Spatial Search, Ranking, and Interoperability

Greg Janée
Institute for Computational Earth System Science
University of California
Santa Barbara, CA 93106-3060
+1 805 893 8453
gjanee@alexandria.ucsb.edu

James Frew
Bren School of Environmental Science & Management
University of California
Santa Barbara, CA 93106-5131
+1 805 893 7356
frew@bren.ucsb.edu

Categories and Subject Descriptors

H.3.7 [**Digital Libraries**]: Standards, Systems issues; H.3.4 [**Systems and Software**]: Distributed systems; D.2.12 [**Interoperability**]: Data mapping.

General Terms

Design, Standardization.

Keywords

Georeferenced, geospatial, distributed, design, interoperability, search, ranking.

1. INTRODUCTION

In any distributed software system that strives to provide interoperability across independently-developed components, the intercomponent interface is a key architectural feature. The design of this interface is typically a compromise between opposing forces. On the one hand, data *providers* find it easiest to disseminate data in representations closely matching their internal, implementation-specific representations, and likewise service providers find it easiest to provide functionality closely matching their underlying implementation functionality. On the other hand, data *consumers* find it easiest if providers are alike in terms of representations and services offered.

Reconciliation of this tension can occur through careful abstraction of implementation differences. Very often, flexibility is added to the interface in the form of capabilities assertions by data and service providers; at the extreme, capabilities assertions grow into full negotiation of capabilities between consumer and provider. However, capabilities assertions and negotiation place large burdens on both providers and (especially) consumers to handle many possibilities that may or may not be encountered in actual use. As a result, what happens *in practice*, either by fiat or simply *de facto*, is that a commonly agreed-upon, base-level interface emerges that allows consumers to make assumptions about the representations and services that will be universally provided, and that gives providers clear requirements as to what must be supported to effectively participate in the system. This base level is apparent in most protocols in use today and is the focus of this paper.

The Alexandria Digital Library (ADL) project has been working to develop lightweight, distributed digital libraries for heterogeneous georeferenced data. *Distributed* means that a library's components may be spread across the Internet, as well as coexisting on a single desktop. *Heterogeneous* means that a library may contain multiple types of digital data, including remotely-sensed imagery, textual documents, executable models, and multimedia instructional materials. *Georeferenced* means that, whenever possible, each item in a library is associated with one or more regions on the Earth's surface. (We refer to the union of these regions as the item's *spatial footprint*.) *Lightweight* means that the burden on library implementations is minimized to allow groups and systems that would not ordinarily be thought of as spatial data providers (traditional library catalogs, for example), nor ordinarily capable of being spatial data providers (small digital library implementations lacking spatial engines, for example) to, in fact, participate in a spatial system. ADL can thus be characterized less as a source of spatial data, and more as a system that provides a spatial orientation to heterogeneous data sources.

Many components make up a distributed digital library like ADL: library clients and library and collection servers, primarily, but query mediation, gazetteer, and map services as well. Given ADL's focus on georeferenced data, a significant aspect of the interfaces between these components is *spatial interoperability*: the ability to communicate geographic regions (specifically, query regions and the spatial footprints of collection items) and to invoke spatial operations (e.g., searching library collections for items having footprints that overlap a given query region). There are two key choices in the design of such spatial interoperability interfaces:

- *Allowable geographic representations*: Supporting many kinds of shapes, especially complex shapes (polygons with holes; collections of regions treated as first-order regions), gives data providers considerable flexibility in describing geographic regions, but places a significant burden on consumers of the representations, which must be coded to handle every possible kind of shape.
- *Allowable spatial query predicates* (OVERLAPS, WITHIN, etc.): Supporting many predicates gives consumers great power in expressing queries and performing other spatial operations, but places a significant burden on service providers. Also, some predicates, such as TOUCHES, are particularly sensitive to the precision and accuracy of footprint specifications, and thus place indirect requirements on data quality.

Given ADL's goal of developing lightweight digital libraries, we are especially interested in identifying a minimal or base level of

Conference'04, Month 1–2, 2004, City, State, Country.

spatial interoperability that supports what we believe are the core georeferenced digital library functionalities: representing the spatial relevances of collection items, and querying for collection items by relevance to a spatially-defined region.

2. SPATIAL INTEROPERABILITY: APPROACHES

There are several existing standards that support spatial interoperability; and they have generally taken a more "heavyweight" approach than ADL has. The GEO application profile of Z39.50 [1] is closest to ADL's approach. It defines representational attributes based on the FGDC Content Standard for Digital Geospatial Metadata [2], and thus includes such complex geographic representations as polygons with multiple holes. The only mandatory search predicate defined by GEO is OVERLAPS testing on bounding boxes, though constraints against individual bounding box coordinates may also be specified – a form of constraint that, paradoxically, is supportable by the simplest implementations but not by many implementations using sophisticated spatial engines.

