Accessing Diverse Geo-Referenced Data Sources with the SAND Spatial DBMS

Jagan Sankaranarayanan Egemen Tanin Hanan Samet Frantisek Brabec

Department of Computer Science
Center for Automation Research
Institute for Advanced Computer Studies
University of Maryland at College Park
{jagan, egemen, hjs, brabec}@cs.umd.edu

Abstract

The Internet has become the most frequently accessed medium for obtaining various types of data. In particular, government agencies, academic institutions, and private enterprises have published gigabytes of geo-referenced data on the Web. However, to obtain geo-referenced data from the Web successfully, systems must be designed to be capable of understanding the data sets published in different data formats. Also, even if the data sets are available in a simple known format, they often have poorly defined structures. With these issues in mind, we have developed an Internet-enabled data collection and conversion utility that interfaces with our prototype spatial database system, SAND. Using this utility, data can be retrieved from many different sources on the Web and converted into a format understandable by the SAND spatial database management system. Our collection and conversion utility is able to import the most popular data formats; namely, ESRI Shapefiles, Microsoft Excel files, HTML files, and GML files. Data in unstructured formats are verified for correct selection of the data types and handling of missing tuples before the insertion operation into the database. Moreover, our utility makes it possible to download any nonspatial data set and combine it internally with a relevant spatial data set. These features are accessible through a spreadsheet-like interface for online editing and structuring of data.

1. Introduction

The Web, with a variety of standards and data formats, comprises of a vast collection of data. Although large volumes of useful data are available on the Web, collecting and representing them in a useful way remains a challenge. Data sets on the Web are available in too many different open and proprietary formats. Also, the data sets may be unstructured, e.g., spreadsheets and HTML files. In addition to this, for our application area, with spatial data, the data sets may not have some of the necessary spatial information bundled with them. For example, data sets may describe a spatial object in textual form but may not necessarily contain the proper locational references. An example for such a data set would be an Excel spreadsheet containing the U.S. county population information and the county names, but without the description of the shapes and the geographic locations of the counties. Thus, there is a clear need for tools that allow online manipulation of data to combine nonspatial with spatial data so that more useful data sets can be created and visualized. Hence, we wanted to design a data conversion utility that will target all the underlying complex issues involved in accessing different data types in an efficient and simple manner for the end-user. In particular, our collection and conversion utility is able to import the most popular data formats; namely, ESRI Shapefiles, Microsoft Excel files, HTML files, and GML files. We built this utility on top of our prototype spatial database system, SAND. With our utility, users of the SAND spatial database system, can easily access multiple data sets located at different sites in different formats and simultaneously work on them for creating meaningful results. The rest of the paper is organized as follows: Section 2 provides an overview of our prototype spatial database system that we built our conversion utilities on. Section 3 provides an overview of some common problems found in various data formats available on the Web and the main approaches to converting the common data formats. Section 4 describes our conversion utility in more detail. Section 5 contains our conclusions.

2. SAND

SAND (Spatial And Nonspatial Data) [Esperanca and Samet, 2002, Samet et al., 2001, Samet et al., 2003] is a spatial database system developed at the University of Maryland. SAND has powerful features that allow sophisticated queries to be issued and useful inferences to be drawn on spatial data. SAND enables the easy exploration of data sets that have spatial attributes in them. It can handle multi-dimensional data sets like river paths, country boundaries, city locations, and can respond to queries involving complex spatial data, e.g., finding the closest counties near a particular spatial location. The SAND Internet Browser is a front end to the SAND spatial database system. It takes in user requests and sends them to the database server over the Internet. Results are then sent back to the client and displayed. The queries are defined interactively and graphically. Our conversion utility is added to the SAND framework through the SAND Internet Browser.

3. Spatial Data on the Web

We have developed many conversion functions in our utility that will allow import of data from various external formats:

3.1 Shapefiles

A rich source for spatial data is the Shapefile format from ESRI [ESRI]. This format is widely used and readily available on the Web. Our converter can import data sets bundled with Shapefiles [GEOTOOLS]. Shapefiles are structured data files where spatial information for geographical locations is stored. So any data set that uses a Shapefile can now be imported into the SAND database system. They are well structured and can easily be understood by most conversion programs.

