Computational Approaches to Mapping and Visualizing Language Data

Daniel Parker and Benjamin Cool
Eastern Michigan University
The LINGUIST List
2000 Huron River Drive, Suite 104
Ypsilanti, MI 28197
{dan, ben}@linguistlist.org

Abstract

LL-MAP is a computational tool designed to integrate language information with data from the physical and social sciences by means of a Geographical Information System (GIS). The system will host a comprehensive set of language distribution maps, along with information on digital language resources, culture, and demographics. LL-MAP is a powerful modeling engine that applies GIS technology to spatially organized language data. This allows researchers to overlay linguistic information and data from other social sciences to create a visual representation of environmental and cultural elements that may affect a language or language group.

1 Introduction

Information about language boundaries and language relationships can provide critical insight into the genetics, culture, migrations, and interactions of populations. However, the full potential of language information to produce advances in the social sciences will only be realized when language data can be perspicuously combined with topographic, political, demographic, and historical data. New spatial social sciences technology, in the form of Geographical Information System (GIS), can organize a wide range of heterogeneous data in a highly flexible and transparent way. Hence, the LINGUIST List, in partnership with the Eastern Michigan University Institute for Geographical Research and Education (IGRE) and 8 international language archives, has developed a distributed GIS system — Language and Location – a Map Annotation Project (LL-MAP) — which dynamically maps language data against georeferenced data from the physical and social sciences in an interactive and sustainable method.

The LL-MAP project is intended as a computational tool for scholarly research, and provides:

- A major database of geospatially-referenced information about existing languages, integrated with a database of genetic relationships among languages through the Multitree project.

- Two user-friendly online interfaces which organize the linguistic, geographic, and social sciences information into customizable map layers and context-sensitive attribute displays

- Flexible tools for querying, annotating, discussing, and collecting language-related data.

- Web Map Service (WMS) and Web Feature Service (WFS) servers to make the data freely available.

LL-MAP is a powerful modeling engine that applies GIS technology to a new realm of scientific investigation. The LL-MAP system spatially organizes language data, allowing researchers to overlay linguistic information and data from other social sciences to create a visual representation of environmental and cultural elements that may affect a language or language group. The LL-MAP

project is intended for a very broad audience, from linguists to geographers. In order to accommodate different users, we have designed two interfaces for viewing data, one based on Google Maps, and the other based on ESRI Geographic Information Systems technology, which supports data querying and analysis.

While it has long been recognized that much linguistic information can be plotted on a map, attempts to integrate language data into a GIS system have been hindered by the fact that the information must be assembled from multiple sources, and identifying these often requires specialized knowledge. The LL-MAP system is able to circumvent this obstacle because of its ability to draw on the resources of a wide range of collaborators, as well as prior work in developing databases of language information.

The linguistic data made available through the LL-MAP project is harvested from a number of different projects (e.g. Ethnologue, WALS, LINGMAP). The non-linguistic data made available through LL-MAP is aggregated from public sources that employ OpenGIS protocols. Via the "Scholars' Workbench," linguists are invited to input geo-referenced data drawn from their own linguistic research; these data can be combined with data already existing in the LL-MAP database to produce new language maps, which can then be stored on the LL-MAP site, or saved and printed on the scholar's local machine.. Furthermore, as an outgrowth of its distributed architecture, the LL_MAP system will acquire unique experience in organizing a virtual community of geospatial data and service providers, and facilitating their cooperation through technical means. The project will also be innovative in showing how standards for GIS metadata can be adapted to linguistic information.

2 Computational Challenges

LL-MAP provides an easy-to-use platform for sharing and displaying linguistic data. As LL-Map is a computational tool for linguists, designing a user interface and database that would meet the needs of the linguistic community was a top priority. However, we also wanted to make the site easy to use for students as well, and thus it was decided early on to provide two interfaces to our application. We refer to these interfaces as the Standard Interface and the Advanced Interface. Both interfaces are built on 100% object-oriented Javascript code, facilitating code management and promoting code reuse. The entire application is a true Web 2.0 application; the interface and associated scripts are loaded once, and then the application runs entirely in the browser, making asynchronous requests back to the LL-Map server when needed to retrieve map images or access the database.

