Data on the Web:
From Relational to Semistructured Data and XML

by Serge Abiteboul, Peter Buneman, and Dan Suciu

Morgan Kaufmann, 1999
258 Pages, ISBN 155860622X

Review by:
Fernando Berzal and Nicolás Marín
Dept. Computer Science and Artificial Intelligence, University of Granada, Spain
nicm@decsai.ugr.es

Since Ancient Days, human beings have exchanged money for goods, or even some goods for other kind of goods. The goods that were exchanged characterized the different ages through History. Today, we are in the so-called Information Age and, therefore, a core part of current interactions take place around data exchanges. Serge Abiteboul et al.'s "Data on the Web" aims at laying the foundations for data exchange in those applications which deal with semistructured data.

A few years ago, software developers used to tame complexity in software systems through the design and implementation of structured data models. They built their applications over well-structured data. The advent of web and object technologies has shown some of the limitations of structured models. The identification of such limitations has led to the development of new approaches to deal with not-so-well structured data.

The exponential growth of web sites has resulted in a huge amount of data which is at our fingertips but lacks the structure that would allow us to use traditional database techniques. During the last decade, researchers and standards organizations have developed new proposals to deal with data management, data fusion, and data interchange in semistructured contexts. In these situations, standards such as the World Wide Web Consortium's eXtensible Markup Language (XML) have proved their suitability.

As stated in the book introduction, XML provides "a simple syntax for data that is both human- and machine-readable". As a logical representation mechanism and not just a data format, XML bridges different views of the data web, namely, that which treats data as collections of documents (that is, sequences of terms with associated tags) and that of database-oriented people who always try to obtain a structured logical model for their data.

The authors present in their book a semistructured data model and relate it with previous data models. Given its current importance, XML and XML-related standards are discussed as particular incarnations of the semistructured data model the authors propose. Semistructured data can be considered as "schemaless or self-describing", that is data which include part of their associated metadata, in contrast to the usual datasets that need external information to be interpreted. Tuples and their relations, as well as objects, are viewed as particular examples of semistructured data in order to show the power of the semistructured data model, which can deal with classical structured data. Before treating XML in depth, the authors comment on some existing data formats which have been used to

represent semistructured data (e.g. OEM and ACeDB).

Apart from the standard XML syntax, the book also includes examples using alternative notations, which are equivalent to each other. LISP-like lists and their corresponding syntax trees are employed to illustrate the nature of semistructured data. The book also includes an introduction to the graph terminology that is used to present alternative technologies in a common framework. For instance, the authors show how a given query can be expressed as a Datalog program, thus establishing connections between semistructured data and logic (in the same sense that SQL can be viewed as a relational-calculus-based query language from a purely logical point of view).

The reader can find a brief description of the XML syntax which, although is relatively comprehensive, is interestingly related to semistructured data representation models. Outdated DTDs (Document Type Definitions) are examined as a means to represent the context-free grammar that determines valid XML documents. More recent standards to describe XML document content (such as XML schema) are only mentioned.

The second part of the book is devoted to query languages for semistructured data. Several languages are explained, from the authors' own Lorel language to Strudel's StruQL, without forgetting specific query languages for XML. XML-QL is presented as a relationally complete language and a novel presentation of XSL (XML Stylesheet Language) as a query language is also included. This part of the book might be too hard for those lacking a solid academic background, since it discusses theoretical aspects and properties of different graph-based and logic-based approaches to semistructured data, something which is really worthwhile for researchers but probably useless for practitioners.

The challenging problem of schema extraction, which consists of typing semistructured data, rounds off the study of the semistructured data model. This third part of the book overviews some schema formalisms and highlights aspects to be considered when trying to extract schemas from data (or even from queries), which is extremely useful for query optimization and efficient data storage.

The final chapters discuss implementation issues focusing on Stanford's Lore System and AT&T's Strudel, two sample systems for semistructured data management the authors have been involved with. Although pointers to other systems are given, a more thorough review of existing systems would be valuable.

As indicated on the back cover, Data on the Web might be considered as a comprehensive and relatively up-to-date examination of effective techniques for retrieving and processing semistructured data. The authors' rigorous style and bibliographic remarks at the end of each chapter make this book outstand in its category, although practitioners might be misled by its title, since they will not find anything about versions, change control, and information retrieval in content management systems. Anyway, database-related academicians should reserve some space on their shelves for this book.