MICHAEL DOMARATZ

Finding and Accessing Spatial Data in the National Spatial Data Infrastructure*

Information about where a thing is or where an event takes place is an important factor in decisionmaking in both the public and private sectors. Spatial data provide a unique context for integrating disparate observations and evaluating competing options. Factors of location, distance, pathways, and other spatial relationships often must be considered when making decisions about economic ventures, environmental and health concerns, responses to emergencies, and other issues.

Public and private sector organizations have quickly realized the usefulness of spatial data in their activities. The nation spends billions of dollars annually on just the collection, management, and dissemination of spatial data. Difficulties in finding and accessing data, and a lack of data documentation, hinder the spatial data community's efforts to work together and leverage this large investment. Through the National Spatial Data Infrastructure, government agencies, private companies, and nonprofit organizations cooperate to develop consistent, reliable means to share spatial data.

THE NATIONAL SPATIAL DATA INFRASTRUCTURE

The National Spatial Data Infrastructure provides a base or structure of relationships among data producers and users that facilitates data sharing. More formally, it is "the technology, policies, standards, and human resources necessary to acquire, process, store, distribute, and improve utilization of geospatial data" (Executive Office of the President, 1994).

The characteristics of the spatial data community greatly influence the approach to developing a "national" infrastructure. The many organizations in the community—including local, regional, State, and Federal government agencies, private companies, and non-profit organizations, have different (and sometimes competing) purposes, abilities, policies, interests, and needs. The many scientific or occupational disciplines in the community have different organizing principles, values, techniques, and terminologies. Some disciplines have a long experience with spatial data. Use of spatial data is quite new in others. All have something to

contribute to the community, although that fact is not always readily apparent to all members of the community! Finally, the members of the community are dispersed geographically, an important factor in evaluating means to link data users and producers. These and other characteristics of the community play a critical role in developing strategies to move forward.

"Spatial data" (sometimes called "geospatial data") identify "the geographic location and characteristics of natural or constructed features and boundaries on the earth" (Executive Office of the President, 1994). Most people would readily identify digital and paper maps, aerial photographs, and remotely sensed images as sources of spatial data. There also are many other types of data, including socioeconomic and demographic statistics, surveys of natural resources, and photographs and videotapes of landscape, that describe the locations and characteristics of geographic features. The spatial component of these data, which might be encoded using geographic coordinates such as longitude and latitude, a street address, or a county name or code, provides a key by which different information sources can be integrated and processed. The infrastructure must accommodate these different data so that potential users can find, evaluate, and integrate them.

An important aspect of the "infrastructure" is technology. Advances in computerized approaches to collect and process spatial data, and decreasing costs for using this technology, have helped spread the use of digital spatial data. Technologies such as the Global Positioning System, geographic information systems, and image processing help organizations that now use spatial data to do so more efficiently and effectively and entice other organizations to use these data for the first time. New, dynamic forms of spatial data are being created. Integration and use of data may result in new data being created.

Advances in telecommunications such as the Internet provide the ability to disseminate these digital data to a large audience. Before this technology was available, many organizations that collected and used spatial data did not have the printing, warehousing, and shipping infrastructure needed to distribute spatial data. The Internet now permits these organizations to make their information widely available, as well as to locate needed data that are produced by others. Traditional relationships within the community are changing rapidly as technology enables the emergence of new data producers and users, and new opportunities for collaborative data collection and use.

Work on infrastructure requires attention to other links within the community. Concerns and views vary widely on issues such as recovering the costs of data collection, freedom of information, and liability. Development of the infrastructure also depends on ensuring that new profes-

sionals are trained well, and that existing professionals keep up with the rapid technological change.

THE NATIONAL GEOSPATIAL DATA CLEARINGHOUSE

One challenge brought on by the changes in the community is that of finding and accessing digital spatial data. The amount of data being produced, the number of organizations producing data, and the decentralization of data production and distribution are growing. Distinctions between data producers and users are being blurred. These data often are not "published" in a traditional way. Mechanisms for finding and accessing information must accommodate these changes in the community.

These concerns are not unique to the spatial data community. Marchionini and Maurer note that "one clear difference between traditional libraries and digital libraries is that digital libraries offer greater opportunity for users to deposit as well as use information" (Marchionini and Maurer, 1995, p. 73). Wilensky provides another view: "For digital libraries to succeed, we must abandon the traditional notion of 'library' altogether. The reason is as follows: The digital 'library' will be a collection of distributed information services; producers of material will make it available, and consumers will find it and use it, perhaps through the help of automated agents" (Wilensky, 1995, p. 60).

