The Web as a Data Source for Spatial Databases

KARLA A. V. BORGES[1,4], ALBERTO H. F. LAENDER[1], CLAUDIA B. MEDEIROS[2],
ALTIGRAN S. DA SILVA[3] AND CLODOVEU DAVIS JR.[4]

[1]UFMG - Federal University of Minas Gerais – Department of Computer Science
Av. Presidente Antônio Carlos, 6627, 31270-901, Belo Horizonte, MG, Brazil
{kavb,laender}@dcc.ufmg.br
[2]UNICAMP - University of Campinas – Institute of Computing
Caixa Postal 6176, 13081-970, Campinas, SP, Brasil
cmbm@ic.unicamp.br
[3] Federal University of Amazonas – Department of Computer Science
69077-000, Manaus, AM, Brazil
alti@dcc.fua.br
[4] Prodabel – Empresa de Informática e Informação do Município de Belo Horizonte
Av. Pres. Carlos Luz, 1275 - 31230-000, Belo Horizonte, MG, Brazil
{karla,Clodoveu}@pbh.gov.br

Abstract. With the phenomenal growth of the WWW, 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 - e.g., addresses, telephone numbers, zip codes and place names. Under this perspective, the Web becomes a large geospatial database, often providing up-to-date local or regional information. In this work 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 these data with high-resolution imagery.

1 Introduction

With the phenomenal growth of the World Wide Web, rich data sources on many different subjects have become available online [15]. Some of these sources are relevant primarily to communities within a specific geographic region. For instance, Web sites containing information on restaurants, theaters, movies, and shops concern mostly Web users who dwell in the neighborhood of these locations [4]. Furthermore, they often include 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 [3]. Indirectly georeferenced data abound on the Internet. Thus, 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, traditional and spatial databases [8, 15]. To manipulate Web data more efficiently, some researchers have resorted to ideas taken from database techniques. Web sources are usually formed as HTML documents in which data of interest (e.g., public facilities) is implicit. Their structure 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. Thus, it is difficult to gather and to use only the data of interest. Furthermore, the structure of most data on the Web is only suggested by presentation features. Therefore, almost all Web data are unstructured or semistructured [1], and cannot be manipulated using traditional database techniques. In order to overcome this problem, a possible strategy is to extract data from Web sources to populate databases for further handling, for instance, by using special programs called *wrappers* [8, 15]. As shown in this paper, an analogous strategy can be applied to extract geographic context from Web pages to populate spatial databases, thus providing means for supporting new location-based Web services.

In this work we focus on using the Web as an important information source for human and urban geographic information. Information about cities is currently being accumulated as online digital contents both in an urban Geo-

graphic Information System (GIS) and in local Web pages. An urban GIS stores information about cities, including geographic attributes; Web pages store daily life information relevant to local Web users [10]. The idea behind our work is that it is possible to enhance an urban GIS using indirectly georeferenced data and information extracted from the Web. The resulting data can be used to build new GIS applications or to update spatial databases.

It is important to notice that both data acquisition and updating in urban environments are costly and time-consuming. In developed countries, especially in the USA, there are usually governmental nation-wide efforts to generate and to maintain address databases. The US Census Bureau, for instance, maintains and distributes at a very low cost its TIGER (*Topologically Integrated Geographic Encoding and Referencing*) files, in which street addresses are coded as a set of attributes for segments of street centerlines. The result is a considerable amount of freely available structured geospatial data. In emergent countries, on the other hand, the situation is quite the opposite because of the associated costs and the lack of policies that enforce the updating and integrity of geographic databases. In Brazil, the collection of geospatial data has been systematically hampered by budget limitations. Brazilian local governments have to oversee large areas, and their budgets must respond to priorities other than updating spatial databases. Therefore, there is a need for ingeniuos solutions to implement low-cost geographic data collections. The fact that Brazil has the largest Web network in Latin America, along with the increasing number of local governments with dedicated Web pages, suggests that our solution can be effectively implemented and will help diminish the lack of up-to-date geographic information at the regional level.

