Issues in Georeferenced/Geospatial Digital Libraries

Greg Janée, Alexandria Digital Library Project
May 15, 2003

Georeferenced information is information that is relevant to a definable and explicitly stated subset of the Earth's surface; we call the subset the information's *spatial coverage*. Any kind of document that is about a particular geographic place (*A Tale of Two Cities*, *History of the Decline and Fall of the Roman Empire*, etc.) is potentially an example of georeferenced information. There's a large subclass of georeferenced information—and here we're thinking of maps, remote-sensing imagery, and the like—that is distributed over the extent of its spatial coverage and that is typically graphically visualized. We distinguish this subclass as *geospatial information* to emphasize its spatial characteristics.

This discussion paper presents some of the issues that arise in storing georeferenced and geospatial information in a digital library, and more importantly, making such information discoverable by and usable to a broad spectrum of library users.

Discovery. The first and most basic issue is: how can georeferenced information be discovered? The traditional approach to georeferenced discovery is to use text-based information retrieval techniques on the metadata associated with the information, and in particular, to base discovery on matching of textual placenames. Using this technique, a user desiring a map of Boulder, Colorado, would use the two underlined words as a query, and the search service would return items whose metadata contained these two words or phrase.

This technique works well for certain classes of information and in certain contexts. Clearly, such a technique would return an item whose title is "Downtown street map of Boulder, Colorado." But as a general technique it suffers from two serious drawbacks. First, there is a whole class of georeferenced information that has *no* associated placenames, namely, data gathered from moving sensors. Examples include satellite imagery and aerial photography, which carry only technical metadata such as the coordinates and attitude of the camera at the time of exposure.

Second, discovery based on textual placename matching is unreliable in the large. Consider for a moment a user desiring a map of the Flatirons rock formations in the Front Range of the Rocky Mountains just outside of Boulder, Arapahoe County, Colorado. The underlined words in the previous sentence all describe the desired area to varying degrees of specificity, but it turns out that the most specific phrase (and pragmatically speaking, the only *useful* phrase) that will retrieve data from the U.S. Geological Survey for this area is Eldorado Springs (!), which happens to be the name of the "quadrant" that covers the Flatirons per the USGS's standardized grid system. The USGS picks one prominent feature within each quadrant to serve as the name for that quadrant. If the USGS had included Flatirons in its cataloging our hypothetical problem would be solved, but the USGS can't be faulted, for there is no easy or manageable way to associate every possible placename with every library item. This is but one instance of a very general

problem, namely, there are *many* names for any given place (administrative names, physiographic names, technical names such as telephone area codes and postal codes, event names such as named earthquakes and hurricanes, etc.), and a document that is about a place can be cataloged by, at best, a handful of those names.

To address the general problem of reliable georeferenced discovery, some kind of structured (non-textual) search technique must be employed. Defining a controlled vocabulary of places (so that, for example, "Boulder, Colorado" would be treated not as a text string but as one of a set of discrete terms) is one structured approach often employed by libraries. This approach resolves the placename multiplicity problem to a certain extent, and an advanced interface to the vocabulary can go a long way toward guiding the user in the selection of the proper term, but it also places a considerable burden on both library catalogers and users to understand, agree upon, and use the vocabulary. Gaining agreement on a vocabulary is a classic problem; it requires designating one among many possible names as the preferred name for a place, which is similar to the USGS choice of one placename to represent a quadrant, but it still leaves the problems of anticipating all of the other names that users will associate with the place and adding them as alternative or related names.

A more general and powerful technique is to support range searching (i.e., searching that employs inequality operators such as BETWEEN and OVERLAPS) over, say, numeric latitude/longitude coordinates. This technique places a large burden on libraries and users to express spatial coverages and query regions using coordinates. But the advantage is that it allows a user to discover information without the user and library having to agree on anything except the coordinate system. Indeed, coordinate-based georeferenced discovery represents the state of the art.