Working with other members of the community, the Federal Geographic Data Committee is encouraging the development of the National Geospatial Data Clearinghouse as a means for the community to find and access digital spatial data. The clearinghouse is a referral service to discover who has what data. Designed with the decentralized distribution of data producers and users in mind, the clearinghouse is comprised of a set of information stores that use computer hardware, software, and telecommunications to link producers and users. To participate in the clearinghouse (see figure 1), producers create descriptions (or "metadata") of their data and make these descriptions available through the Internet. The resulting form of the clearinghouse is the "constellation" of sites, linked through the Internet (see figure 2).

Producers also may make their spatial data available directly through the clearinghouse. Many government organizations are taking advantage of this option to disseminate data that are in the public domain. Use of this option by other organizations will grow as methods for commercial transactions on the Internet mature.

To find and access data, a user communicates with the sites through the Internet, retrieves the metadata, and evaluates the metadata to determine the usefulness of available data for the planned application. When

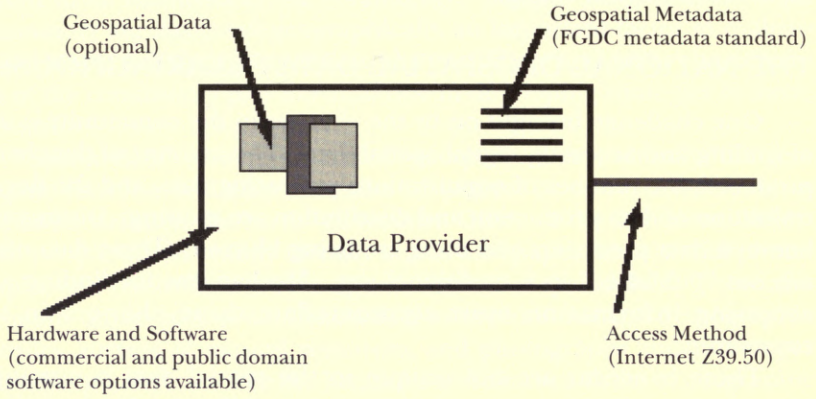

Figure 1. Components of a clearinghouse site.

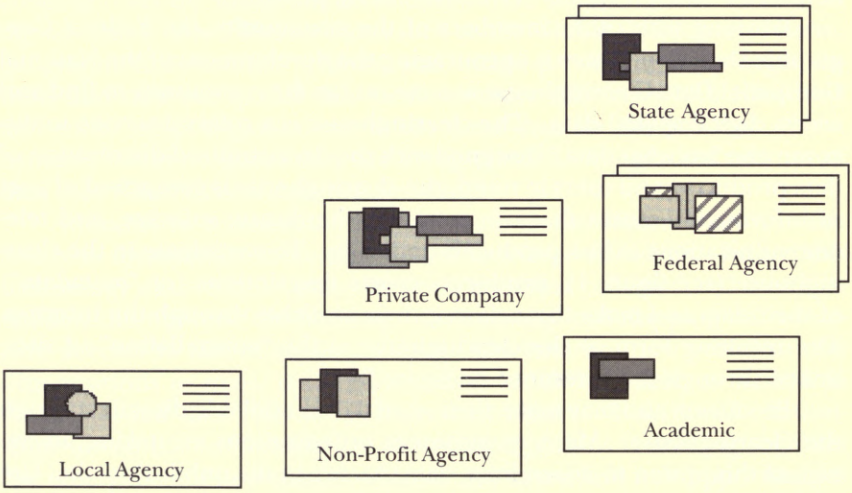

Figure 2. The clearinghouse is a "constellation" of sites from which data producers provide information.

useful data are identified, the user follows the instructions in the metadata for retrieving or ordering the data.

Another perspective on the clearinghouse is to consider different ways through which spatial data can be accessed using the Internet (see Figure 3). Of these alternatives, the clearinghouse is based on data producers providing access to metadata through Z39.50-compliant servers. Popular options include providing access to metadata through World Wide Web servers, and providing access to spatial data through File Transfer Protocol (FTP) services.

This approach to building the clearinghouse has been well received in the spatial data community. Producers value the ability to provide information about their data directly to potential consumers. The approach also encourages producers to keep the metadata current. Users appreciate the ability to access metadata and spatial data from their desks, and the ability to determine for themselves which data are the most suitable for their applications. Federal agencies, required to participate in