In this paper we present an environment that allows the extraction of geospatial data from Web pages. The solution also makes the conversion to a suitable format (in our implementation, XML) and uploads the converted data into spatial databases for later use in urban GIS. Our approach focuses specifically on the integration of data from distinct Web sources, urban GIS, and high-resolution imagery, using street addresses as the basis for integration. This solution has several advantages. First, addresses are natural keys to accessing urban GIS and they are most common the form of spatial localization used by general public in Web pages. Second, in urban planning it is difficulty to closely monitor the evolution of citywide activities and phenomena and, by using our solution, Web data sources can be used as geographic knowledge bases. Furthermore, it provides a simple and inexpensive solution for keeping geospatial databases up-to-date.

The remainder of this paper is organized as follows. Section 2 discusses related work. Section 3 describes our approach to extract address data from Web sources. Section 4 describes an implemented case study, which integrates Web data into GIS in a real application for the city of Belo Horizonte, Brazil. Information from Web sources is combined with high-resolution imagery, to enhance an urban GIS. Finally, Section 5 presents conclusions.

2 Related Work

Recently, many research efforts have been conducted on the recognition and use of geospatial information from Web sites. It has been demonstrated that discovery and exploitation of geographic information in Web pages is quite feasible, and exploitation of such information provides a useful new paradigm for the navigation and retrieval of Web information [17]. Some techniques are proposed for extraction of the geographical context of web pages, based on the occurrence of text address and post codes, place names and telephone numbers. Kambayashi et al. [13] divide these efforts into three categories. The first category – *map-enhanced Web applications* – uses maps as a user-friendly interface for the Web, thus making it possible to handle geographic data through usual Web browsers [10, 12, 17, 18]. The second category exploits geographic location information that is found on Web pages. This information consists of place names, latitude/longitude pairs, postal addresses, and so on, used to classify and to index Web pages [3, 4, 6, 12, 18]. The third category focuses on the integration of Web information and geographic knowledge [13, 12]. Some approaches belong to more than one category. In addition to the initiatives referred in this section, there are some commercial sites which have recently started offering a geographic search capability. In these sites it is possible to locate places of interest in the vicinity of a given address and to navigate to their Web sites [7, 9, 11, 19, 20]. Yet, these sites have been built to locate business Web pages, which are previously stored in the search site's database. Furthmore, they cannot recognize alternative names for the same place or informal names.

Our approach differs from the ones just mentioned because we use Web data as a source for the improvement of geographic knowledge on the city, which includes being able to populate and enrich urban GIS databases using information extracted directly from the Web. Our main motivation is to take advantage of Web data, a rich source of local information, as well as to offer an alternative for the creation of new GIS data, since data acquisition costs are a very important issue [16].

3 Obtaining Spatial Information from Web Sources

To make it possible to create an environment to integrate Web pages to spatial location information, we had to meet several challenges. The first was to extract indirectly georeferenced data in textual form (such as postal addresses)

from the contents of Web pages. We stress that such information, when available, is implicit and occurs as any other ordinary string mixed with HTML markup. In GIS, the process of recognizing geographic context is referred to as *geoparsing*, and the process of assigning geographic coordinates is referred to as *geocoding* [17]. This section discusses the efforts to geoparse and to geocode Web pages. The extracted addresses act as keys to the geocoder.

The second challenge was to establish ways for transforming the extracted spatial location information in the form they are provided by the generic public to the form they are stored in a typical GIS. After that, a set of geographic coordinates corresponding to the addresses can be obtained, using an address matching function. Finally, the extracted information was inserted into the GIS database and superimposed on high-resolution imagery or maps.

The basic procedure for our application is (1) to crawl Web sites to collect the pages containing data of interest, (2) to geoparse the collected pages to extract geographic indication and the relevant data, (3) to make the data available in an suitable format (in our case, XML), (4) to geocode the addresses into a coordinate system, (5) to update the GIS database and, finally (6) to integrate information from several geospatial data. The resulting system can be used by municipalities, users with some kind of urban GIS database, or geographic database designers.

This section describes how step 2 can be accomplished by deploying the DEByE (Data Extraction By Example) [14] example-based approach to automatically extract semistructured data. This approach is more convenient for our application because it lets the user specify a target structure for the data to be extracted. Furthermore, the user might be interested in only a subset of the information encoded in the page. Moreover, DEByE does not require the user to describe the inherent structure of a whole page.