The Open GIS Consortium's specifications provide far richer representations and services. The OpenGIS Simple Features Specification [3] defines nine representations, including such complex shapes as polygons with multiple holes and arbitrary geometry collections, and it includes all five predicates (DISJOINT/INTERSECTS, TOUCHES, CROSSES, CONTAINS/WITHIN, and OVERLAPS) based on the Dimensionally Extended Nine-Intersection Model (DE-9IM) [4]. The OpenGIS Web Feature Service [5] uses the same DE-9IM-based suite of predicates, but its representations are based on the Geography Markup Language [6], an extraordinarily complex specification that defines hundreds of XML elements describing roughly twenty possible geometric shapes.

In contrast to these other standards, ADL has taken a deliberately lightweight approach to spatial interoperability. Spatial footprints of collection items may be described as boxes (i.e., as latitude/longitude-aligned rectangles defined on the sphere; a more detailed definition is given below) or as simple (i.e., non-intersecting and hole-free) polygons or polylines [7]. Query regions may be specified as boxes or polygons [8]. Polygons and polylines must be accompanied by their bounding boxes, to support components that can only operate on boxes. Spatial predicates are limited to INTERSECTS, WITHIN, and CONTAINS.

This more minimal approach supports the basic operations of a georeferenced digital library. The allowable representations can describe collection items' spatial footprints with sufficient fidelity to yield recognizable shapes for both areal (counties, states, etc.) and linear (rivers, streets, etc.) features. The mandatory inclusion of bounding boxes with the more complex kinds of shapes supports a notion of *spatial fallback*, in which less-capable components can fully and automatically participate in the system, albeit at lower spatial fidelity. The predicate most often used is INTERSECTS, but the WITHIN predicate is useful in excluding irrelevant data and the CONTAINS predicate is useful in formulating coverage queries.

3. SPATIAL INTEROPERABILITY: CHALLENGES

We find it very interesting, however, that even ADL's lightweight approach to spatial interoperability has proven to be surprisingly difficult to implement. We detail three reasons why.

Polygons and other complex shapes add substantial complexity to component implementations. Spatial predicates involving polygons (polygon-polygon intersection, point-in-polygon testing, etc.), or other complex shapes such as curves and geometry collections, require specialized algorithms typically found only within sophisticated spatial engines. Polygons have impacts on other components as well. For example, rendering polygons (as opposed to, say, boxes only) in graphical user interfaces requires more complex programmatic interfaces capable of accepting possibly voluminous polygon descriptions.

Spatial interoperability must be based on a geodetically continuous topology. Topologically speaking, the Earth's surface is a 2-sphere manifold, which is to say that it is everywhere locally homeomorphic to a plane, but globally, it is a sphere. Planar algorithms are vastly simpler and more plentiful than spherical algorithms, and many data providers work within only limited regions of the Earth's surface. Therefore, both providers and users of geographic information are able, and find it convenient, to work within various kinds of planar cartographic projections, even though use of a planar projection necessarily introduces one or more discontinuities in the representation of the Earth's surface. But a system trying to provide spatial interoperability over *all* data providers must recognize and accommodate the continuous sphericity of the Earth's surface. We call this requirement *geodetic continuity*. Failure to support geodetic continuity can make it impossible, or at least burdensome, to describe and operate on geographic regions that cross a discontinuity.

One approach to geodetic continuity is to give data providers freedom to use any locally-appropriate planar projection, so long as the particulars of the projection are declared via some formal mechanism. This introduces three significant complications:

1. The non-trivial task of re-projecting data into a common query space is simply "punted" to the data user.
2. Geographic regions represented solely by their vertices cannot be trivially re-projected, since the implied lines connecting the vertices may project into complex curves.
3. Re-projection can cause simple regions to become geometry collections, such as when the region crosses a discontinuity in the destination projection.

Another approach to geodetic continuity is to define spatial interoperability in terms of spherical regions. Here the problems are in dealing with the complexity of, and generally poor support for, spherical regions. Spherical polygons (i.e., polygons whose edges are great circle arcs) are difficult to work with. Even computing the bounding box of a general spherical polygon is quite difficult, and as a result few contemporary spatial engines are capable of operating on spherical topologies.

