Research Directions in Georeferenced IR based on the Alexandria Digital Library Project

Linda L. Hill
UC Santa Barbara
Alexandria Digital Library Project
Santa Barbara, California 93106
01-805-893-8587
lhill@alexandria.ucsb.edu

Michael F. Goodchild
UC Santa Barbara
Department of Geography
Santa Barbara, California 93106
01-805-893-8049
good@geog.ucsb.edu

Greg Janée
UC Santa Barbara
Alexandria Digital Library Project
Santa Barbara, California 93106
01-805-893-8453
gjanee@alexandria.ucsb.edu

Categories and Subject Descriptors
H3.3; H3.7

General Terms
Design

Keywords
Georeferencing; geospatial, gazetteers; knowledge organization systems

1. INTRODUCTION
The Alexandria Digital Library (ADL) Project began designing and implementing georeferenced digital libraries in 1994, building on the Davidson Library's Map and Imagery Lab's collection and services at the University of California, Santa Barbara. From the beginning, we have incorporated gazetteers to map between placenames and spatial footprints (longitude and latitude coordinates for the location on the surface of the Earth). We have used various database software packages to perform merged textual and spatial searching and developed a series of user interfaces and protocols for accessing distributed collections, gazetteers, and thesauri. We have an intimate, in-the-trenches acquaintance with the issues of designing and implementing effective geographical information retrieval and visualization of results in digital library environments—that is, outside of geographic information systems (GIS). In addition, we have developed a conceptual design for a test collection for the evaluation of the effectiveness of spatial searching.

2. GEOREFERENCING
Georeferencing in information resources and systems supports the gathering, understanding, and use of widely distributed and varied data, text, and image resources with relevance to a particular location on the surface of the Earth. Georeferencing methods include the formal space methods of geodesists and cartographers based on precise measurement with respect to a well-defined Earth frame [2], and the less formal place methods used every day by humans (e.g., placenames, spatial prepositions such as "near", and such classes or types as "mountains" and "schools"). Formal methods support analysis based on mathematically defined continuous spaces (e.g., calculations of distance between pairs of coordinates), while communication between humans involves reasoning about informally defined discontinuous places that often include vague references (e.g., about ten miles north of downtown; near Phoenix), vernacular names, names in multiple languages, and names for areas that are non-administrative and have fuzzy boundaries. Integrating the two georeferencing methods amounts to a new and powerful functionality for information systems. The ADL project has focused on the integration of space and place georeferencing in networked information systems.

We define information about an object (e.g., article, book, image, map, dataset, news report, manuscript, catalog record) as geographic information when it includes a georeference. In recent decades, there has been explosive growth in the use of information technologies to handle geographic information and the greater part of this growth is associated with formal space georeferences. Billions of dollars are spent every year on the collection of precise geographic information in digital form (by remote sensing from satellites and aircraft or by surveying on the ground). Geographic information systems (GIS) and related software applications allow such information to be processed, analyzed, and used to support a range of activities [4]. The challenge for digital libraries is to provide geospatial and placename access to the range of information resources, including text documents, hardcopy maps and images, and geospatial data and images through a common set of methods and interfaces.

2.1. Gazetteers
Our research into digital gazetteers has provided the key component necessary for unlocking collections of objects with informal georeferencing and making them accessible through spatial queries and understandable through visualizations of distributions on base maps [1]. Gazetteer lookup services, when widely available through the Internet, will enable georeferenced query translation (between place and space), query expansion (e.g., more specific placenames for "Southern California"), and labeling named features on maps and images. These services also underlie many areas of intelligence gathering and news analysis, where text and speech can be mined for placename references, and where these in turn can be used to integrate knowledge and to infer geographic relationships.

A gazetteer is a set of gazetteer entries defining natural and cultural features with one or more names, sets of coordinates, feature types (classes), relationships (e.g., administrative containment), time periods, and supplemental toponymic, descriptive, and documentary information. A basic gazetteer entry is formally defined as a triple <**N,t,g**>, where **N** denotes a feature name, **t** denotes a feature type, and **g** denotes coordinates for the location and extent of the feature. The content of g might be a simple coordinate pair (no information on extent), or a bounding box (generalized representation of extent), or a representation of a complex polygonal or polylinear shape (detailed representation of extent). We include in **g** longitude/latitude, and any scheme that can be transformed reversibly into longitude/latitude through an algorithmic procedure (e.g., the Universal Transverse Mercator (UTM) grid system). These three basic elements of a gazetteer entry support discovery by names, coordinates, and types (e.g., searching for schools in or near Goleta, CA).