The Standard Interface is built on the public Google Maps APIs. These APIs provide a clean object-oriented way to build rich map applications. LL-Map further extended this with an abstraction layer to encapsulate Google map displays inside standard user interface components that behave (programatically) the same as any other UI component of the application, as well as to enhance Google Maps with new functionality, including:

- **WMS Harvesting:** The interface allows harvesting of WMS Maps from any public server in the world. Given the URL of a WMS server LL-Map will collect the necessary data from the server to generate an interface for viewing the map layers.

- **WMS layering:** Map layers from multiple services can be overlaid on top of the Google base map and even on top of each other, with customizable translucency levels, allowing the user to correlate data sets from independent servers.

- **Correlating Location with Language Data:** LL-Map utilizes Google Maps markers to display locations of endangered, extinct and living languages based on Ethnologue language locations; while Ethnologue is used, the project is extensible enough to allow the use of any dataset of ISO 639-3 codes to geospatial data. The use of Google markers to display the location of languages allows the interface to display brief data about any given language within the interface itself, while providing hyperlinks to more in-depth collections of data on the language available on the web. The most notable link provided via the Google Map Interface is the LL-Map Data Browser.

While the Standard Interface is powerful in itself, it lacks several key features that the LL-Map Advanced interface offers; in particular, it lacks the ability to perform queries on the map data, and you cannot create new layers or edit existing layers.

For the Advanced Interface, we chose to work with the state-of-the-art ArcGIS, a software product produced by ESRI. This software allows the researcher to view spatial data, create maps, and perform high-level spatial analysis. The advantage we gain from ArcGIS is that it provides not only the server software for generating map images, but also desktop tools for creating maps.

In developing the Advanced Interface, the primary challenge was creating an up-to-date, standard-compliant web interface to ArcGIS. Although ESRI provides a software package called Designer, which will generate a web interface for mapping spatial data, the software was not readily adaptable to display data contained within the LL-MAP system. In particular, the interface generated by Designer uses outdated technologies, such as HTML frames (which are deprecated in the current HTML standard), and generates javascript that is not object-oriented. This interface, which uses a large number of global variables, results in an opaque interface that only works in Microsoft Internet Explorer. Since LL-Map must sustain itself in the future, every precaution was kept to ensure that the technology the system relies on conforms to the latest standards in web technologies.

At first we tried to rectify the output of the Designer program. Despite moderate success in this area, the code proved to be too complicated and hard to maintain. To create the Advanced Interface, the LL-Map team developed a new ArcGIS interface from scratch using Javascript toolkits and libraries, asynchronous requests, and the latest web standards in XHTML. An Object-Oriented approach to creating an API for the ArcGIS system was a necessary step toward maximum accessibility. The creation of Javascript objects, with constructors to represent types of map interfaces, allows for the easy integration of a GIS interface into any style web page. The use of this object-oriented design allows for proper function and variable scoping (i.e. multiple map interfaces can be created on one web page without having the chance of interference, while still having the ability for the driver program to synchronize the interfaces). As a result, we were able to create equivalent user interface components that encapsulate an ArcGIS map, just as we did with Google Maps.

The Advanced Map interface has the following advantages over the Google interface:

• **Faster and More Reliable Image Service:** The Advanced ESRI based interface uses both ArcIMS and WMS to display map images without tiling. Google uses tiling (breaking images into smaller pieces) to reduce the strain on their servers providing the images. By reducing the number of connections that the interface makes with each server involved, the chances that one request returning slowly affects the users experience is greatly reduced. Less simultaneous connections to the browser in general will always provide a speed benefit. The disadvantage to this approach is an excess use of bandwidth since the entire display must be redrawn with each movement of the map.

