Finders and Keepers In Search of Spatial Data

Jonathan W. Lowe

This column covers the role of emerging technologies in the exchange of spatial information.

In North Carolina, when the legislature directed the state mapping agencies to produce new flood maps, they took the opportunity one step further. They will be offering the entire state's very high resolution elevation, flood zones, feature data and orthophotography to the public over the Web for free. Some 17 to 20 TB of data will be freely downloadable at www.ncfloodmaps.com.

North Carolina's flood mapping example is one of many such data explosions. In British Columbia, for instance, the Ministry of Sustainable Resource Management's BMGS branch manages a library containing approximately 2,000,000 aerial photos to which they add 50,000 new shots every year.

Among the challenges of managing a large data collection is communicating its existence to potential users. As their holdings grow, data providers (both public and private) need metadata publishing techniques so their users can easily become aware of new offerings. On the receiving end, users need effective searching tools that return appropriate results.

The ever-growing volume of both spatial data and its related metadata are changing the way our industry describes, archives, and searches for information. This column examines data suppliers' techniques for broadcasting their assets, and users' techniques for finding appropriate spatial data.

Net Results columnist Jonathan W. Lowe is the owner of Local Knowledge Consulting (Berkeley, California), where he designs and implements spatial Web sites. Lowe can be contacted at info@giswebsite.com.

Filtering the search engine

BMGS envisions a system that unites eager seekers with appropriate data automatically. Sifting through a collection of two million air photos is easy enough (for a fast computer), but this collection is just one of many such spatial data stores. No search engine is powerful enough to scan all the Web's data collections and return only appropriate results in a reasonable period of time. How do worldwide search engines decide which collections are worth sifting in the first place?

If BMGS follows the FGDC's (www.fgdc.gov) model, their solution will include standard metadata, and clearinghouse nodes. Alternatively, emerging metadata publication models driven by Web services may supplant the FGDC model.

Standard metadata. At the heart of any metadata standard is simply a way to answer "who, what, when, where, why, how?" questions, called elements, about any data. One of the most spartan metadata standards, Dublin Core (dublincore.org), contains only 15 required metadata elements, while the FGDC has 334. Consider Dublin Core's single geographic element, "Coverage":

"The extent or scope of the content of the resource [which] will typically include spatial location (a place name or geographic coordinates), temporal period (a period label, date, or date range) or jurisdiction (such as a named administrative entity). Recommended best practice is to select a value from a controlled vocabulary (for example, the Thesaurus of Geographic Names [TGN]) and that, where appropriate, named places or time periods be used in preference to numeric identifiers such as sets of coordinates or date ranges."

The nice thing about a standard like Dublin Core is its low cost of entry; it only takes a moment to populate 15 metadata elements. Even sparse metadata are better than no metadata at all.

Glossary

ASP: Active Server Pages

BMGS: Base Mapping and Geomatic Services

FAQ: Frequently asked questions

FGDC: Federal Geographic Data Committee

GILS: Government Information Locator Service

HTML: Hypertext markup language

ISO: International Organization for Standardization

MARC: Machine-readable cataloging

NCSU: North Carolina State University

OAI-MHP: Open Archives Initiative Metadata Harvesting Protocol

RFI: Request for information

SGML: Standard general markup language

SMMS: Spatial Metadata Management System

XML: Extensible markup language

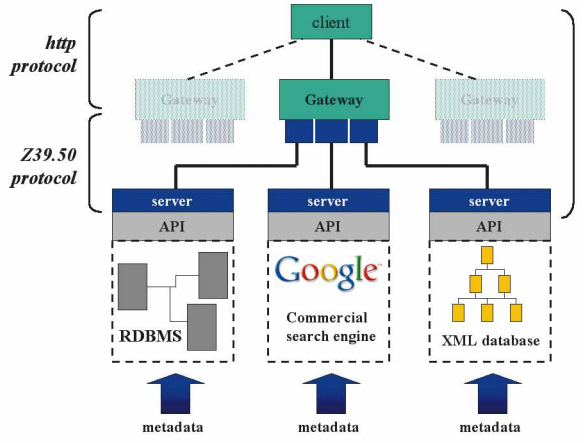


FIGURE 1 This generalized clearinghouse or "catalog" architecture shows how several discrete nodes can participate in an interoperable metadata search. The Web browser submits a search request to a proxy server called a Gateway, which in turn searches the records of multiple metadata servers in parallel using the Z39.50 protocol.