3.1 The DEByE Tool

DEByE is a tool that has been developed by the UFMG Database Group to generates wrappers for extracting data from Web pages. It is fully based on a visual paradigm, which allows the user to specify a set of examples for the objects to be extracted. Example objects are taken from a sample page of the same Web source from which other objects (data) will be extracted. By examining the structure of the Web page and the HTML text surrounding the example data, the tool derives an *Object Extraction Pattern* (OEP), a set of regular expressions that includes information on the structure of the objects to be extracted and also the textual context in which the data appear in the Web pages. The OEP is then passed to a general-purpose wrapper that uses it to extract data from new pages in the same Web source, provided that they have structure and contents similar to

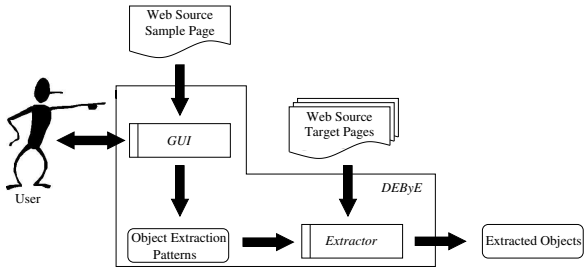

Figure 2: Modules of the DEByE tool and their role in the data extraction process

the sample page, by applying regular expressions and some structuring operations [14]. DEByE currently operates as a Web service, to be used by any application that wishes to provide data extraction functionality to end users. For general data extraction solutions, a DEByE interface based on the paradigm of nested tables is used, which is simple, intuitive, and yet powerful enough to describe hierarchical structures that are very common in data available on the Web. The sample pages are displayed in the upper window, also called the *source window*. The lower window, also called the *table window*, is used to assemble example objects. The user can select pieces of data of interest from the source window and "paste" them on the respective columns of the *table window*. After specifying the example objects, the user can select the "Generate Wrapper" button to generate the corresponding OEP, which encompasses structural and textual information on the objects present in the sample page. Once generated, this OEP is used by an extractor module that will perform the actual data extraction of new objects and then will output them using an XML-based representation. DEByE is also capable of dealing with more complex objects, by using a so-called *bottom-up* assembly strategy, explained in [14]. Figure 1 shows a snapshot of a user session with an example object in the lower window (*table window*) and the extracted objects showed in HTML format in the upper window. Figure 2 presents an overview of the whole DEByE approach. The two modules called *Graphical User Interface* (GUI) and *Extractor* compose the DEByE tool.

3.2 Geocoding Process using Addresses

One obvious source of geospatial information is the postal address, which is universally used to facilitate the delivery of physical mail to a specific location around the world. Though recognition of addresses is a fairly studied problem, it is complicated by the fact that formatting standards vary considerably from one country to another [17]. Furthermore, in the same country it is common to have variations for the encoding of the same address. The postal address may or may not have fields for each addressing component,

