Chapter IX

Integrity Issues in the Web: Beyond Distributed Databases

José F. Aldana Montes, Mariemma I. Yagüe del Valle and Antonio C. Gómez Lora
Universidad de Málaga, Spain

INTRODUCTION

Issues related to integrity in databases and distributed databases have been introduced in previous chapters. Therefore, the integrity problem in databases and how it can be managed in several data models (relational, active, temporal, geographical, and object-relational databases) are well known to the reader. The focus of this chapter is on introducing a new paradigm: The Web as the database, and its implications regarding integrity, i.e., the progressive adaptation of database techniques to Web usage. We consider that this will be done in a quite similar way to the evolution from integrated file management systems to database management systems.

In any case, this will be a much more difficult goal and quite a lot of work is still to be done. The special features of the Web make things which are necessary on a database system just optional in this environment. On the other hand, some other things which are usually considered as essential parts of any database, are now disassembled into its building blocks and used as needed (Silberschatz & Zdonik, 1996; Bernstein et al., 1998).

At first glance, the Web is a huge repository of information without any structure whatsoever. Nowadays, this is changing quickly. The consolidation of the Extensible Markup Language (XML, 1998) as a new standard adopted...
by the World Wide Web Consortium (W3C) has made the publication of electronic data easier. With a simple syntax for data, XML is, at the same time, human and machine understandable. XML has important advantages over HTML (HyperText Markup Language). While HTML is a data visualization language for the Web (this was not its initial intended purpose), with XML, data structure and rendering are orthogonal. We can represent meta-information about data through user-defined tags. No rendering information is included in an XML document.

It could be considered that the main feature of XML is that of being a data exchange format, but we will show that it is much more than this in this chapter.

Thinking about the Web as a huge, highly distributed database, we may consider different dimensions to conceptually describe it. Özsu and Valduriez (1999) defines a classification of database systems with respect to: 1) their distribution; 2) the autonomy of local systems; and 3) the heterogeneity of database systems. The autonomy concept is considered as the distribution of control, not of data. This indicates the degree to which individual DBMSs can operate independently. Whereas autonomy refers to the distribution of control, the distribution dimension deals with the physical distribution of data over multiple sites. With respect to heterogeneity, this can range from hardware heterogeneity, differences in networking protocols, variations in DBMSs, etc., to the data model or the policy for managing integrity on the database.

Obviously, the Web is on the distribution plane, and, as shown in figure 1, we think that “it falls out” of the cube because it presents the highest degree of distribution, heterogeneity, and autonomy, and therefore, traditional distributed database techniques must be further extended to deal with this new environment. It is within this context that we are going to study the different issues related to integrity and its maintenance on the Web. We are also going to introduce the reader to other related and open issues, such as the query problem and query optimisation on the Web, since the special features of the Web environment make techniques for querying or maintaining the Web, different to those of traditional databases.

SEMISTRUCTURED DATA AND XML: AN OVERVIEW

With respect to the information available on the Web, we can distinguish between data which is completely unstructured, such as images, sounds, and
Related Content

A Database Project in a Small Company (or How the Real World Doesn't Always Follow the Book)
Efrem Mallach (2009). Selected Readings on Database Technologies and Applications (pp. 134-147).
www.igi-global.com/chapter/database-project-small-company-real/28550?camid=4v1a

Toward a Formal Semantics for Control-Flow Process Models
www.igi-global.com/article/toward-formal-semantics-control-flow/65542?camid=4v1a

A Framework for Analyzing Mobile Transaction Models
www.igi-global.com/chapter/framework-analyzing-mobile-transaction-models/4349?camid=4v1a