We further generalize and formalize this gazetteer model as a type of knowledge organization system (KOS) that interrelates formal and informal representations—the formal world of precise measurement and the informal world of human reasoning and discourse. In generalizing the definition of the gazetteer entry **g** becomes a location in any space-time frame, for which we use the term space-time referencing or ST referencing for short. ST referencing also applies to named space-time events, such as hurricanes, and to named time periods, such as the Iron Age.

2.2. Geospatial representation and ranking

The specification of geospatial location and extent is currently covered by the standards developed for geographic information systems [6] which support complex and varied geospatial description in their Geometry Markup Language (GML) [5]. For more general use, we need a less complex standard that can more easily be implemented in catalogs and searching processes. We have an initial specification of a simple geometry language, developed as a profile of version 3 of the GML. Issues remain, however, such as how to represent unambiguously a bounding box (i.e., a four-sided box whose sides enclose the maximum longitude/latitude extent of the primary geometry) that extends over the ±180 meridian using the GML specification, and adopting a general treatment of the complications of projection and datum for our spherical Earth.

When a geospatial area is used as a search parameter, the search system should be able to rank the retrieval set by a spatial similarity measure that includes the degree of overlap and the relative size of the query area compared to the matching item's area of coverage. Such a ranking will put at the top items whose spatial coverage and size are most like the query area and move to the bottom of the list items whose area of overlap is small and/or whose size is much smaller (e.g., a city block) or larger (e.g., a map of the world) than the query area. We are experimenting with a set of spatial ranking methods for this purpose. A further research issue is how to integrate text-based and space-based rankings for a digital library system.

3. TESTBED FOR THE EVALUATION OF GEOSPATIAL MATCHING PERFORMANCE IN IR

For spatial retrieval testing, we have proposed the creation of a test collection that can be used to test various methods of spatial information retrieval and approaches to the generalization of the spatial footprint. For example, the assertion is made that bounding boxes perform well enough for spatial information retrieval matching operations and that the extra cost of storage and processing, and the limitations of widely deployed software, do not justify the use of polygons in digital library environments. The problem can be stated in the form of a question: How well do generalized polygons (e.g., bounding boxes) perform for spatial matching operations in information retrieval systems compared to polygonal representations that more nearly represent the shape and extent of the location? Other uses of such a test collection include the testing of various spatial similarity measures; testing of other forms of geospatial generalizations, such as various grid systems and generalized polygons; and testing database and spatial searching software.

Such a test collection would contain records representing places/locations within one geographic area which are (a) of varying levels of coverage (e.g., like the difference between countries, counties-provinces, cities, and airports); and (b) of varying shapes (e.g., circular, square, multiple polygons, diagonal, panhandle). Each record in the set would contain three spatial representations: centroid, polygon, and bounding box. Multiple levels of polygon resolution may also be included. The shape of an area is a factor in the closeness-of-fit of a corresponding bounding box. Therefore, the test collection should include locations to test the extremes of the fit of the bounding boxes, measured by the similarity of the bounding box area to the polygonal area.

The spatial similarity of each record to all other records could be calculated using the most detailed polygons available. For each record, all other records can be ranked according to a spatial similarity value indicating degree of spatial overlap and relative size. These ranked sets are the reference ranking for each record against which all other rankings are compared. Spatial similarity can be recalculated for each record based on other representations (e.g., their bounding boxes). The ranked lists based on other representations can be evaluated through comparison to the reference ranking lists, using various techniques such as correlations and precision and recall based on rank position for relevant records [3].

4. References

[1] *ADL Gazetteer Development Page.* Available: www.alexandria.ucsb.edu/gazetteer [2004, May 6].

[2] Bugayevskiy, L. M., & Snyder, J. P. *Map projections : a reference manual*. Taylor & Francis, London; Bristol, PA, 1995.

[3] Hill, L. L. *Access to Geographic Concepts in Online Bibliographic Files : Effectiveness of Current Practices and the Potential of a Graphic Interface.* Unpublished Ph.D. Dissertation, University of Pittsburgh, Pittsburgh, PA, 1990.

[4] Longley, P., Goodchild, M. F., Maguire, D., & Rhind, D. *Geographic Information Systems and Science*. Wiley, Chichester ; New York, 2001.

[5] Open GIS Consortium Inc. *Geography Markup Language (GML) Implementation Specification (version 3).* 2003. Available: http://www.opengis.org/techno/documents/02-023r4.pdf [2003, September 9].

[6] Open GIS Consortium Inc. *OGC Home Page*. Available: http://www.opengis.org/ [2004, May 6].