Chapter XXXVI
Spatial Data Integration Over the Web

Laura Diaz
Universitat Jaume I, Spain

Carlos Granell
Universitat Jaume I, Spain

Michael Gould
Universitat Jaume I, Spain

INTRODUCTION

Spatial data are increasingly becoming available on the Internet in applications such as routing portals that involve map-based and satellite imagery backgrounds, allowing a large audience to access and share the rich databases that are currently used by the specialized geographic community. These spatial data are heterogeneous, being available in various formats, and stored in disparate formats (flat files, relational or object-oriented databases, etc.). Some data are structured according to well-established data modeling techniques such as the relational or object-oriented data models; other data, such as data maintained in various information systems, spreadsheets, or Internet repositories, are in proprietary formats, semi-structured, or unstructured. In practice, this situation of multiple models and schemas combined with the difficulty for establishment agreements for data representation in the application domains becomes spatial data in special regarding other types of scientific data, making the interoperability problem a nontrivial task (Lemmens, Wytzisk, de By, Granell, Gould & van Oosterom, 2006). In addition to the scale of data integration, the complex and heterogeneous query processing and domain-specific computational capabilities supported by these sources make spatial data integration a real challenge (Boulcema, Essid & Lacroix, 2002; Devogele, Parent & Spaccapietra, 1998; Goodchild, Egenhofer, Fegeas & Kottman, 1999).

Historically, due to specialized characteristics and the nature of spatial data, geographic information systems (GISs) were managed separately from existing database systems. As first steps to spatial data integration in the mid-1990s, advances in database technology enabled accommodating spatial data in relational databases, allowing organizations to take the first steps toward enterprise GIS and the elimination of organizational "spatial data islands" (ESRI, 2003). Some examples are the appearance of Oracle Spatial (www.oracle.com), PostgreSQL with the PostGIS extension (www.postgresql.org), and MySQL (www.mysql.com). The early work for spatial data integration in database systems focused on sharing simple spatial features in a relational database. Then, standard data manipulation languages such as SQL (Structured Query Language) began to adopt common spatial functionalities to embed, for example, spatial selections and topological queries in SQL statements. The arrival of the first relational models capable of storing both spatial and attribute data led to spatial databases (Rigaux, Scholl & Voisard, 2001), which provided methods for spatial data modeling, algorithms, access methods, and query processing extending traditional database systems.

The success factor of Web services technology has permitted promoting service integration and interoperability among heterogeneous distributed information sources. The GIS approach to service-oriented architecture (SOA) is represented by the Spatial Data Infrastructure (SDI) paradigm, which offers the possibility to access distributed, heterogeneous spatial data through a set of policies, common rules, and standards that facilitate interconnecting spatial information users in an interoperable way (Granell, Gould, Manso & Bernabé, 2007b).

Several approaches and techniques exist for data integration over the Web (Fileto, 2001); however, some of the most representative are the following: *Gateways* as middleware that allows an application running in one DBMS (Data Base Management System) to access data maintained by another DBMS; and *Data Warehouses*, which are separated databases built specifically for decision support, useful when decisions depend on heavy analysis of large amounts of data collected from a variety of possibly heterogeneous data sources. In the GIS domain, most data integration approaches have included one or both of the more common techniques: *wrapper* (Roth & Schwarz, 1997*)* and *mediator* (Wiederhold, 1992). Wrappers provide interfaces for accessing concrete data sources, retrieving the results, and translating them into a common scheme. Mediators are software components in the middleware in charge of specifying such a common scheme that provides an integrated interface to a set of wrappers; indeed, underlying data sources. Data integration approaches use mediators to handle client queries, submit them directly to the wrappers, and integrate the results before delivering the response to client applications. GIS systems manage data integration mostly by using wrappers and mediators implementing standard interfaces specified by Open Geospatial Consortium (OGC), which provides open standards and specifications for supporting standardized information exchange, sharing, accessing, and processing geospatial data.

The demand for interoperability has boosted the development of standards and tools to facilitate data transformation and integration. Furthermore, this chapter focuses on interface standards as key to spatial data syntactical integration over the Web. Nevertheless, there are still many challenges to be met, especially those concerned with data semantics and harmonization of interoperating systems.