The Map Library's Emerging Role in the Dissemination of Cartographic Information on the Internet

Patrick McGlamery
Homer Babbidge Library

Robert G. Cromley
Department of Geography
The University of Connecticut
Storrs, CT 06269

The Internet is allowing a range of cartographic products from images of map documents to numerical databases of cartographic content to be transmitted to a global user community. This technological innovation is forcing map libraries to rethink the manner in which to provide their services since libraries have traditionally had the responsibility for the storage of, and access to, information by society. The functions of a map library that allow a patron to search the holdings, go to the storage location, browse the document, and ultimately copy it in-house or check out the document can now be provided online. This paper describes the efforts and problems of collection development, assessment of user community needs and access policies associated with an internet-based map library.

INTRODUCTION

With the advent of the Internet in the 1990s, a range of cartographic formats from images of map documents to numerical databases of cartographic content can now be transmitted to a global user community. The form and method of this dissemination is of particular concern to libraries that traditionally have had the responsibility for the storage of, and access to, information by society. In 1999, a Mapping Sciences Committee Workshop, "Distributed Geolibraries" described a vision of the future of cartographic information http://cartome.org/distributed-geolibraries.htm#summary:

"A distributed geo-library is a vision for the future. It would permit users to quickly and easily obtain all existing information available about a place that is relevant to a defined need. It is modeled on the operations of a traditional library, updated to a digital networked world, and focused on something that has never been possible in the traditional library: the supply of information in response to a geographically defined need. It would integrate the resources of the Internet and the World Wide Web into a simple mechanism for searching and retrieving information rele-vant to a wide range of problems, including natural disasters, emergencies, community planning, and environmental quality. A geo-library is a digital library filled with geo-information—information associated with a distinct area or footprint on the Earth's surface—and for which the primary search mechanism is place. A geo-library is distributed if its users, services, metadata, and information assets can be integrated among many distinct locations."

"A distributed geo-library is a vision for the future."

The challenge for libraries is to evolve the "operations of a traditional library" from *Library as Institution to Library as Function* by integrating its knowledge base of how collections and users interact.

Libraries are not information producers; a library collects, catalogs, stores and disseminates data but it does not create the data. Traditionally, a library patron enters a facility, and makes a query of the information stored by the library through and organized query system. This used to be a card catalog but now is an On-line Public Access Catalog (OPAC). An OPAC is a relational database that leads the patron through a process of

search and discovery to an item on a shelf. The patron can then go to the bookshelf and retrieve the document, browse it in-house, copy it in-house or check out the document.

Historically, libraries have organized the storage of information by media type: books, microfilm, serial publications, maps and other formats. It reflects the economic imperative of storage units, keeping "like with like". Within the media type, such as books or maps, the organization is thematic, but the primary categorization is media type. For example, map libraries typically have held maps in flat and vertical files, folded on shelves, bound into atlases and on micro media. Other cartographic information such as gazetteers is shelved with atlases because they are bound as books even though the information is in a very different form from a map. Regardless of media type or thematic content, the OPAC contains the storage location for each item.

"Increasingly, producers of spatial data are distributing it only as digital images and databases."

Increasingly, producers of spatial data are distributing it only as digital images and databases. In 1990, the U.S. Bureau of the Census stopped printing census tract maps. For the 2000 census the Census is producing these maps as Adobe PDF images via the Internet. This transformation of information from paper to electronic form has required that libraries redesign their information delivery service. Buckland (1992) has argued that since library materials in electronic form lend themselves to remote access and shared use, the assembling of local collections becomes less important. Coordinated collection development and cooperative collections are now more strategic. This concept of coordinated collections underlies most attempts to provide access to cartographic information over the Internet.

USING THE INTERNET FOR MAP LIBRARY FUNCTIONS

Two types of Internet sites involving cartographic information have evolved: 1) sites in which the cartographic information is ancillary to another purpose such as promoting tourism and travel, and 2) sites in which the cartographic information itself is the main topic of interest. Sites in the latter case provide differing sets of functions traditionally associated with a map library. The site may just be the equivalent of an OPAC that enables an Internet patron to search for selected items, or the site may also have stored information that the patron can browse on-line and even download to a local desktop.

The Alexandria Digital Library, http://webclient.alexandria.ucsb.edu/ a part of the University of California Digital Library (UCDL), is a cartographical search engine that queries several of the map libraries cooperating in the UCDL. A user of the website makes queries using a spatial index (a latitude/longitude bounding rectangle). The search engine ascertains the location of holdings regarding the selected place, but the patron can neither browse, copy nor check-out any of these holdings. A descriptive catalog record is provided, however few opportunities for data downloads have been implemented. When data downloads are available, for example to download a Digital Raster Image, those users who are not students, faculty and staff of the University of California system are required to pay a fee.

