IDEA GROUP PUBLISHING
701 E. Chocolate Avenue, Suite 200, Hershey PA 17033-1240, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.idea-group.com
ITB10779

Chapter II

Integrating Web Data and Geographic Knowledge into Spatial Databases

Alberto H.F. Laender, Federal University of Minas Gerais, Brazil

Karla A.V. Borges,
Federal University of Minas Gerais, Brazil & PRODABEL, Brazil

Joyce C.P. Carvalho, Federal University of Minas Gerais, Brazil

Claudia B. Medeiros, University of Campinas, Brazil

Altigran S. da Silva, Federal University of Amazonas, Brazil

Clodoveu A. Davis Jr.,
PRODABEL, Brazil & Catholic University of Minas Gerais, Brazil

Abstract

With the phenomenal growth of the World Wide Web, rich data sources on many different subjects have become available online. Some of these sources store daily facts that often involve textual geographic descriptions. These descriptions can be perceived as indirectly georeferenced data – for example, addresses, telephone numbers, zip codes and place names. In this chapter we focus on using the Web as an important source of urban geographic information and propose to enhance urban Geographic Information Systems (GIS) using indirectly georeferenced data extracted *from the Web. We describe an environment that allows the extraction of geospatial data from Web pages, converts them to XML format and uploads the converted data into spatial databases for later use in urban GIS. The effectiveness of our approach is demonstrated by a real urban GIS application that uses street addresses as the basis for integrating data from different Web sources, combining the data with high-resolution imagery.*

Introduction

With the popularization of the Web, a huge amount of information has been made available to a large audience (Abiteboul, Buneman, & Suciu, 2000). In some cases, the information available in a Web site, such as pages containing information on restaurants, theaters, movies and shops, concern mostly communities that dwell in a specific neighborhood (Buyukkokten, Cho, Garcia-Molina, Gravano, & Shivakumar, 1999). Furthermore, these sites often provide indirectly georeferenced data such as addresses, telephone numbers, zip codes, place names and other textual geographic descriptions. By *indirectly georeferenced data* we mean spatial data with no associated coordinate (x,y) data. Nevertheless, this kind of data can be converted to positional data, using, for example, address matching functions (Arikawa, Sagara, & Okamura, 2000). Indeed, it can be observed that indirectly georeferenced data abound on the Web. Thus, under this perspective, the Web can be seen as a large geospatial database that often provides up-to-date, regionally relevant information.

In spite of being publicly and readily available, Web data can hardly be properly queried or manipulated as, for instance, data available in traditional and spatial databases (Florescu, Levy, & Mendelzon, 1998). Almost all Web data are unstructured or semistructured (Abiteboul et al., 2000), and cannot be manipulated using traditional database techniques. Web sources are usually constructed as HTML documents in which data of interest (for example, public facilities) is implicit. The structure of these documents can only be detected by visual inspection and is not declared explicitly. In most cases, such data are mixed with markup tags, other strings and in-line code; the structure of most data on the Web is only suggested by presentation features. Besides, when looking for specific information on the Web, users are generally faced with the problem of having to access various distinct and independent Web sites to obtain scattered complementary pieces of information. Typically, this occurs in situations where the required information cannot be found in a single Web source. For example, suppose many different Web sites provide information about restaurants in a city, each site with its own informational content. Someone who wants to get