In contrast, the FGDC metadata has individual elements for coordinates, place name, date, extent, and many other specific spatial descriptors. Though labor-intensive to populate, the end result goes a long way. Computer programs can later cross-reference a subset of rich metadata content like the FGDC's to another, less rigorous metadata standard such as MARC (a library standard), thereby exposing the content to a different community of searchers. Different formats can represent the same metadata content. The USGS (www.usgs.gov), for example, offers a tool called *mp* (metadata parser, at http://geology.usgs.gov/tools/metadata/tools/doc/mp.html) that disassembles FGDC metadata to recognize its components, then checks its syntactical structure and outputs it to other formats such as XML, SGML, HTML, or even FAQ-style HTML, for use with a variety of applications.

While flexible formating is an asset, the FGDC's flexible content may not be. Metadata authors use descriptive text rather than numeric codes to populate FGDC metadata elements. This later makes searching more difficult, since one person's stream is another's creek, for example. The emerging ISO19115 metadata standard addresses this problem by recommending not only metadata elements, but also valid content — code 100 representing creeks or streams, for instance — enabling standardized searches.

Harvesting the clearinghouses. In the mid-1990s, after agreeing to spatial metadata form and content standards, the FGDC, borrowing from the library community and others, designed a system for searching against the metadata. They rejected a central warehouse setup in favor of distributed clearinghouse nodes so that each data provider could keep their own metadata current. Using a protocol called Z39.50, each node exposes its metadata database to the geospatial community using an industry-accepted profile called GEO. Any software capable of recognizing this profile can search every node's database through a single interface (**see Figure 1**).

The idea caught on. Today, there are approximately 250 clearinghouse nodes serving a wealth of metadata. Even though each clearinghouse stands alone, software such as Blue Angle Technologies' (www.blueangeltech.com) Metastar can search against multiple clearinghouses as if they were a single warehouse thanks to the common protocol. According to some in the library community, however, querying every node may not be the most effective approach.

Imagine searching for maps related to the California Gold Rush. Others, often librarians, may already have special collections on this specific topic. Searching only within their metadata subset, it's safe to use a keyword like mine to find maps for gold mines without also getting maps of land mines. Such a focused metadata collection might be built by selectively downloading topical data to a special collection site, or, more efficiently, by harvesting only the metadata from its original sources. Another protocol, the OAI-MHP (www.openarchives.org), supports harvesting of Dublin Core metadata and is easier to implement than the Z39.50 protocol.