"The site may . . . enable an Internet patron to search for selected items, or the site may also have stored information that the patron can browse on-line and even download to a local desktop."

The Harvard Geospatial Library (HGL) http://geodesy.harvard.edu/servlet/MainGeodesyMap is a developing cartographical catalog whose goal is to:

> "alleviate the most common challenges users face when they embark upon a geospatial analysis project: finding interesting data, obtaining that data in a useable form, learning to use new data analysis tools, and accessing appropriate computing platforms."

The HGL's user interface is more intuitive than the ADL interface, but it is still basically a catalog which refers the user to a static dataset of information object which might then be downloaded, or not. Although Harvard is a private university, it makes more of its data available to the web-user, then it can reasonably interpret its license agreements to data.

Although the Federal Geographic Data Committee's (FGDC) clearing house program is not a library, it uses library functions in a similar approach.

> "The Geospatial Data Clearinghouse http://130.11.52.184/FGDC-gateway.html is a collection of over 250 spatial data servers, that have digital geographic data primarily for use in Geographic Information Systems (GIS), image processing systems, and other modeling software. These data collections can be searched through a single interface based on their descriptions, or 'metadata'."

The user retrieves the metadata records that describe the spatial data and indicate availability. The metadata indicates to the user where the holdings are and what the mechanisms are for acquiring the spatial data.

"Fundamental to library management is a keen understanding of the user community."

ESRI's Geography Network http://www.geographynetwork.com/ is another example of non-libraries offering a library function. The Geography Network is:

> "a global community of government and commercial data providers who are committed to making geographic content easily accessible… Through the Geography Network, you can access many types of geographic content including live maps, downloadable data, and more advanced services. The Geography Network content is distributed at many locations around the world, providing you access to the latest information available directly from the source."

Fundamental to library management is a keen understanding of the user community. Both ADL and HGL are grounded in the primacy of the user. In comparison, the Geographic Network and FGDC Clearinghouse invite participation from the geospatial data producing communities, aiming to aggregate collections, but they lack that key component of a library—collection building with a special user community in mind. These programs rely on a 'scatter-shot' approach to collection building based upon available data, not on user needs. These approaches differ from the more focused library strategy for map librarians in an age of accessing of machine-readable information. McGlamery (1989) has outlined a 'plan of action' for the information age that focused heavily on: 1) collection development, 2) user community needs, and 3) access policies—the underpinnings of modern library science. The next sections describe how this plan has been implemented in building an Internet-based map library.

THE DEVELOPMENT OF MAGIC WEBSITE

In 1991, the Map and Geographic Information Center (MAGIC) FTP site at the University of Connecticut evolved from that 'plan of action' and the site has not strayed far from those tenets of basic librarianship (after the introduction of HTML the FTP site became a website http://magic.lib.uconn.edu). The basis of the plan for the map librarian's dilemma with respect to machine readable information was to 're-bind'—making odd things fit into a standard collection—public domain spatial databases into formats required by the University of Connecticut's user community. While many agencies are producing digital databases, rarely do they produce data in a geographic and software format that is directly compatible

with the needs of most data users. Clarke (1995) has commented that most items of geographic interest seem to lie on the border of two or more map sheets. For example, in Connecticut the basic geography used by most policy and decision makers is that of 169 town municipalities. However, the digital line graph files produced by the U.S. Geological Survey are organized spatially by quadrangles, whereas TIGER line graph files are organized spatially by county. Neither of these geographic units have much utility for most users of Connecticut data.

Therefore, the first stage in the development of a digital collection was establishing the spatial domain and geographic unit analysis within that domain. The map library established the town in addition to the state, county and quadrangle as its basic domains and counties, towns, census tracts and census block groups as the geographic units within the appropriate domains. The 169 towns in Connecticut were extracted from the TIGER line graph files of the eight counties in the state to create census geographies and street coverages. The same towns were extracted by the state Department of Environmental Protection from the DGL files of the 118 quadrangles comprising the state for other features such as hydrography and roads. The files were then projected into Connecticut State Plane NAD 27, the state standard at the time. Finally, the files were converted into the ARC/INFO coverage interchange file format (E00) and MapInfo interchange file format (MIF/MID) from their native formats. The files were zipped, put up on an open FTP site and made available to the public user, retaining their public domain status.

"The first stage in the development of a digital collection was establishing the spatial domain and geographic unit analysis within that domain."

Use of this site now averages 9,000 zipped data files downloaded each month. The data on MAGIC are still primarily of Connecticut and are still in the public domain. Over time, MAGIC has established connections with local producers of state data and there has been strong cooperation and trust in sharing spatial data. State agencies recognize the public's need for high-quality data and the resources required for distributing the data themselves. MAGIC now has over 20,000 files from the U.S. Bureau of the Census, the U.S. Geological Survey, Connecticut's State Departments of Environmental Protection, Transportation, and Public Safety, the National Resources Conservation Service, and the U.S. Fish and Wildlife Service. Towns are also just beginning to provide their data for library distribution.

