Integrating Spatial Metadata and Data Dissemination Over the Internet

by Robert G. Cromley & Patrick McGlamery *

Abstract
As more spatial databases are compiled and made available for dissemination via the Internet, there is an increasing need for metadata descriptions of the downloaded data to be available on demand. This paper describes a system in development at the University of Connecticut that not only allows users to select their data dynamically but also to prepare FGDC compliant metadata records for these data that can be downloaded in the same session. In this manner, users will have a more complete knowledge of how to use this information effectively.

Introduction
Thematic maps have been a part of research library collections for the past century. Only since the later 1970s though has the choropleth map become a ubiquitous tool for visualizing demographic information in the United States and have these thematic maps found their way in growing numbers into the map library. The 1970 Census and the DIME (dual, independent, matching, encoding) files, dependent though they were on large mainframe computers, set the stage for our expectations of network delivered, on-demand demographic mapping and data dissemination.

During the 1990s, radical innovations in technology increased the computing power of the personal computer (PC), developed the CD-ROM as a distribution media, and created a burgeoning network infrastructure and browsing software, resulting in the Worldwide Web (WWW) as we know it today. While all of these technologies have resulted in the growth of data dissemination from map libraries as well as social science data libraries, perhaps the most significant technological change was the replacement of magnetic tapes by CDs. This particular shift 'democratized' social science data by moving it from the constraints of mainframe computing and putting the data in libraries and hence on scholar's workstations.

In the meantime, Geographic Information Systems (GIS) have been changing the ways in which we look at geographies. Technology has had a huge impact on GIS, freeing it first from the mainframe, then from high-end UNIX workstations. Powerful PCs and our networked environment on the WWW have created a dynamic, online versions of GIS. (Green and Bossomaier, 2002). Now GIS is expanding the user base of who is looking at geography… and maps. These innovations have put pressures on the map library. In the past two years over a dozen new "GIS Librarian" positions have been established and staffed in research libraries around the United States. In fact, the GIS Librarian is a geodata librarian focusing primarily on TIGER (topologically, integrated, geographic, encoded, reference) line graph files, demographic data and mapping; this is a result of the Association of Research Libraries' (ARL) GIS Program that began in 1993 (ARL, 1995).

> The goals of the ARL GIS Literacy Project were designed to meet the current needs of libraries and users while addressing the changes that libraries are facing during this time period of experimentation, transition, and transformation to networked- based services. These goals include the following:
>
> - Introduction of GIS to a variety of libraries (e.g., public, state-based, academic, and university libraries in public and private institutions) to address diverse user information needs;
> - Development of a team of GIS professionals in the research library community who are willing to lend time and expertise to applications, user training, and education programs;
> - Encouragement of connections among federal, state, and local GIS users and information;
> - Promotion of research, education, and the public right-to-know through improved access to government information;
> - Initiation of library projects to explore new applications of spatially referenced data and to evaluate the introduction of these services in research libraries; and
> - Implementation of programs to allow institutions that have invested in networking capabilities to leverage the sharing of resources via networks.

Working with ESRI, a GIS software developer, the ARL

program worked to reorient libraries toward considering the geoprocessing of social science data, especially Census data.

Traditionally, map libraries have had the responsibility for the storage of and access to cartographic information. A library collects, catalogs, stores and provides access to information but does not produce the data itself. Increasingly, the producers of spatial data are only distributing it as a digital database; not as paper or even as viewable information. In 1990, the U.S. Bureau of the Census stopped producing census tract and other printed maps (for the 2000 census some maps can now be downloaded as PDF files). This almost total conversion of information from a paper to an electronic format has forced map libraries to rethink how services are to be provided.

Metadata

With the development of the Internet in the 1990s, an opportunity was created for transmitting digital cartographic and social science information to a global user community. However, as more databases are compiled and made available for dissemination via the Internet, there is also an increasing need for descriptions of the downloaded data to be available on demand. These 'data about data' or metadata traditionally play four roles in the archiving of data (FGDC, 1997):

- availability – data to determine what data exists
- suitability of use – data to determine proper data uses
- access – data to acquire a set of data, and
- transfer – data needed to process a set of data.

Metadata regarding the content, quality, and other characteristics of disseminated data are critical not only in transferring data from an external source but also in interpreting that data by the end user; that is, it acts as a codebook. Codebooks describe the structure, contents and layout of a datafile. The purpose of this paper is to discuss the issues surrounding the construction of metadata records for customized databases, especially dynamic spatial databases with demographic attributes, disseminated over the Internet.

