PERSPECTIVE

Browsing Metadata

Where Do We Go From Here?

MICHAEL F. GOODCHILD

Michael F. Goodchild, professor of geography at the University of California, Santa Barbara, is this month's Perspective contributor. In the Perspective column, geospatial community leaders voice their opinions about issues facing those who use geospatial technologies. Here, Goodchild discusses the past, present, and future of metadata. In this month's First Impressions, he also reviews SMMS 3.0, one of the many metadata creation and management products on the market. Goodchild can be reached by e-mail at good@ncgia.ucsb.edu, Internet www.ncgia.ucsb.edu, (805) 893-8049, or fax (805) 893-3146.

Great progress has been made in recent years in getting metadata standards adopted and used. New products provide tools for creating, managing, and publishing metadata; standard GIS products increasingly support metadata; the Federal Geographic Data Committee has funded many metadata creation projects; and standards are being discussed at the international level through the International Standards Organization.

Metadata allow us to describe data sets, to advertise their existence to potential users, and to evaluate the fitness of data sets for use. They also enable us to find data in large collections and clearinghouses. So, is the problem of describing, advertising, and evaluating geospatial data solved?

During the past few years, I've been involved with the Alexandria Digital Library (alexandria.ucsb.edu), a major effort funded by the National Science Foundation and other federal agencies to make the resources of a map and imagery collection available over the World Wide Web. I've also participated in broader discussions about the NSDI and related topics as a member of the National Research Council's Mapping Science Committee. Based on those experiences, here are a few thoughts about where the development of metadata needs to go next.

Collection-level metadata

The library is often a useful framework for thinking about information description, search, and related tasks. The development of the printing press in the fifteenth century provoked an explosion in the production of information, and books found their way into the collections of institutions, scholars, and private citizens. Putting the title and author on a book's spine allowed the user of a collection to hunt quickly for suitable information and to avoid opening a book until there was some confidence that it might contain something useful. As libraries grew and it became impossible for anyone to visit all of a large library's shelves, the same information was extracted onto catalog cards and sorted for easy search. Eventually, the information was digitized. Geospatial data sets are undergoing a similar evolution, in that metadata allow us to search large collections based on suitable indicators, including title, spatial coverage, date of validity, and theme.

There is one major difference, though, between a large library collection and a collection of geospatial data. Although one expects a major library to have a copy of any important book, users have very different expectations for geospatial data collections. A given data set, for example, is likely to be found in only one collection, most likely the collection of the agency that developed and maintains the data. After all, one major objective of the NSDI is to reduce duplication in the production and storage of geospatial data. Imagine looking for a book, knowing that it is in one of the major libraries, but with no basis for guessing which one.

In reality, the situation is not quite so discouraging. The National Geospatial Data Clearinghouse (www.fgdc.gov) provides a single point of access to many different collections, but by no means all.

After years in the geospatial data business, one acquires a useful knowledge of where to look for certain types of data. Such knowledge is also metadata, but of a different sort — metadata not about individual data sets,

Abbreviations

NSDI: National Spatial Data Infrastructure

USGS: U.S. Geological Survey

but about entire collections. For example, most of us know that Microsoft's Terraserver features various types of imagery. Although GIS professionals may have acquired such knowledge, it is stored almost entirely in our heads. Someone new to the world of geospatial data would not have such an advantage.

Because geospatial data sets are often managed and served by the same agencies that produced and maintain them, the contents of collections tend to reflect familiar arrangements for the production of geospatial data. For example, digital orthophoto quads are available from servers maintained by the USGS, the primary producing agency for this type of data. A user collecting all of the data available for a given area would likely have to access and search a number of servers — each one offering particular types of data — because collections tend to be organized by theme more than by geographic coverage. Because each collection may use a different search mechanism and may structure data using different formats, it can be enormously time-consuming and costly to assemble data in this way.

A collection-level metadata standard would allow us to describe the contents of geospatial data collections — their geographic, thematic, and temporal coverage — to support users who must identify the servers most likely to contain needed data. It would help in overcoming the problems associated with integrating data obtained from different servers. And it could also lead to a more rational and orderly development of the world's geospatial data resources.

The value of browsing

Any habitual library or bookstore visitor is familiar with the concept of browsing — the process of happening upon a book of interest simply by spotting it on a shelf. Browsing has elements of vagueness and serendipity in it, both concepts that are largely alien to computing systems, which know only how to do things precisely.

Bookstores and libraries both support browsing by how they choose to order books on shelves, usually by subject and sometimes by author. Emulating such a system in the digital realm turns out to be surprisingly difficult. One approach to accomplish this is to devise goodness-of-fit measures between a request and a data set's metadata, such that data sets that don't quite fit the request still get returned but with lower scores. But that assumes that someone can devise suitable metrics in advance and in ways that emulate the results of a vague browsing.

Another approach is to sharply limit the number of metadata items that are used in a search. The Alexandria Digital Library uses only a limited selection of metadata items for searching, but gives the user access to many more items for any data set returned by the search. Similarly, the Online Computer Library Center (www.oclc.org) Dublin Core metadata standard, which has been developed for digital materials in general and not only geospatial data sets, focuses only on a limited number of items; not all of the items, however, that might conceivably be useful for describing data.

User interfaces will have to be designed to show increasingly precise views of collection contents, starting with very general overviews and honing in on specifics, just as the library user does in browsing the shelves.

In practice, the human process of searching a library or bookstore is dynamic — requests are refined in stages as the user browses and gains a general understanding of what might be available and finally settles on a given book. Such negotiations often occur in human interaction as precision gradually replaces vagueness in everyday tasks like providing driving directions or identifying people. Implementing these types of negotiations in computing systems will require something very different from a routine standard query language search of a database table. User interfaces will have to be designed to show increasingly precise views of collection contents, starting with very general overviews and honing in on specifics, just as the library user does in browsing the shelves.

Other issues

We humans are very efficient in everyday communication, making use of gestures, voice inflections, and other ways of adding to the content of pure text. Sarcasm is mostly conveyed through inflection, and "yeah, yeah" is often cited as an instance of how two positives can make a negative. As institutions like the library move enthusiastically into the digital age, it is often worth stopping to ask what the digital world discards. Technologies such as GIS still rely heavily on humans to augment what they are good at and to make their precision useful in a human world.

As a scientist, it makes me profoundly uncomfortable to imagine that one day I could submit a paper for publication based on geospatial data that I "found on the Web," and I am sure reviewers would raise strong objections. Despite its claims of objectivity, science still relies heavily on trust, interpersonal networks that are now facilitated by e-mail but still essentially personal, and on brands like USGS, which automatically imply quality. In that sense, the phone number of the person who last tried to use a data set may be its most valuable item of metadata. How exactly these issues can be factored into automated search technologies that more successfully emulate the process of browsing is a major research and development challenge.

In short, we have made great progress in recent years in solving the problems of describing geospatial data sets and in getting solutions widely adopted within the community. But compared with humans, and their highly evolved skills of communication and strategies for survival and success in a complex world, digital systems remain painfully clumsy in many respects and powerful only in certain strictly limited tasks. ❖