Although digital databases were the first form of spatial information to be collected and stored on MAGIC, over time map documents and other images were scanned and stored as raster graphic files. The University does not hold the maps that the scanned images represent; most reside in private libraries of libraries far from Connecticut. Through the Internet, MAGIC has been able to build a public collection of Connecticut's cartographic lineage. Recently members of the School of Engineering reference a 1764 survey of the town of Lebanon to a GPS cadastral survey. Faculty and students were surprised to witness the high quality of the surveying done in Connecticut 250 years ago that was completed using only astronomical observations, pencil and paper. Linking the image of the manuscript map alerted the engineers to ancient controls, enabling them to build out their survey. The MAGIC website brought historical data together with current state of the art spatial data for their use.

"Through the Internet, MAGIC has been able to build a public collection of Connecticut's cartographic lineage."

Throughout its existence, profiling the MAGIC user has been an important aspect of maintaining its digital geolibrary. The use of the data, which data are used most, who uses the data, and which level of geography is most used are key bits of information for the collecting librarian. It is simply not enough to passively collect materials, whether they are books, journals, maps or data. Libraries monitor use statistics and assign budgets

accordingly. MAGIC has eight years of comprehensive transaction logs that chronicle the use of the collection. Decisions based on this transaction data have directed MAGIC to acquire more statewide data and directly lead to building the collection of historical map images. The image files of historical maps now account for thirty percent of the data accessed from MAGIC.

"In the conversion from paper to digital media, one of the primary gains has been the separation of the storage and display functions of maps."

Although the MAGIC website has expanded its collections considerably in the past decade, it is still basically a site for passively downloading compressed ASCII export files. Vector data have been augmented with image data; digital orthophotography, scanned historical maps and other remotely sensed data. However, even with these additions, some fundamental structural flaws in the data organization emerged. The University research user community's need to search for, discover, view and acquire timely, and historical social science attribute data and associative digital cartographic data has lead to a complementary website for the dissemination of geographically referenced attribute data for Connecticut.

THE DEVELOPMENT OF THE CTDATA WEBSITE

In the conversion from paper to digital media, one of the primary gains has been the separation of the storage and display functions of maps (Marble, 1987). The storage of spatial databases in the vector data model also separated the storage of the geographic base file (GBF) of the spatial entities from their associated attributes. This latter separation forms the basis of the hybrid architecture used by some geographic information systems in which the spatial entities are independently stored in a different module from the non-spatial attributes (Worboys, 1995). This structure is referred to as the geo-relational model when the non-spatial attributes are stored in a relational database that interfaces to the corresponding spatial entities (see Morehouse, 1985; Waugh and Healey, 1987). The geo-relational model is the basis for the organization of many software formats such as ARC/INFO coverages and ARCVIEW shape files.

When MAGIC built its digital collection of spatial databases, the geo-relational model was the basis for the stored database. A coverage was created for each theme of data. When building a coverage for each theme, however, the question arises as to what attributes will be included in he relational table associated with the geographic base file. The answer to this question is straightforward for coverages based on continuous field data. For field data, the attribute is assumed to vary continuously over space (Burrough and McDonnell, 1998). The elements of the geographic base file are determined by the spatial distribution of the associated attribute. For example, in vector-based soils coverage the polygon outlines only refer to geography of soil classes and no other attribute. Thus, only one non-spatial attribute is associated with the geographic base file, although many attributes can be included in a table that names the elements or defines spatial relationships to other geographies. The answer to this question is less obvious for coverages based on entity data. For entity data, the object exists independently of the attributes that are used to describe it; many polygon coverages or entity data are merely collection zones for which summary attributes are compiled. For the geography of Connecticut towns, the U. S. Census collected approximately 5000 attributes in the 1990 census in just the STF3A file for population and housing. The same number of attributes exists for the county, tract and block group geographies. In preparing town coverages of population for the MAGIC collection, the small set of basic demographic attributes was preselected for inclusion in every town coverage. Similarly, in preparing town coverages of housing, a small set of housing attributes was also preselected for each coverage.

"When MAGIC built its digital collection of spatial databases, the geo-relational model was the basis for the stored database."

The result is a duplication of geographies in the collection; in this case the geography of towns, census tracts and block groups is repeated in the population and housing coverages.

The geo-relational model handles this problem by only requiring that the geographic base file contains a unique identifier field that can be used in a relational join operation to attach any associated attribute table. In a collection of coverages, it is only necessary to store each GBF once as long as the associated attribute table has a minimum of attributes that name each object and provide unique identifiers for subsequent joins. The attributes for describing the objects can be stored in separate sites as long as the proper key fields are present. For coverages of true field data, a separate site is not necessary because only one thematic attribute relevant to the coverage should be included in the table for that coverage. For entity data, however, no site existed that could easily provide the attribute information in a proper table format for all of the geographies that are specific to the Connecticut user community needs.