• **Feature Querying:** ArcIMS provides advanced querying functionality with direct access to the geocoded information in the database. A user can select a map layer to query, select a point on the map and receive information about the given layer and information about any features within the selected spatial region. The process involves an XML request to ArcIMS, which returns XML. The XML is converted into a Javascript object and sent back to the browser, which interprets the response and displays the information in an easy to read fashion. If a query returns an ISO language code, the interface can direct the user to the Data Browser to conduct further research on that language.

• **Data Input:** The Advanced Interface will provide users the ability to dynamically add new information via the map interface. A user may define a project, which contains the following hierarchy:

1. **Feature:** The lowest level of the hierarchy, the building blocks of

points, polygons and polylines. These features can be language locations, villages, country boundaries or even language features among a given population.

2. **Layer:** A layer is a collection of features; the ability to turn on and off the visibility of a layer is the lowest level of control that a user has when displaying a map.

3. **Map:** a map is a collection of layers; maps are used to organize layers into coherent models. A project may have many maps which can be useful for showing versions of data.

A modified snowflake database schema has been implemented to store feature information entered into the system, which allows for great scalability and extensibility. New features can be created with ease, and the display interface can be scaled based on the definition of the data in the database. To connect the database to the client application, we developed a Perl server to receive and respond to JSON-formatted requests from clients using REST (Representational State Transfer) URLs.

In developing the server-side code, the biggest challenge was developing an efficient and easy-to-use way of accessing the database. The data for a single object in the LL-MAP schema is stored in multiple tables, which roughly correspond to a class hierarchy. Thus accessing a single object involves accessing multiple database tables. We were able to utilize an existing Perl application framework, called Zefy, to automatically generate Perl object classes that wrap this complex schema in what is call an active record pattern, allowing the server code to deal with objects as real objects, and leaving the framework to determine which tables need to be accessed to fulfill a request.

3 Research Applications

By making accessible in one place a substantial portion of scholarly research, the LL-MAP system encourages collaboration between linguists, historians, archeologists, ethnographers and geneticists, as they explore the relationship between language and cultural adaptation and change. The wide range of information integrated into the system, allied to the researcher's ability to overlay data from different sources onto a single map, can manifest co-occurrences that may have previously passed unnoticed. From such overlays of linguistic and non-linguistic data, for example, LL-MAP partners at the University of Stockholm have discovered that Urdu speakers in the Detroit metropolitan area tended to settle in dense Arabic speaking areas rather than in dense Hindi speaking areas. Linguistically, Urdu and Hindi are closely related, and are largely mutually interpretable, but the factor that seems to play the most important role in determining spatial speaker distribution is religion, not language. Most Arabic and Urdu speakers are Muslim, while most Hindi speakers are Hindu (Veselinova and Booza, 2006).

Indeed, the most important benefit of the LL-MAP system is that it will produce "new" knowledge. In bringing together scientists and students, LL-MAP will benefit the larger scientific community. It will serve as a model not only for collaborative research projects, but also for the application of computational technologies to research in the social sciences. Overall, it fills a need for general dissemination of authoritative information about human languages, and will increase public knowledge of lesser-known languages and cultures, underlining the importance of language and linguistic diversity to cultural understanding and scientific inquiry.

The mapping (visualization) and integration capacities of a GIS system is particularly useful in organizing massive amounts of data in a flexible, context-sensitive information display. Not all data need be displayed at once, and data can be manipulated and recognized in response to user queries. The use of GIS to join physical and social science data derived from different sources is a continuing research thrust in the scientific community. This novel integration of data is expected to trigger new research questions both in linguistics and its sister fields.

References

Veselinova, Ljuba and Jason C. Booza. 2006. Using GIS to Map the Multilingual City. In Proceedings of the 26th ESRI International User Conference. San Diego, CA, August 7-11, 2006.