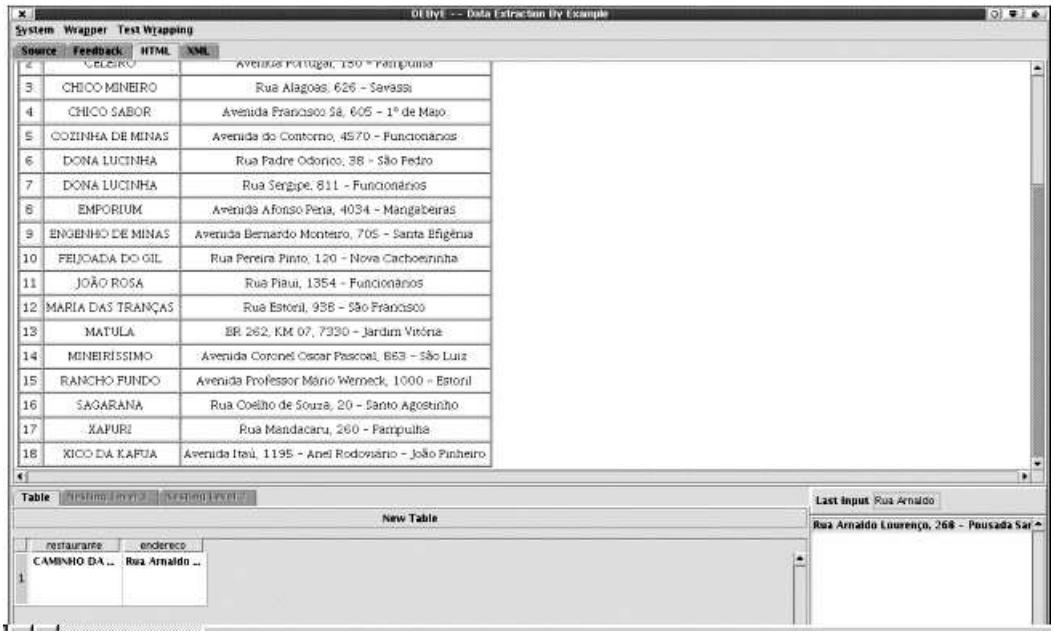

Figure 1: Snapshot of an example specification session with the DEByE Interface

such as thoroughfare type (street, avenue, plaza, and so on), thoroughfare name, building number, neighborhood, city, state, country, zip code, and possibly others. Thus, recognizing address data from Web pages is complicated, since there are no fixed rules for specifying address fields.

Addresses in Web pages are typically segmented into comma-delimited fields or line breaks, and sometimes information such as the country is omitted. This broad variation in abbreviations, punctuation, line breaks, and other features that are used to express the same address makes the parsing process more complicated [3, 17]. Even though our approach takes advantage of user-provided examples to recognize and to extract addresses it is not easy to separate the fields that compose the address correctly. Therefore, our strategy is to extract the address without worrying about dividing it into fields. Once an address has been extracted, it must be parsed into a consistent format in order to be searched for in the addressing database. Thus, we postpone the parsing problems that normally arise until we get to the geocoding step. The most frequent problems include misspellings, format variations, different names used for the same location, and coincidental names for different thoroughfares. The geocoding process includes three phases: (1) the treatment of the semi-structured alphanumeric addresses that have been extracted from the Web, (2) the establishment of a correspondence between the structured address and the addressing database (the *matching* phase), and (3) the actual assignment of coordinates to the event (the *location* phase).

Starting from structured addresses, actual geocoding can be performed in several ways, depending on the available addressing information. In order to be able to perform the parsing, matching, and locating tasks, the geocoding process needs to have access to a database in which information about the addressing system are stored. There are two basic categories of information in such a database. The first category is comprised of the actual addressing infrastructure, with objects such as point-georeferenced individual addresses and street centerlines with address ranges. The second includes any additional information items that can be used to resolve ambiguities or as a rough geographic reference in case the address, for any reason, cannot be located in the first category. This includes elements such as all sorts of spatial reference units (area objects that correspond to artificial borders, such as neighborhood limits, districts, ZIP areas, municipal divisions, and so on), along with a catalog of reference points known by the citizens. This catalog can contain what we call "reference places", i.e., popularly known spots in a city that can be referenced by name only, points that are so easily recognized by the population that their location does not require a formal address. Of course, the addressing database can be rather incomplete, depending on the available data about a given city or location. Our goal is to accommodate this by trying to geocode at the most precise level first, and, if that is not possible, successively resorting to less precise geocoding methods until some location can be established.

Since we do not assume any particular structuring in the incoming address, we must be able to determine the structure by analyzing the string of text corresponding to