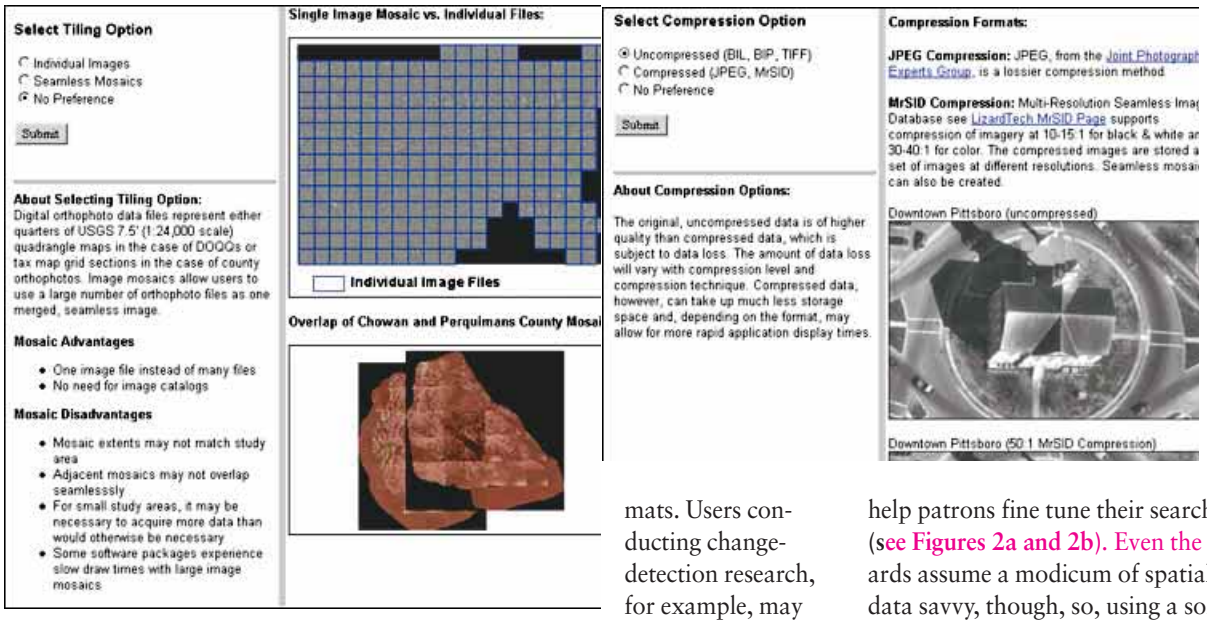
The danger with such harvested metadata collections, though, is that their content can get stale. For example, a date-dependent filter might miss current 2002 data if its stale metadata claims its last update was in 1999. Keeping metadata subsets fresh with frequent reharvests, however, can put too much pressure on the source servers.

Even though spatial metadata harvesting technology emerged only recently, harvested metadata collections' speedier searches may appeal to organizations with multiple management tiers. For instance, a state government could harvest metadata from its county and local governments, then offer the whole collection through a single search point.

Virtual stacks

The technical approaches to metadata, clearinghouses, and harvesting impact the end users, though we may not always realize it. The more metadata there are, the harder it is to craft a search that will return just a few useful results. The increased sophistication in search strategies is most evident at libraries, the longtime centers for search and retrieval. As has always been the case, managers of library collections devote considerable attention to acquiring, cataloging and storing data, and then building flexible interfaces for finding it again later.

Steve Morris, who directs Digital Library Initiatives, NCSU Libraries (www.lib.ncsu.edu/stacks/gis), says library geospatial information services have changed significantly over the past decade. Libraries have moved "from [paper] map collections to [dig-



FIGURES 2A AND 2B NCSU librarian, Steve Morris, created online wizards to refine his patrons' spatial searches. Tiling and compression are just two of the criteria that narrow a search for orthophotography, improving the likelihood of appropriate matches.

ital] data collections to provider of map services, with the library increasingly becoming a portal to other map services. . . " rather than the steward of actual hard copy or digital data archives. The move toward portal function is a response to the flood of new data. Finding room to archive each North Carolina county's 20 to 120 GB of orthophotography, for instance, is a daunting challenge. With data volumes growing ever larger, Morris admits that, "It's attractive to use services because they eliminate the hassle of storing the growing volume of data." And 49 out of 100 North Carolina counties already distribute spatial data using Web services.

Oldies but not goodies. But although Web services reduce the incentive for the library to actually acquire the materials, substituting a link for the actual digital file it represents raises concerns about longer term access and preservation. Counties or orthophotography vendors don't always preserve the old imagery in an easily retrievable place and there is no guarantee storage media will be refreshed or data migrated to newer file formats. Users conducting change-detection research, for example, may not be able to find those important old images as easily as the more recent ones.