Digital Map Formats

In a recent book on web-based cartography, Kraak (2000) defined two types of web maps: static maps and dynamic maps. Static maps are images displayed on a browser whose elements cannot be changed. On the other hand, users can not only change elements of dynamic maps but also interact with them. However, although the visualization of some maps may be static, the process of their construction may be dynamic. A static map may exist as a predefined image or, using the Common Gateway Interface (CGI), a Server-side utility program can create a customized image (Kobben, 2000).

The same dichotomy can be applied to spatial databases being disseminated over the Internet. Some spatial databases may be static - predefined files that are downloaded - or, some may be dynamic - customized files that are created by the user in real time. For static spatial databases, metadata descriptors are contained in separate files that can be viewed and/or downloaded with the database. Dynamic spatial databases create new questions for metadata records that are not adequately covered by existing standards because the descriptors of the data set being transferred are in part created in real time. In constructing dynamic spatial databases, both the geography and/or the attribute information can be defined by the user, and the user can select one or both. One cannot anticipate what the user will select. Therefore, descriptors need to be created *a posteriori* for dynamic spatial databases whereas descriptors for static spatial databases can be created *a priori*.

The Nature of Metadata

Metadata is a dual-use concept. The metadata content in an HTML (hyper-text, markup language) or TXT (plain text) formatted file is a codebook for the user. As an SML (standard markup language) or XML (extensible markup language) formatted file, it is tagged to create an index for search, query and discovery in a clearinghouse network. If the content elements are stored in a database, a PRINT statement can generate metadata information in either an HTML, TXT, SML, XML or any other format depending on the needs of the user. However, a complete storage of content elements is only possible for static spatial databases. For dynamic databases some content elements that are known a priori can be stored in a database whereas other elements that are defined by user choices and are created "on-the-fly". Since most users downloading data are mainly interested in the codebook aspects of the metadata, this paper focuses on the codebook concept for the a posteriori construction of metadata for dynamic databases.

Metadata Standards

This discussion is on the application of Federal Geographic Data Committee (FGDC) standards to dynamic spatial databases because these are the national standards used by most distributors of spatial metadata. There are seven major components of this metadata standard (FGDC, 1997):

- Identification information which contains basic characteristics of the data set including a description of its content, its spatial domain and its time period of content;
- Data Quality information that provides a general assessment of a data set's quality and suitability of use,
- Spatial Data Organization information that describes the mechanism used to represent spatial information in the data set;

- Spatial Reference information that describes the reference frame used to encode spatial information;
- Entity and Attribute information that outlines the characteristics of each attribute including its definition, domain, and unit of measure;
- Distribution information that identifies the data distributor and the options of obtaining the data, and;
- Metadata Reference information that describes the currentness of the metadata and the party responsible for maintaining it.

In addition to these main components, **Citation**, **Time Period**, and **Contact** information are important sub-components that are repeated under different primary components. While the FGDC standards provide for extensive description of spatial characteristics, its abilities with Entities is less robust. The Social Science data community's Data Documentation Initiative (DDI) standards are more definitive. While the developing DDI is actively adopting geodata descriptors, FGDC, bearing the burden of a seven-year legacy, has been less responsive.

Spatial Data Issues

There are different issues associated with the spatial database if the underlying geography is area-class or choroplethic. For an area-class geography, the geo-units are defined by an attribute for which the associated attribute set is well defined and closed (probably better suited for a static database). For a choroplethic geography, the geo-units are usually defined as political or administrative entities for which the associated attribute set is more open-ended (probably better suited for a dynamic database). These differences have implications for metadata records. For example, an area-class database is more likely to have the same originator for the geographic and the attribute information because these are intertwined. A choroplethic geography is more likely to have different originators - one for the geography and perhaps multiple originators for the attributes.

Likewise, the geography and attribute lineage information for an area-class database should be the same, whereas these lineages should be different for a choroplethic database.

There are numerous examples then where a clearer distinction should be made in the construction of the geography from the construction of the attributes.

There are two additional considerations. Given that the Internet is a distributed network, the data behind a customized database can be distributed over many organizations and locations. The metadata for the customized database needs to capture the complexity of this system. Some users will only want attribute data and not the geography for use in non-visual analyses. Because the attributes are geo-referenced, though, some description of the geo-units is still necessary.

Constructing a Dynamic System

At the University of Connecticut, we have been working on developing a system to generate customized spatial databases and their associated metadata records. The system is designed to create a full spatial database (in progress) or just a geo-referenced attribute database. In this system, metadata are generated both from existing meta-databases as well as the user's own query responses. The metadata in these databases are also used by the system to compile and retrieve the customized database. In addition the user defines: 1) a study area (the spatial domain), 2) a geographic unit of inquiry (the spatial resolution), and 3) the set of attributes.

The tasks of meta-database organization involve: 1) determining the FGDC metadata elements relevant to a geography database and those relevant to a geo-referenced attribute database; 2) separating those elements with both geography and attribute descriptors into separate tags; and 3) deciding which elements cover a whole database and which cover elements of the database (Figure 1).