Gazetteer integration. The above discussion should not be construed as implying that placenames are unimportant or not useful in georeferenced discovery. On the contrary, placenames play a critical role in georeferenced discovery, and in georeferenced digital libraries in general, because human spatial cognition relies on relationships to and among known features, and to the extent that those features are named, to placenames. (Proof: think about how you would answer the question, "Where are you right now?") Consequently, it is necessary that a *gazetteer* (a kind of dictionary that supports translation between placenames and coordinates) be integrated into a georeferenced digital library to help translate between the user's mental model of geographical space and the library's discovery system.

Ideally, one might like to integrate a gazetteer into a georeferenced digital library in a deep and significant way, so that the translation from user queries to the library's discovery system is hidden and seamless: the user enters Flatirons and the library automagically performs an appropriate coordinate-based search and returns data for quadrant Eldorado Springs. Unfortunately, at the time of this writing, publicly available gazetteer data is too incomplete, inaccurate, and imprecise to support this level of functionality. Gazetteer data is often gathered from existing maps, meaning that administrative and geographic features are often represented only by points, even when

the features have significant extents (a single point location for an entire county, for example). Furthermore, the point locations may reflect the locations of the map labels for the features, not the features themselves.

But while seamless gazetteer integration may remain elusive, *human-mediated* gazetteer integration is possible and, moreover, effective and easily achieved. It simply requires that gazetteer lookup services be treated as an extension of the library's user interface. In this model, the user enters Flatirons, interacts with a gazetteer and background map to locate the desired place (in case there are multiple Flatirons, as there almost always are), and from there forms an appropriate coordinate-based spatial query region to submit to the library.

Heterogeneity. In the world of georeferenced and geospatial information there are many metadata formats and content standards in use, and they all describe the spatial coverages of georeferenced information slightly differently. There are multiple geometry languages, coordinate systems, and georeferencing techniques in common use. A digital library that accommodates this heterogeneity will have to provide mapping mechanisms at some level.

Data typing. Digital libraries are typically text-based in that they treat the metadata they operate on as either undifferentiated text or text that is subdivided into named fields. Furthermore, the discovery and ranking functionality digital libraries provide is typically based on finding and counting occurrences of words in metadata text. Adding a more structured discovery technique to a digital library (specifically, range searching over geographic coordinates) adds a number of complications. If we think of text as being one type of data, adding georeferenced discovery means adding both a second data type and the more general notion that there are now multiple data types to be distinguished and handled. Some of the ramifications of this:

- Input validation is required: geographic metadata (descriptions of the spatial coverages of library items, for example) cannot be treated as undifferentiated text.
- Relatively complex internal structures and external representations are required to describe spatial coverages and geographic query regions.
- The library must provide the means to express, and of course execute, different types of query constraints (spatial constraints, textual constraints, etc.) as well as boolean combinations of different types of constraints.
- Type-specific ranking methods are required, as well as methods for combining different types of rankings. Georeferenced discovery requires a specialized type of ranking to avoid the common problem of the World Atlas (with its entire-world spatial coverage) being returned as the first answer to any and every query. One technique is to rank library items based on the "spatial similarity" (a function of the overlap in size and location) between the items' spatial coverages and the user's query region. Given the state of Colorado as a query region, such a ranking would return state-size items first, and would return very large (the World Atlas) and very small (city-size) items last. The point here is that such a ranking scheme is quite

different in nature from classical information retrieval rankings and requires special consideration.
- Specialized user interface components are required to input and to view typed data; in the case of geographic coordinates, this means interactive map browsers, which in turn require underlying map servers.

Scalability. It is easy to accumulate geospatial data since it is often generated by automatic means (satellites, sensors, etc.), and thus scalability in terms of accommodating large numbers of library items can become an issue.

Georeferenced discovery techniques are nicely scalable in theory (the two-dimensional R-tree-based index structures in common use today offer logarithmic search time), but in practice, scalability is more limited. At the time of this writing, commonly available, commercial georeferenced search engines reach the limits of practical use (say, one day of compute time and several gigabytes of RAM) to index as few as 10^6 items. Larger numbers of items can be accommodated, of course, but only with exponentially increasing amounts of expense and custom engineering.