Hide and seek. Higher volumes of available data are also changing the degree of sophistication required for a successful search, the first step in many spatial projects. The typical search returns either too many inappropriate results or none at all. For NCSU's library patrons, though searching is not so frustrating.

In person, Morris interviews patrons about their spatial data needs and helps them craft a sophisticated search filter. In most searches, the biggest challenge is knowing how to phrase the question to create a filter that sifts through the voluminous metadata, eliminating all but the truly relevant matches. For instance, the more tightly a user can specify the desired data format, delivery methods, map projection, tiling strategy, project scale, and other criteria, the more appropriate the results will be.

When NCSU's digital catalog became available online, however, Morris discovered that library patrons were conducting searches around the clock, at hours when no librarian could guide them. His solution was to create online search wizards — a series of educational Web pages to help patrons fine tune their searches (see Figures 2a and 2b). Even the wizards assume a modicum of spatial data savvy, though, so, using a software package by QARBON (www.qarbon.com), Morris also offers his patrons introductory online animations (called viewlets) that explain the basics of spatial data's origins and use.

Tipping the scale. According to Morris, the most useful (but often missing) element of metadata for a spatial data search is one he calls "appropriate scale for use." A user studying individual trees for an urban forestry project, for example, wouldn't want polygons delimiting nationwide ecoregions, even if some of those regions overlapped her city's study area. Similarly, a user working at the scale of county demographics may not want census block data.

The same scale problem arises when the filter is a bounding box. Even though the box encloses your own little neighborhood, the result set usually includes overlapping data with worldwide coverage, created at a scale inappropriate for your neighborhood study. Scale-based limitations could solve the search box problem, but digitally born products—not being created from an analog product to which a scale may be easily assigned—typically are not provided with a scale statement in the metadata (for example, a remotely sensed or GPS-captured data). And even if statements about scale are part of meta-

data's lineage section, what if the data are an aggregate of multiple products, each of a different scale?

Scale is such a useful filter, though, that in NCSU's system, Morris assigns an implied scale based on what he knows about the data (but doesn't share his guess with the user). The search logic assumes that larger scale is better, since NCSU researchers are often working on land grant–based local projects. According to Morris, one possible catalog application would involve harvesting metadata from clearinghouse nodes, selectively pulling back metadata based on bounding coordinates and making assumptions about appropriateness of scale based on the data extent. Global extent usually indicates a small scale of data detail; local extent suggests large scale.

Browsing. For fine-grained searches, the NCSU search tools include the ability to browse. Librarians use the terms precision and recall to distinguish, say, a search for "rivers" from a search for "rivers and streams and water and. . . " In the NCSU thesaurus browse system users may begin with a precision search to identify potentially fruitful nodes, followed by manual browsing of the sublist for broader, narrower, related, or substituted terms (that is, recall). Thesaurus systems accommodates data seekers who start with too broad or too narrow a term, or the wrong synonym. Without thesaurus browsing, for instance, the precision filter of "rivers" might ignore data described with substitute terms such as "streams," narrower terms such as "scenic rivers", broader terms such as "surface hydrography," or even related terms such as "hydrologic units" or "navigable waterways."

Where oh where?

Even if you don't plan on opening your own library anytime soon, these metadata broadcasting and searching techniques may be part of your organization sooner than you think; spatial vendors such as ESRI (www.esri.com) and Intergraph already include metadata publication and harvesting capabilities into their product suites. For example, the State of Kentucky uses ESRI's ArcIMS 4.0 to publish metadata from ArcCatalog with a geography-network-lookalike search interface (kygeonet.state.ky.us/). Intergraph's (www.intergraph.com) SMMS allows users to create and edit FGDC-compliant metadata from the GeoMedia environment and to publish it using an Oracle or SQL Server database, GeoConnect 1.01, and an ASP web interface. For example, Intergraph's SMMS drives the State of New Jersey's spatial search pages (njgeodata.state.nj.us). There's an awful lot of data out there; may all these approaches succeed in uniting spatial users with the data they seek.