e.g. as attempted in the RDF standard. This affects both the representation as well as the processing of semantic information and will require a fruitful collaboration among the fields of data management and knowledge representation and processing.

- *Multimedia*: XML data processing as of today is still mostly taking a very data and text-centred viewpoint. The role of XML in multimedia data processing is currently the representation of structural and semantic information on multimedia data, as with SMIL and MPEG7. The integration of data management systems for streaming media and static data has been a research challenge over many years. Taking up the message-oriented paradigm in XML processing could allow obtaining a more uniform view of processing of multimedia and XML data streams. Attempts in that direction are under way [6].

- *Mobility*: Mobile data management is an established research area [3]. The asynchronous, message-oriented communication paradigm that is implied by the use of XML, might turn out to be particularly suitable for those applications. Also, as resources in mobile computing will remain inherently scarce, personalization and localization of information, requiring adequate management of the corresponding metadata, will play a major role.

- *Security and Trust*: Electronic commerce is a driving application for using XML. This was at RIDE-DM'2001 the topic of both a keynote speech by Tamer Ozsu on "Document Management Issues in E-Commerce" and a panel entitled "What Do You Need to Set Up an E-Commerce Platform?". Consequently, problems of security and trust play an important role in the management of (sensitive) XML data. This opens a wide range of interdisciplinary questions ranging from the integration of cryptographic methods with XML data management infrastructures, the mining of XML databases for supporting business decisions, to the management of customer and business partner interaction histories for obtaining data for trust assessment.

Probably (and hopefully) we will see in a few years a new special issue of SIGMOD RECORD focussing on these particular questions. In the meantime we hope the reader can take from this issue some interesting information on the current state-of-the art in the field and some inspiration for his or her own, future work.

I would particularly thank the Editor of SIGMOD RECORD Ling Liu for opening the opportunity for setting up this special issue and to the authors for their cooperation in providing their articles on relatively short notice.

References

[1] D. Suciu (Ed.): SIGMOD RECORD, Special Issue on Management of Semi-structured data, Vol 26, No 4, December 1997.

[2] M.H. Dunham, A. Helal (Eds.): SIGMOD RECORD, Special Issue on Data Management Issues in Mobile Computing, Vol 24, No 4, December 1995.

[3] Karl Aberer, Ling Liu (Eds.): Eleventh International Workshop on Research Issues in Data Engineering: Document Management for Data Intensive Business and Scientific Applications, Heidelberg, Germany, 1-2 April 2001, IEEE Computer Society, 2001.

[4] Hong Su, Diane Kramer, Li Chen, Kajal T. Claypool, Elke A. Rundensteiner: XEM: Managing the Evolution of XML Documents. RIDE-DM 2001: 103-110, 2001.

[5] Lex Wedemeijer: Engineering for Conceptual Schema Flexibility. RIDE-DM 2001: 85-92, 2001.

[6] Calton Pu, Karsten Schwan, Jonathan Walpole: Infosphere Project: System Support for Information Flow Applications. SIGMOD Record 30(1): 25-34, 2001.