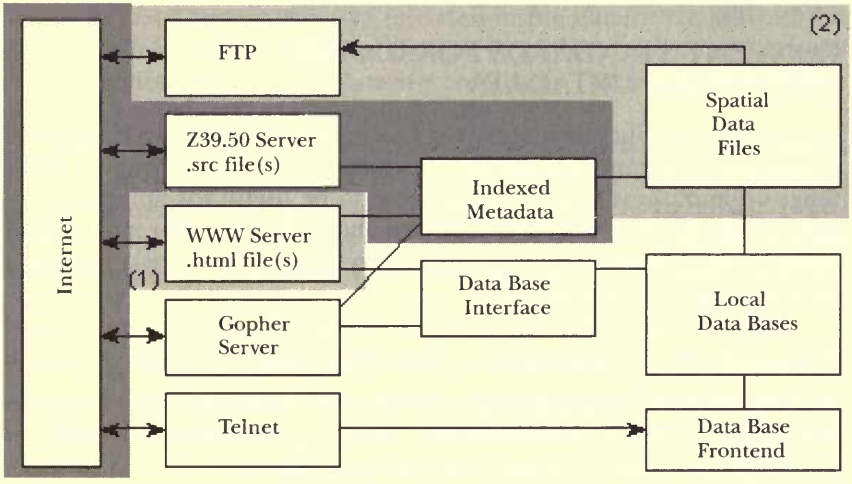

Figure 3. Of the methods through which the Internet can be used for spatial data, the clearinghouse uses Z39.50-compliant servers to provide metadata. Popular options include the use of the World Wide Web and the File Transfer Protocol.

the clearinghouse by Executive Order 12906, have been pleased with the initial response to their efforts. For example, the U.S. Fish and Wildlife Service reported that, in the first month of operation, approximately 29,000 digital maps from the National Wetlands Inventory were retrieved. The U.S. Geological Survey reported a similar volume of interest, with 40,000 files of spatial data retrieved in the first three months of operation (Federal Geographic Data Committee, 1994a).

The implementation of the clearinghouse is just beginning, and there are many challenges ahead. It has been noted that "the Internet is starting to provide the largest library humankind has ever had. As true as this may be, the Internet is also the messiest library that ever has existed" (Marchionini & Maurer, 1995, p. 72). This concern has been noted about the clearinghouse, and the Federal Geographic Data Committee is working within the spatial data and Internet communities to develop means to find and evaluate data more efficiently. New tools for using the Internet will result in new ways to implement the clearinghouse. The committee also sponsors a competitively awarded cooperative agreement program to encourage collaborative experimentation and implementation of the clearinghouse.

CONTENT STANDARDS FOR DIGITAL GEOSPATIAL METADATA

The lack of documentation for existing spatial data also hinders the spatial data community's ability to leverage its data investments. Many times organizations find data that seem to be useful for an application only to discover that very little is known about the data. For most organizations, concern about failure and liability is too great to risk the use of data that are not documented.

Data documentation, or metadata, describe the content, quality, condition, and other characteristics of data. Metadata for spatial data include:

- *Identification Information*—basic information about the data. Examples include the title or other identifier, the geographic area covered, currentness, and rules for acquiring or using the data.
- *Data Quality Information*—an assessment of the quality of the data. Examples include positional and attribute accuracy, completeness, logical consistency, and lineage (the sources of information and methods used to produce the data).
- *Spatial Data Organization Information*—identification of the mechanisms used to represent spatial information in the data. Examples include the method used to represent spatial positions directly (such as raster or vector) and indirectly (such as street addresses or county codes), and the number of spatial objects in the data set.

- *Spatial Reference Information*—description of the reference frame for, and means of encoding, coordinates in the data. Examples include the name of, and parameters for, map projections or grid coordinate systems, horizontal and vertical datums, and the resolution of the coordinate system.
- *Entity and Attribute Information*—information about the thematic content of the data, including the entities types, their attributes, and the domains from which attribute values may be assigned. Examples include the names and definitions of features, attributes, and attribute values.
- *Distribution Information*—information describing how to obtain the data. Examples include the means available to contact a distributor, available data formats, information about how to obtain data online or on physical media (such as cartridge tape or CD-ROM), and fees for the data.
- *Metadata Reference Information*—information about the metadata. Examples include the date the metadata were created and the means to contact the organization that created the metadata.

The data producer is the best source of this information. Details about the boundary of the area encoded in the data, the quality of the data, coordinate systems, data dictionaries, and other elements of metadata are all available when spatial data are produced. The best time to collect metadata is when the data are being collected.

Can data producers be persuaded to collect metadata? Fortunately, a major use of metadata is of great interest to data producers. Metadata provide a means to organize and maintain a producer's internal investment in data. Metadata help organizations to insure themselves from loss of knowledge about their data caused by personnel changes or by the passage of time. Metadata also help to protect organizations from conflicts caused by misuse of data.

In addition to maintaining internal investments in data, there are two other uses of metadata important to the spatial data community: (1) enabling participation in data clearinghouses, and (2) supporting transfers of data. Metadata are the core information of a data clearinghouse. Through the clearinghouse, an organization can find useful data that are available from others and make its data known to new customers. An organization also can identify other organizations with similar interests that may be potential partners in data collection and maintenance activities. Metadata also are essential information during the transfer of spatial data between organizations. For data to be useful to an organization, the organization must be able to integrate them into its holdings and applications. Metadata provide critical information needed to process and ingest new data.