Assigning the basic FGDC metadata elements is relatively easy. Identification, Data Quality, Spatial Data Organization, Entity and Attribute, Distribution, and Metadata Reference information all belong to both, whereas Spatial Reference information is only relevant to the geography. Separating elements is more difficult. For example, within Identification Information one can separate Citation, Description, Time Period of Content, Status, and Native Data Set Environment into Spatial Data and Attribute Data Tags. On the other hand, Spatial Domain, Keywords, Access Constraints and Use Constraints only have one Tag.

Deciding which elements cover a whole database and which ones cover elements of the database is also more difficult. Customizing the geography means that Spatial Reference Information is not at the level of a whole geography database but at the level of the elements within the database whereas Spatial Reference is at a whole geography level. Customizing attributes means that Originator and Lineage may belong at both the dataset and individual field level.

While our current development has been for FGDC content standards, we are creating a Meta-Tag Database for cross-walking between different metadata standards such as FGDC, DDI, the FERRET Project's MIF (metadata interface file), Dublin Core, and the library communities MARC formats. For example, the name of an attribute has the following Meta-Tags: Attribute_Label in FGDC; attr name in DDI or M in MIF. This will enable users from a variety of research communities to access metadata codebooks in familiar formats. It will also streamline

metadata creation for clearinghouse indexing.

Conclusions

When designing a dynamic spatial database dissemination system, there are more considerations necessary in the construction of metadata records than for a static system. The nature of the dynamic system requires that a metadata codebook reflects the impromptu nature of the data query and extraction. Some queries will be directed to obtain only attribute data that can be used in statistical analyses; other queries may want the full geography as well. The metadata needs to reflect these choices made by the user. In addition, demographic and other social science attribute data, may have different source organizations and locations. The metadata codebook needs to capture the complexity of the source information. By providing full narrative information for numeric datafiles, users have a value-added product.

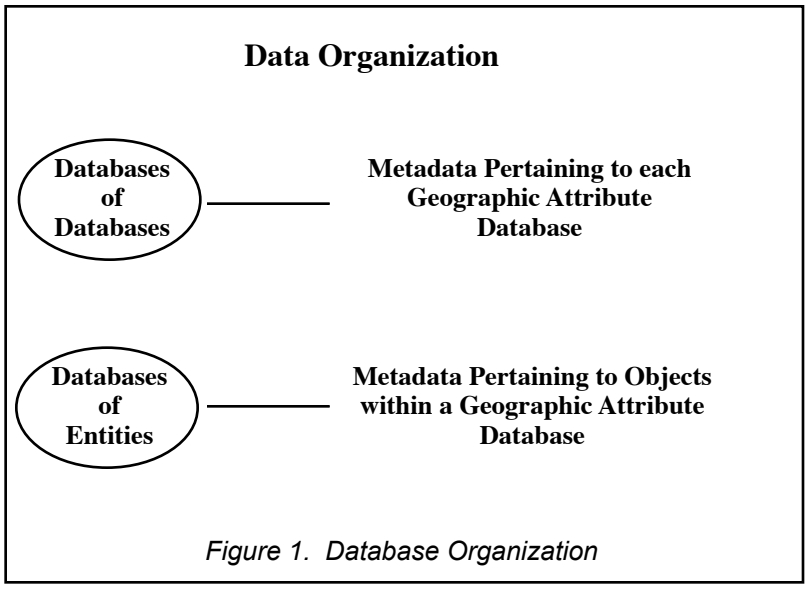

Figure 1. Database Organization

Bibliographic References

Association of Research Libraries (1995), The ARL GIS Literacy Project, ftp://www.arl.org/info/gis/gis.descrip.

Federal Geographic Data Committee (1997), FGDC Standards Reference Model, http://www.fgdc.gov/standards/refmod97.pdf.

Green, D. and T. Bossomaier (2002), Online GIS and Spatial Metadata. New York: Taylor and Francis.

Kobben, B. (2001). Publishing maps on the web. Chapter 6 in Web Cartography, M-J. Kraak and A. Brown (eds.), New York: Taylor and Francis.

Kraak, M-J. (2001), Settings and needs for web cartography. Chapter 1 in Web Cartography, M-J. Kraak and A. Brown (eds.), New York: Taylor and Francis.

* Paper presented Robert G. Cromley[+], Department of Geography U-4148, 215 Glenbrook Rd., University of Connecticut, Storrs, CT 06269 USA, Phone (860)-486-2059, cromley@uconnvm.uconn.edu. and Patrick McGlamery, Homer Babbidge Library, University of Connecticut, Storrs, CT 06269 USA [+]All correspondence should be directed to this author.