Generating and Reintegrating Geospatial Data

Robert F. Chavez
The Perseus Project
Eaton Hall 124
Tufts University
Medford MA 02155
E-mail: rchavez@perseus.tufts.edu

ABSTRACT
The process of building a geospatial component to access existing materials in the Perseus Digital Library has raised interesting questions about the interaction between historical and geospatial data. The traditional methods of describing geographic features' names and locations do not provide a complete solution for historical data such as that in the Perseus Digital Library. Very often data sources for a spatial database must be created from the historical materials themselves.

KEYWORDS: Geography, geospatial integration, GIS.

INTRODUCTION
The challenges of building a geographic information system that provides access to digital library materials have been highlighted by the work of the Alexandria Digital Library (ADL) and the Geospatial Knowledge Representation System (GKRS) prototype among others. Geographically referenced data permits information retrieval from a digital library by means of geospatial queries and descriptions of geographic locations.

Two common types of geospatial queries are often applied in this types of information retrieval: “what” and “where” queries. For example, a “what” query might be phrased: “What texts, images, or other objects related to location X can be found in this library?” A “where” query might be phrased: “Where in the world is location X?” The tool that links the requested information to the geographic context defined in the query is a metadata schema. Through the metadata mapping a spatial query can locate information that corresponds to, overlaps, or is contained by a spatial extent.

The ADL provides the primary exemplum for this type of digital library architecture. The ADL's geolibrary model is based on a standardized metadata schema created in part from data collected from established sets of geographic metadata sources. Information is spatially defined by mapping objects in the library's collection to this schema. In the ADL model, the primary attribute of an object is its geographic location. Classes of information about a geographic location are mapped to a footprint, a latitude and longitude value that represents "a complex polygonal boundary"[1].

The Perseus Project is currently implementing a geographic information system for the Perseus Digital Library similar in nature to that of the ADL, but which also accounts for the special circumstances and problems encountered when dealing with a digital library composed primarily of historical materials, such as literary sources, papyri, inscriptions, and image collections of ancient material culture. Geographically referencing data in the Perseus Digital Library unifies disparate classes of data and allows users to access the entire collection through geospatial query.

THE CHALLENGES OF BUILDING A SCHEMA
Geographically referencing data that interacts with the various components of the Perseus Digital Library is dependent on the development of a simple metadata schema. This initial schema is composed of a geographic placename; a geographic location represented by point, line, bounding box, or complex polygon coordinates; and where available a type designation. This schema resembles the minimum gazetteer entry components outlined by the ADL[2].

While these attributes allow Perseus to begin to make the connection between spatial data and textual geographic location data, the nature of the historical information itself has affected the way we build our schema and forced us to re-examine the way we think about spatial queries.

Dealing with a Lack of Spatial and Descriptive Data
The ability to identify geographically an object in the Perseus Digital Library depends entirely upon the availability of some kind of spatial data. If data is available it can be added to the metadata schema and the object can be geographically-referenced.

It is often the case that no spatial data source exists for ancient, medieval or Renaissance materials. Thus, the object itself may be the only available data source, or the object itself is simply a reference to a feature or location that no longer exists or has been lost in real geographical space. Whereas the ADL has available sources such as the Geographic Names Information System (GNIS), the National Imagery and Mapping Agency's (NIMA) Geographic Names Processing System (GNPS), and the

Digital Libraries, San Antonio, TX.

Getty Thesaurus of Geographic Names, these do not provide sufficient coverage for all the materials in the Perseus Digital Library. While Perseus is incorporating NIMA and Getty gazetteers into its metadata schema, the primary sources of descriptive and spatial data for our schema are actually the implicit georeferences in the existing textual database and the maps, site plans, surveys, and field reports.