3.2 Excel Files

Another common format that we explored is Microsoft's Excel spreadsheet format. The spreadsheet format of Excel is quite similar to SAND's internal binary format. The conversion process is made to look simple to the end user by automating many interim steps. Yet, there are a few issues that we have to deal in Excel files. Excel does not enforce strict type checking or any form of data integrity. Excel data sets may have compound tables. The attribute values of some of the tuples may be missing or may not have consistent data types. Some of the data types supported by Excel may not have their equivalents in SAND (e.g., date). Also, Excel files are used for many purposes other than spatial data. Hence, these files do not necessarily have to have a proper spatial data component in them and they may contain many other non-convertible items. For generality, we assumed that the users familiar with spreadsheets will store data in tables arranged in rows and columns. Typically, a row represents a tuple in the data set and a column represents an attribute. Each column has a data type (like string for city names, integer for populations, etc.) that should map to the internal data types of SAND. To provide a successful conversion process, some data modification would have to do be done online by the user. Hence, we had to design an online editing facility where multiple functions are provided to the users to enable them to resolve discrepancies in the conversion process. This forms the basis of our approach to converting such unstructured data.

3.3 HTML Files

HTML files form the most common data source on the Web. Currently, we restrict ourselves to porting data that are stored as HTML tables. Similar to the Excel files, HTML data do not provide any type checking or data integrity checking. Similar measures should be taken in conversions like the ones taken for Excel files.

3.4. GML Files

Geographic Markup Language (GML) [GML] like its general purpose variant XML, is designed to become the de-facto standard for spatial data on the Web. GML is an open standard for exchange, storage, and retrieval of GIS data. GML data contain both spatial and nonspatial data bundled together. We have added GML importing capabilities to the SAND system. With its growing acceptance, useful data sets are expected to be soon available in the Web in GML format. Data files in GML may also need user guidance in some of the conversions. Currently we can convert static GML data saved as data files. Database servers that can dynamically generate GML data are not yet considered in our work. Hence, considering the rest of the research arena for XML/GML, we plan to address this issue in our future work.

4. The Conversion Utility

We have developed our utility so that it will guide the user through the conversion process. The main steps of the process are (Figure 1):

a. Locate the data source and choose the appropriate data converter,

b. Edit in the spreadsheet interface to enable users to move the data into the desired form.

c. Verify the spreadsheet and port the data.

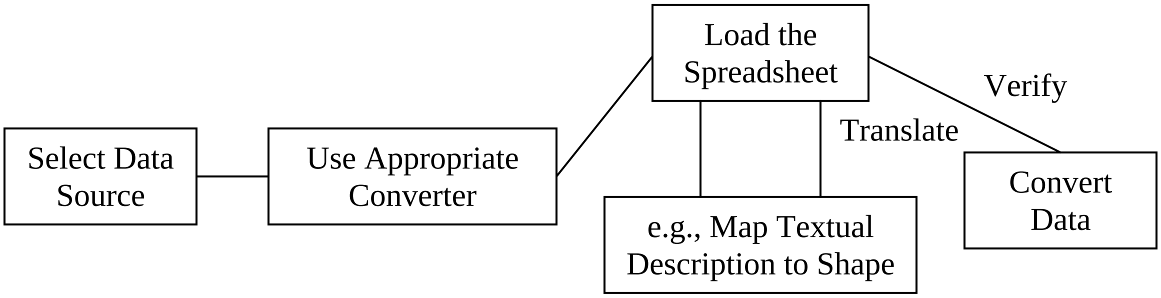
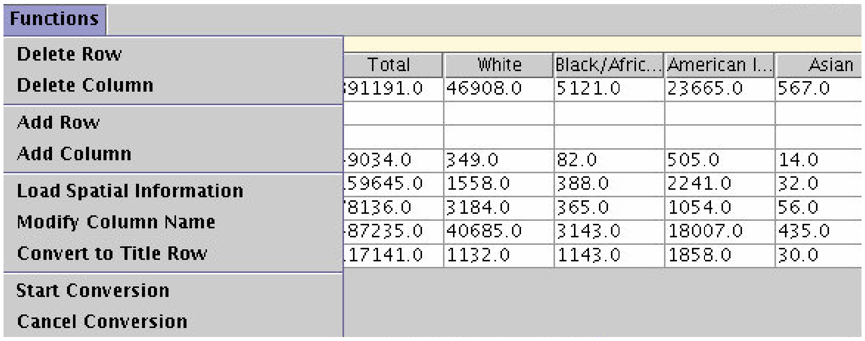


Figure 1: The conversion steps.