As the spatial data community recognized the value of metadata, interest in standards for metadata grew. The Federal Geographic Data Committee sponsored a forum on the subject in 1992. At the forum, the participants heard descriptions of different approaches to setting standards for metadata, and agreed on the need for a standard on information content for metadata for digital spatial data. Volunteers drafted a standard which the Federal Geographic Data Committee offered for public review from October 1992 to April 1993. Extensive comments were received from the public. The committee revised the draft based on these comments and tests conducted by its member agencies. The committee also coordinated its efforts with those for related activities, such as the Machine Readable Cataloging (MARC) and the Government Information Locator Service (GILS). The standard was approved by the committee on June 8, 1994 (Federal Geographic Data Committee, 1994b). Geographic information coordination committees in several states also have adopted the standard.

The standard supports the common uses of metadata: to enable an organization to protect its internal investments in data, to support data clearinghouses, and to support data transfer. The standard provides for the encoding of information needed to satisfy common uses of metadata:

- availability—data needed to determine the sets of data that exist for a geographic location.
- fitness for use—data needed to determine if a set of data meets a specified need.
- access—data needed to acquire a set of data.
- transfer—data needed to process and use a set of data.

The standard provides a common set of terminology and definitions for the documentation of spatial data. The standard establishes names for data elements and groups of data elements, the definitions of these data elements and groups, and information about the values that are to be provided for the data elements. Information about elements that are mandatory, mandatory under certain conditions, and optional (provided at the discretion of the data producer) also is provided.

The standard specifies the information content of metadata, but does not specify how this information is organized in a computer system or in a data transfer, nor the means by which this information is transmitted or communicated to the user. The variety of means for organizing data in a computer or in a transfer, the different institutional and technical capabilities of data producers, and the rapid evolution of means to provide information through the Internet provided the basis for this decision.

Recognizing the different needs and abilities within the spatial data community, the standard provides leeway in a number of implementa-

tion decisions. Elements of metadata can be encoded for different levels of granularity of data, ranging in size from large collections of data files to individual lines and areas. The amount of detail that is encoded may vary. Different data may vary in their significance or value, and the effort expended in developing metadata should correspond to the value of the data. Decisions about documenting existing and new data should be considered carefully. Organizations with large holdings of undocumented "legacy" data are concerned about the costs of documenting these holdings, and sometimes allow the consequences of past practices to control their decisions about documenting new data.

The Federal Geographic Data Committee recognizes the need for the standard to evolve with the changing needs of the community, and is working with the community on improvements. Current activities include identifying a core set of metadata to facilitate searches, developing means of providing "lite" amounts of metadata, and developing means of adding locally used extensions to the standard.

FUTURE DIRECTIONS

The factors which have fashioned activities to develop the National Spatial Data Infrastructure will continue to challenge current views about how to collect and share digital spatial data. Successful approaches will be those that allow the community to contribute, share, integrate, and use spatial data for varying units of space, periods of time, and thematic detail. It is difficult to know exactly what may emerge from this dynamic environment. Chrisman (1994) has identified some things that a digital library of geographic information should *not* be: it is not a collection of map sheets; it is not just one snapshot in time; and it is not modeled on a digital library of books or scientific publications.

The Federal Geographic Data Committee sponsors projects to develop the National Spatial Data Infrastructure and to improve the community's ability to work together. For more information about the committee's activities, visit the committee's World Wide Web site at <URL: http://fgdc.er.usgs.gov>, or contact the committee by electronic mail at gdc@usgs.gov or by postal mail at the FGDC Secretariat, c/o U.S. Geological Survey, 590 National Center, Reston, Virginia 22092, USA.

REFERENCES

Chrisman, N. (1994). A vision of digital libraries for geographic information, or how I stopped trying to find the on-ramps to the information superhighway. *Geo Info Systems*, 4(4), 21-24.

Executive Office of the President. (1994). Coordinating geographic data acquisition and access: The National Spatial Data Infrastructure (Executive order 12906). *Federal Register*, 59(71), 17671-17674.

Federal Geographic Data Committee. (1994a). *Content standards for digital geospatial metadata.* Reston, VA: Federal Geographic Data Committee.

Federal Geographic Data Committee. (1994b). *The National Geospatial Data Clearinghouse: A Report on federal agency activities within the First Six Months of Executive Order 12906.* Washington, DC: Federal Geographic Data Committee.

Marchionini, G., & Maurer, H. (1995). The roles of digital libraries in teaching and learning. *Communications of the ACM,* 38(4), 67-75.

Wilensky, R. (1995). UC Berkeley's digital library project. *Communications of the ACM,* 38(4), 60.