it. The objective of this process is to create a structured tuple containing every significant piece of information from the original address string. If necessary, addressing elements found in the string are normalized or adjusted before becoming fields in the tuple. This process is called the *geoparsing* of the original address. The algorithms that can be used in the parsing of the address are very similar to the ones used in programming languages in order to assess the syntax of a language construct. The string gets initially divided into tokens, considering whitespace characters (blanks, commas, points, hyphens, and so on) as delimiters. The resulting set of tokens is then analyzed sequentially, in an attempt to determine the function of each one of them. The analysis of each token uses the addressing database, in order to establish hypotheses as to what is the function of each term (token) in the original address. The token functions we look for are (1) Thoroughfare type: street, avenue, plaza, boulevard and so on, along with their usual abbreviations; (2) Thoroughfare name: the name popularly associated with the thoroughfare; thoroughfare names can also have shortened versions, popular nicknames, and older names that need to be taken into consideration; (3) Street number: number that is usually posted at the door of each building, to indicate a sequence within the thoroughfare; (4) Neighborhood : name of any intra-municipal division that is used to identify distinct regions within the city's limits and (5) additional data, such as city name, state, and postal code. The result of the parsing is a set of fully structured addresses, in which there are fields for each of the components identified. One a postal address has been recognized and parsed it must be geocoded into coordinate system such as latitude and longitude. The matching function requires a thoroughfare code for each of address extracted and parsed. All the problems previously mentioned must be solved in this phase. This is important because there can be redundant names, i.e., more than one street can have the same name, possibly in different neighborhoods. Also, there is no guarantee that the thoroughfare names resulting from the parsing process are complete, correct, or even existent. However, there can be situations in which the thoroughfare name alone cannot determine a single thoroughfare code, like in the case of homonymous streets. We must then be able to resort to additional information, such as a postal code or a neighborhood name, in order to establish the correct thoroughfare code association.

After the geocoding task, data objects extracted from Web sites can be stored in a spatial database. These objects represent entities in the real world, like restaurants, hotels and museums. Each object has a set of attributes (e.g., name, street, phone, URL) and a position in the GIS.

4 An Application Experience - The Case of Belo Horizonte

We chose to work with a spatial database from the local government of Belo Horizonte. Belo Horizonte was one of the first Brazilian municipal administration to develop an urban GIS. The city's database includes 400,000 individual addresses and 0.40-meter resolution images. These two factors, plus the vast amount of information about the city on the Internet, enabled us to develop a prototype application for Belo Horizonte to validate our proposal.

The data integrated by our application comes from six different sources: Belo Horizonte's high-resolution imagery (available at www.belohorizonte.com.br), Belo Horizonte urban GIS data, and four distinct Web sites (www.passeiolegal.com.br, www.terra.com.br, www.inbh.com.br, www.comidadibuteco.com.br). The selected sites provide information about hotels, eleven categories of restaurants, pubs, museums and other cultural attractions, consulates, advertising agency, flower shops, movie theaters, theaters, clothing stores, libraries, hospitals and emergency room. A subset of pages available in each site was collected amounting 65 pages and 540 objects.

We next used the DEByE tool to parse the collected pages, extracting the names of points of interest and their addresses (Figure 1). The set of extracted data were then coded in an XML-based format.

In order to geocode the extracted addresses, we must transform them into the format in which they are stored in the Belo Horizonte's addressing database. The result is a set of fully structured addresses containing (1) thoroughfare type (street, avenue, plaza, boulevard and so on), (2) thoroughfare name, (3) building number, (4) neighborhood or other types of complementary information.

Each postal address recognized and parsed was next geocoded using an address matching function. Finally, all data extracted from the Web pages where these addresses were recognized were stored in a spatial database and assigned (x,y) coordinate points.

As a result, we extended Belo Horizonte GIS database with twenty eight news tables. The attributes of these tables were: place name, thoroughfare type, thoroughfare name, street number, neighborhood and individual address assigned code. The results of experiments are summarized in Table 1. Column "Pages" contain the number of pages collected in each site, column "Objects" contain the number of objects in each site, column "Extracted" contain the number of objects extracted, column "Exact" contain the number of the addresses matched with exact locations in the addressing database, column "Close" contain the number of addresses which were placed at the numerically closest address on the same street or at an approximate location, based on a reference point, and the last column "Not Found" contain the number of addresses which could not be