A more significant problem is the difficulty of *joining* (in the database sense) boolean combinations of spatial constraints with textual and other types of constraints. A typical and reasonable gazetteer query such as "find a city named Boulder near the Rocky Mountains" is a conjunction of a discrete constraint (find a city, where "city" is one of a set of discrete types), a textual constraint (named Boulder), and a spatial constraint (near the Rocky Mountains). When large numbers of items are involved, generating efficient query plans becomes paramount, but query plans are unavoidably sensitive to the queries and to the distribution of the data, and the data is inevitably far from uniformly distributed. Whether the library is working within the framework of a single relational database or across multiple, distributed search indexes, the heterogeneous join problem is a significant practical problem and appears to still be a research problem as well.

Spatial context. In any information space, users need context to understand where they "are" in that space and what the information is "about" in order to formulate queries and interpret query results.

In the case of a georeferenced digital library, the geographic context is especially critical because, as noted previously, humans reason about space symbolically and rely heavily on relationships to known features. Consequently, users typically want to express georeferenced queries either symbolically ("Boulder") or by relationship to known features, say, by indicating a region of interest on a context-providing background map. The implication is that georeferenced digital libraries must provide interactive background maps that are integrated into the library's user interface and that have sufficient detail to allow users to relate their location to the landmarks they recognize. As counter-proof of this, try using a library or other system that requires input of numeric geographic coordinates—without relying on any external, context-providing device!

Context is also critical in interpreting and evaluating query result sets and individual library items. Suppose for a moment that we have gotten past the first spatial query that everybody does ("let's see if I can find my house"), an exercise that is highly misleading because we are intimately familiar with our own surroundings. Instead, suppose that the aerial photograph to the right has been returned as a response to the previous query about the Flatirons rock formations near Boulder, Colorado, an area we will suppose we are unfamiliar with. Does this give us the information we were after? More to the point, where *are* the Flatirons in this image? Clearly we could answer this question by importing the photograph into a geographic information system (GIS), layering map and placename data over it, but that assumes a high level of sophistication and resources on the part of the user. To be useful to the general user, a georeferenced digital library must provide a certain amount of its own (lightweight) GIS functionality to be able to show this image positioned over a reference map, perhaps, or labeled with the features within the image.

Content access and integration. In this last section we consider the implications on digital libraries of specifically geospatial information.

Geospatial information has complex structure and is often quite large. It can consist of multiple parts, require specialized viewing tools, and be accessible via multiple formats, protocols, and interfaces. For example, geospatial imagery is often distributed as a pair of files: an image proper (say, a TIFF file) together with a file containing georeferencing information. A client may have to independently access one or both of these files to use the information successfully. To take another example, the industry-standard ESRI "shapefile" format is in fact a set of 4–7 coordinated files of different types. And geospatial information is also often accessible via programmatic services such as the OpenGIS Web Map Server (WMS) protocol or ESRI's ArcIMS interface. Access via service becomes more critical as the size of the geospatial information increases because service access typically allows navigation to and interaction with just a desired subset of the item.

Geospatial information is also closely tied to geographic information systems and other types of data exploration environments. In some cases GIS systems are simply a common means of viewing and working the information; in other cases, the information is entirely unusable outside the appropriate analysis or visualization package. Either way, the ability to meaningfully process geospatial information strictly within the confines of a standard Web browser is limited.

These characteristics of geospatial information—its complex structure and its close ties to analysis environments—call out the need for geospatial digital libraries to be able to

describe, in detail and for the benefit of both programmatic clients and human users, the components that make up a library item and the different modes by which it may be accessed. It is insufficient to describe the content of a geospatial library item by, say, a single, unadorned URL ("click here"). A more detailed description can allow the user to make an informed decision about the best way to access the item, and can allow the library to seamlessly hand the item off to a more capable analysis or visualization package.

The author would like to thank Linda L. Hill for her helpful comments in preparing this paper.