ADL's solution to providing geodetic continuity is to base spatial interoperability on only the simplest spherical regions: geodetic boxes. A *geodetic box* is a region of the sphere delimited by northernmost and southernmost latitudes and easternmost and westernmost longitudes. It is thus equivalent to the familiar axis-aligned box in many planar projections, though it should be noted that identifying the explicit eastern and western boundaries, as opposed to minimum and maximum coordinate values (as is the practice in several other standards), avoids discontinuities at the anti-prime meridian. Geodetic bounding boxes can be easily computed from polygons in most cylindrical and polar projections, and support for ADL's spatial predicates is manageable (though, admittedly, non-trivial) by implementations capable of performing operations in planar topologies only.

Spatial predicates other than INTERSECTS *are difficult to support*. While it is not surprising that predicates such as TOUCHES are difficult to support, predicates such as CONTAINS and WITHIN are difficult, too. We mentioned earlier that re-projection of geometric shapes can trigger the creation of geometry collections. Another source of geometry collections are footprints of library items that are relevant to multiple, spatially discrete phenomena such as hurricanes or earthquakes. Few spatial engines support spatial predicates on geometry collections. Other spatial engines allow multiple, independent geometries to be associated with an entity. These latter implementations can still support the INTERSECTS predicate; can support the WITHIN predicate with some judicious query-rewriting; but still cannot support the CONTAINS predicate.

4. CONCLUSION

We conclude that providing spatial interoperability that is both geodetically continuous (i.e., is defined in and operates within the topology of a 2-sphere manifold) and lightweight (i.e., is easily supportable by data and service providers) requires two architectural design characteristics, neither of which is sufficient by itself. First, the set of allowable geographic representations must be restricted to just the most basic representation, namely, geodetic boxes. Second, the set of allowable spatial predicates must be limited to only the most basic predicate, INTERSECTS.

However, reliance on only the INTERSECTS predicate brings up the need for spatial ranking. A collection item with whole-world spatial extent, such as a world atlas, answers every INTERSECTS query, since its footprint intersects every possible query region. Without ranking of query results, such an item may well be the *first* answer to every query. While such behavior is technically correct, in practice it is decidedly user-unfriendly.

ADL has been researching one particular notion of ranking query results based on the *spatial similarity* of their footprints to the query region. For this purpose we define the spatial similarity of two geographic regions to be the Hausdorff distance between the regions, i.e., a function of the regions' sizes, shapes, and locations [9]. For example, given the state of California as a query region, this definition of spatial similarity would rank highest those resources that approximately match California as a whole (e.g., a map of the state of California, a California statewide dataset). A resource such as a map of the western United States would rank lower, and very large (e.g., a world atlas) and very small (a city street map) resources would rank lowest. Other types of rankings, such as those based on items' absolute sizes or minimum feature resolution, also merit investigation.

5. REFERENCES

[1] Douglas D. Nebert. Z39.50 Application Profile for Geospatial Metadata or "GEO." Version 2.2 (May 27, 2000). http://www.blueangeltech.com/standards/GeoProfile/geo22.htm.

[2] Federal Geographic Data Committee. FGDC-STD-001-1998. Content Standard for Digital Geospatial Metadata. 1998. http://www.fgdc.gov/metadata/contstan.html.

[3] Open GIS Consortium, Inc. OGC 99-049. OpenGIS Simple Features Specification for SQL. Revision 1.1 (May 5, 1999). http://www.opengis.org/docs/99-049.pdf.

[4] Eliseo Clementini and P. Di Felice. A Model for Representing Topological Relationships Between Complex Geometric Features in Spatial Databases. Information Sciences **90**(1-4) (1996): 121-136.

[5] Open GIS Consortium, Inc. OGC 02-058. Web Feature Service Implementation Specification. Version 1.0.0 (September 19, 2002). http://www.opengis.org/docs/02-058.pdf.

[6] Open GIS Consortium, Inc. OGC 02-023r4. OpenGIS Geography Markup Language (GML) Implementation Specification. Version 3.00 (January 29, 2003). http://www.opengis.org/docs/02-023r4.pdf.

[7] Greg Janée. XML document type definition for ADL bucket reports. http://www.alexandria.ucsb.edu/middleware/dtds/ADL-bucket-report.dtd.

[8] Greg Janée. XML document type definition for the ADL query language. http://www.alexandria.ucsb.edu/middleware/dtds/ADL-query.dtd.

[9] Greg Janée. Spatial Similarity Functions. 2003. http://www.alexandria.ucsb.edu/~gjanee/archive/2003/similarity.html.