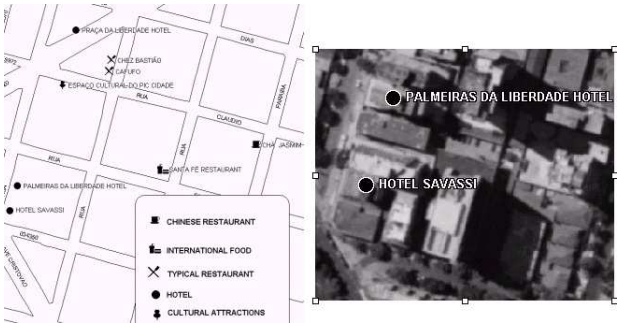

Figure 3: Integration of urban information from Web pages with geographic databases and high-resolution imagery in a GIS environment

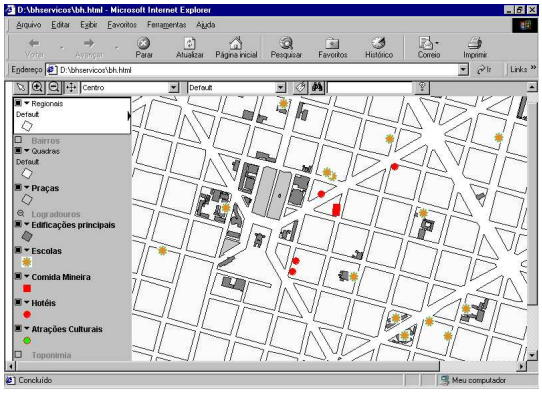

Figure 4: Web browser showing hotels, typical restaurants and cultural attractions extracted from Web sources and Schools extracted from spatial database

located. As for the geocoding, 90% of the addresses were matched by exact locations. The remaining 8% were placed at the numerically closest address or at a reference point. Some objects could not be located because their addresses were incomplete or inexistent (2%).

The new geographic features were overlapped to high-resolution imagery in the GIS, thus allowing for many kinds of maps to be produced by integrating points located from the Web sites to existing GIS data (Figure 3). This allowed integrated browsing of all facilities within an area. For example, the Web site www.passeiolegal.com.br (one of our Web data sources) provides information about restaurants and hotels. However there is a specific page for each cuisine type and a separate page for hotels. This prevents a user to get information on all restaurants near a hotel, since the site allows the user to search only one subject at a time. Our solution solves this problem. Another form to visualize all extracted data is a new geographic page. This approach provides an opportunity to create a new information space for everyday life. We use Alov Map [2], a free software, to publish our GIS data in the internet (Figure 4).

5 Conclusion and Future Work

In this work we focus on using the Web as an important source of urban geographic information and propose to enhance urban 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. Our proposal is centered on the integration of urban information from local Web pages with geographic databases and high-resolution imagery in a GIS environment. All Web pages that refer to public spaces including, for instance, restaurants, schools, hospitals, shopping centers, and theaters, can be collected, their data extracted and associated with the city's map. Integration with existing GIS data will allow, for instance, urban planners to have a more realistic view of the city, with the actual distribution of its services. 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 them with high-resolution imagery. Although still preliminary, the results obtained with our application prototype are encouraging.

Our future work includes the use of ontology-based tools to automatically recognize the indications of the urban geographical context of Web pages, including the recognition of addresses, ZIP codes, reference place names, and popular names for urban locations, for which the user would not have to provide examples as shown here. Another line of work involves proposing a way to assign geographic locations to local Web pages, in which the location of the page's subject is stored within the HTML code. This can provide means to index Web pages according to their geographical location(s). In this approach, coordinates or other forms of geographic reference can be retrieved from Web pages and be included in a spatial index. This spatial index can be used to improve the retrieval process; users would be able to provide the usual keywords, along with place references in which their interest lies. This can also have an important effect on the tools and resources that can be used to update spatial databases, using information available in the Internet for that. We are also working on integrating data extracted from multiple sources [5] in order to improve the geocoding process. Web sources might overlap (an attribute may be available from several sources) or contain replicated data. An appropriate integration of such sources will contribute to decrease the efforts of geocoding, thus improving data quality.

Table 1: Experimental Results

Site	Pages	Objects	Extracted	Exact	Close	Not Found
www.passeiolegal.com.br	12	122	122	120	-	2
www.terra.com.br	2	52	52	48	4	-
www.inbh.com.br	38	277	277	237	33	7
www.comidadibuteco.com.br	13	89	89	83	6	-

References

[1] ABITEBOUL, S., BUNEMAN, P., AND SUCIU, D. *Data on the Web - From Relations to Semistructured Data and XML*. Morgan Kaufman Publishers, San Francisco, California, 2000.