The textual database is comprised of Text Encoding Initiative (TEI) conformant SGML texts, which cover a span of time stretching from the works of the ancient Greek poet Homer to Renaissance drama to city guidebooks of nineteenth-century London. These texts can function as data sources since the text editor can encode geographical, temporal, and prosopographical information. In order to place historical objects in real space Perseus has developed an authority list of geographic locations from standard sources of geographical information in its collections, such as the Princeton Encyclopedia of Classical Sites and Thucydides' history of the Peloponnesian War. The authority list enables Perseus to automate the identification of geographic locations in any Perseus text. These toponyms are then translated into geospatial coordinates for display in the Perseus atlas, the mechanism which permits the querying and visualization of spatial data.

Looking Toward Temporal Geospatial Queries

Assuming an object has sufficient spatial and descriptive data and can be geographically referenced, it is a relatively simple matter to perform a spatial query that provides access to that object or displays the location of that object in the Perseus atlas interface. But, spatial queries that ask "what" and "where" are not always sufficient to access and retrieve historical data. A spatial query of the Perseus Digital Library should also ideally have the option of adding a temporal component.. A useful query might be phrased "What objects illustrate the changes in the topography of location X from the first century BCE through the first century CE?" Many objects in the Perseus Digital Library are naturally understood in terms of time and space. For example, in our text of the hundreds of letters written by the Roman orator Cicero each letter begins with a date and a place. One retrieves a superset of spatial information when one poses the question: "Where was Cicero when he wrote his letters?" With the proper interaction between spatial and textual databases however one could pose a more precise question: "Where was Cicero when he wrote his letters in 43 BCE?" or "In his letters of 43 BCE, where did the events Cicero writes about take place?"

Objects other than texts are also understood in a temporal and spatial context as well. In cases of city level topography objects move from place to place over a span of time. For example, if one were to formulate a query with a spatial extent that covered the Roman Forum or the city of London the resulting information would include objects that may or may exist on a standard static map. Topography changes with time. In city level queries there is a need to define spatial extents in terms of dates. A more precise query might output a map of features in the Roman Forum in 14 BCE or a map of features in the city of London corresponding with the time during which the events in Bleak House take place.

Just as the texts themselves are the primary source for geographic description, they are also the primary source for the temporal information that could be applied to geospatial queries. One of our goals is to develop a temporal database based on the date information contained in the texts

IMPLEMENTING THE GEOSPATIAL COMPONENT

An initial schema covering approximately 2100 ancient geographic locations is in place in the Perseus Digital Library and new data is added as it becomes available. Spatial data, that is, geographic coordinates and extents, have been collected from various sources, primarily NIMA, and the U.S. Board of Geographical Names, or has been extracted from ESRI Shapefile format files and georeferenced site plans and is stored in a PostGreSQL database. Some data remains in the form of ESRI Shapefiles. PostGreSQL enables Perseus to maintain the spatial and textual databases in a simple clean format that is easy to maintain, portable, and easy to integrate with our mapping application, MapSever [3].

MapServer (an OpenSource CGI-based application, developed by the University of Minnesota and the Minnesota Land Management Information Center) provides the tools to build interactive web interfaces that query and visualize spatial data on the Perseus web site. Our use of OpenSource software encourages longevity of the data. MapServer's perl module (MapScript) enables Perseus to query, and process data from the RDBMS, build maps based on this data, and build customizable interfaces to deliver these maps over the Internet, such as the Perseus atlas.

CONCLUSION

The Perseus Digital Library's geographic information system serves to unify all classes of historical data contained in the library. Because of the historical nature of our data we have leveraged the materials in the digital library to create a geospatial database that now supports simple geospatial queries and will soon support geospatial queries with a temporal modifier. Since the database is built around an OpenSource architecture it is easy to maintain and distribute this data over time.

REFERENCES

1. Hill, Linda L., James Frew, and Qi Zheng. Geographic Names: The Implementation of a Gazetteer in a Georeferenced Digital Library. D-Lib Magazine, 5.1.

2. Hill, Linda L., James Frew, and Qi Zheng. Geographic Names: The Implementation of a Gazetteer in a Georeferenced Digital Library. D-Lib Magazine, 5.1.

3. Mapserver Hompage: http://mapserver.gis.umn.edu/index.html