The spreadsheet interface provides all the functionality needed for simple editing of the data, e.g., deleting and inserting rows and columns (Figure 2), and forms the core of our converter. The spreadsheet also provides an interface to map/combine the textual information to/with spatial information.



Total	White	Black/Afric...	American I...	Asian
91191.0	46908.0	5121.0	23665.0	567.0
9034.0	349.0	82.0	505.0	14.0
59645.0	1558.0	388.0	2241.0	32.0
8136.0	3184.0	365.0	1054.0	56.0
87235.0	40685.0	3143.0	18007.0	435.0
17141.0	1132.0	1143.0	1858.0	30.0

Figure 2: A data spreadsheet allows easy manipulation of online data.

Mapping textual information to spatial information poses the most interesting issues. For example, names of places could map to many locations and shapes spatially. For example, the term Washington may map to a state in the U.S., one of more than two dozen counties across different states, the capital of the U.S., as well as numerous other entities. Our user interface allows for the selection of the correct geographical value among the different options which are displayed using their shapes (Figure 3). The resulting selection is stored in the SAND internal format.

Before the data can be ported into the database, it has to be verified. The verification step ensures that:

a. Every tuple in the table has an associated spatial component,

b. The values in a column belong to the same data type,

c. There are no missing attributes in tuples.

It then generates a list of errors, warnings, and instructions for suitable actions.

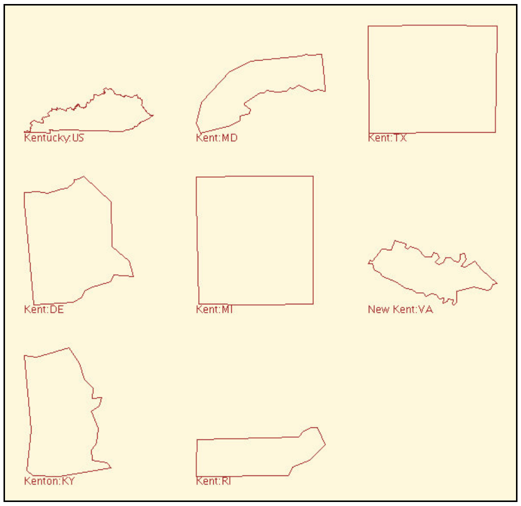


Figure 3: Shapes found for an abbreviated name; Kent. The utility helps select the right spatial option.

5. Conclusion

SAND is an ongoing research project to develop spatial browsers and spreadsheets. Importing data from various sources would prove as a useful tool to harness the full power and capabilities of the system. Hence, we designed and implemented a conversion utility for SAND. The following are the notable features of our utility:

a. The ability to combine and map non-spatial data with spatial data in a meaningful way,

b. A user-friendly spatial spreadsheet-like interface for online data editing and content creation,

c. The verification of the unstructured data for correctness of data types, lack of missing fields, etc.,

d. Converters for importing data from ESRI Shapefiles, Microsoft Excel files, HTML files, and GML files providing the users with a potentially large variety of data sets that can be downloaded from numerous sources.

We are currently working to add other data formats to our program and generalize the conversion process.

Acknowledgments

This work was supported in part by the U.S. National Science Foundation under Grants EIA-99-01636, EIA-99-00268, IIS-00-86162, and EIA-00-91474.

References

[Esperanca and Samet, 2002] C. Esperanca and H. Samet. Experience with SAND-Tcl: A scripting tool for spatial databases. Journal of Visual Languages and Computing, 13(2):229-255, April 2002.
[ESRI] Tiger Census Data in Shapefile format. http://www.esri.com/data/download/census2000_tigerline/.
[GEOTOOLS] Open source mapping toolkit. http://www.geotools.org/.
[GML] Geography Markup Language, 2.0, specification document. http://www.opengis.net/gml/01-029/GML2.html.
[Samet et al., 2001] H. Samet, F. Brabec, and G. R. Hjaltason. Interfacing the SAND Spatial Browser with FedStats Data. In 1st National Conference on Digital Government Research, Los Angeles, CA, May 2001, pages 41-47.
[Samet et al., 2003] H. Samet, H. Alborzi, F. Brabec, C. Esperanca, G. R. Hjaltason, F. Morgan, and E. Tanin. Use of the SAND Spatial Browser for Digital Government applications. Communications of the ACM, 46(1):63-66, January 2003.