[2] ALOV. Alov Map. In *available online at http://alov.org/index.html*.

[3] ARIKAWA, M., SAGARA, T., AND OKAMURA, K. Spatial Mesia Fusion Project. In *Proceedings of Kyoto International Conference on Digital Library: Research and Practice* (Kyoto, Japan, 2000), pp. 75–82.

[4] BUYUKKOKTEN, O., CHO, J., GARCIA-MOLINA, H., GRAVANO, L., AND SHIVAKUMAR, N. Exploiting Geographical Location Information of Web Pages. In *Proceedings of ACM SIGMOD Workshop on the Web and Databases, WebDB'99* (Philadelphia, Pennsylvania, 1999), pp. 91–96.

[5] CARVALHO, J. C. P., AND DA SILVA, A. S. Finding Similar Identities among Objects from Multiple Web Sources. In *Proceedings of the Fifth International Workshop on Web Information and Data Management* (New Orleans, LA, 2003 (to appear)).

[6] DING, J., GRAVANO, L., AND SHIVAKUMAR, N. Computing Geographical Scopes of Web Resources. In *Proceedings of the 26th International Conference on Very Large Database, VLDB'00* (Cairo, Egypt, 2000), pp. 545–556.

[7] DOTGEO. .geo. In *available online at http://www.dotgeo.org*.

[8] FLORESCU, D., LEVY, A., AND MENDELZON, A. Database Techniques for the World-Wide Web: A Survey. *SIGMOD Record 27*, 3 (1998), 59–74.

[9] GEOSEARCH, N. L. NorthernLigth Search Engine. In *available online at http://www.northernlight.com*.

[10] HIRAMATSU, K., AND ISHIDA, T. An Augmented Web Space for Digital Cities. In *Proceedings of Symposium on Applications and the Internet (IEEE), SAINT-2001* (San Diego, California, 2001), pp. 105–112.

[11] INFOSPACE. infospace.com. In *available online at http://www.infospace.com*.

[12] JONES, C. B., PURVES, R., RUAS, A., SANDERSON, M., SESTER, M., VAN KREVELD, M., AND WEIBEL, R. Spatial Information Retrieval and Geographical Ontologies. An Overview of the SPIRIT Project. In *SIGIR 2002: The 25th Annual International ACM SIGIR Conference on Research and Development Information Retrieval* (Tampere, Finland, 2002).

[13] KAMBAYASHI, Y., CHENG, K., AND LEE, R. Database Approach for Improving Web Efficiency and Enhancing Geographic Information Systems. In *Proceedings of International Conference on Internet Information Retrieval, 2001 IRC* (Korea, Japan, 2001).

[14] LAENDER, A. H. F., RIBEIRO-NETO, B. A., AND DA SILVA, A. S. DEByE Data Extraction By Example. *Data and Knowledge Engineering 40*, 2 (2002), 121–154.

[15] LAENDER, A. H. F., RIBEIRO-NETO, B. A., DA SILVA, A. S., AND TEIXEIRA, J. S. A Brief Survey of Web Data Extraction Tools. *SIGMOD Record 31*, 2 (2002), 84–93.

[16] LAURINI, R. *Information Systems for Urban Planning - a hypermedia co-operative approach*. Taylor Francis Inc., New York, New York, 2001.

[17] MCCURLEY, K. S. Geospatial Mapping and Navigation of the Web. In *Proceedings of WWW10* (Hong Kong, 2001), pp. 221–229.

[18] THAKKAR, S., KNOBLOCK, C. A., AMBITE, J. L., AND SHAHABI, C. Dynamically Composing Web Services from On-line Sources. In *Proceedings of AAAI-02 Workshop on Intelligent Service Integration* (Edmondon, Canada, 2002).

[19] WHEREONEARTH. whereonearth.com. In *available online at http://www.whereonearth.com*.

[20] YAHOO. Yahoo yellow pages. In *available online at http://www.yp.yahoo.com*.