". . . it is only necessary to store each GBF once as long as the associated attribute table has a minimum of attributes that name each object and provide unique identifiers for subsequent joins."

The U. S. Census website provides access to the 1990 census of population and housing and other intercensual population estimates, but is not organized in a relational format. Because census geographies are hierarchically nested, the census website was designed to retrieve the attribute information in a hierarchical manner. Because the geographical hierarchy skips from state to county to tract to block group, a table can only be built for all objects nested within the next higher geographic level. This means that to extract block group attributes for the city of Hartford, fifty separate tables must be extracted (one for each census tract). Extracting tract data is less cumbersome for Connecticut towns because each town is contained in only one county and all of any tracts in one county can be retrieved in one table. However, a user must know the census codes for the tract identifiers associated with any town in order to retrieve just that town's tracts—a situation that rarely occurs within the general user community.

To overcome these problems of duplicating geographic base files as well as simplifying the extraction of Connecticut attribute data, the Map Library partnered with he University of Connecticut's Center for Geographic Information and Analysis (UCCGIA) to develop the Connecticut Data Server (CTDATA) http://ctdata.lib.uconn.edu. Attribute information is extracted from this site by first defining the relevant study area of inquiry. At present, data can be extracted for the entire state, a county, a town, a labor market area (LMA), a regional planning district (RPD), a congressional district, a state service delivery area (SDA), or a tourism district. After the study area is defined, the user then defines the geographic units of resolution—some subdivision of the study area. All of the previous units are subdivisions at the state level as well as census tracts (block groups are in preparation). Towns and census tracts are valid subdivisions also of counties, LMAs, RPDs, SDAs, tourism districts, and congressional districts (towns split among two or more congressional districts are assigned to the district in which the largest portion of the town's population resided in 1990). After the geographic parameters are chosen, the user is then presented with a choice of databases that are relevant to the chosen geography. Once the desired database is selected, the user can select the specific data fields of his/her interest to be prepared in the data table.

"After the geographic parameters are chosen, the user is then presented with a choice of databases that are relevant to the chosen geography."

While data can be extracted from this site for many purposes, it was designed to distribute information about regions within Connecticut at different levels of geographic resolution in a format that could later be linked to geographic data for later use in a geographic information system. Simultaneous to the development of CTDATA, ARCVIEW shape files

for each possible geography were prepared for inclusion on the MAGIC site. Each shape file only contained in its attribute table a name field and appropriate key field(s). For towns, more than one key field was included because towns are referenced as minor civil divisions in census geographies, and the state has developed its own identification number for each town. CTDATA has two output formats: a table format for display on the browser, and a comma delimited format that can be saved as a text file (extension .txt) by the browser. The text file that is created can then be imported into ARCVIEW as a table and joined to a compatible geography.

Thematically, the data currently on the system cover U. S. Census population estimates, historic town population counts, employment data, the 1998 town profile series and the 2000 Census Public Law data used in redistricting. Most of these data are at the town level because that is the most important geography for Connecticut. New school district geographies and their associated attribute tables are under present construction. These new data present new problems because school districts change more frequently and are specific only to certain grade levels.

CONCLUSIONS

The process of a networked map library for the dissemination of cartographic information at the University of Connecticut is a constant evolution. The current geo-relational approach between the MAGIC and CTDATA websites is the product of many trials and errors over time. It falls short of the vision of a distributed geolibrary's goal of efficient information storage and use oriented retrieval of cartographic information. The next level of consolidation will involve the movement to a full geodatabase approach in which the GBF information is also stored as attributes of an object in a relational table. The geography of objects would be retrieved based on the selected study area. This would reduce the need to store multiple coverages that have the same basic geographic units, for example, storing both a county coverage and regional planning district coverage of towns when only a geodatabase of towns is necessary.

"A long-term goal (and benefit) of a web-based map library is increasing the awareness within the state of the need for data standards, more metadata descriptions, coordination of production efforts, and proper archiving."

A second limitation of the current system is that the user can only retrieve data from one database in the construction of the attribute table. The U. S. Census' FERRET project is attempting to overcome this problem by allowing users to enter keywords that can be used to search the metadata of many databases for the possible retrieval of data from different databases that can be merged into one table. MAGIC and UCCGIA are working with the U. S. Census to prepare their attribute data sets in the appropriate format for inclusion in the FERRET search engine.

A long-term goal (and benefit) of a web-based map library is increasing the awareness within the state of the need for data standards, more metadata descriptions, coordination of production efforts, and proper archiving. Many agencies lack the capacity for storing historical information and only retain current data. Having a central location such as the MAGIC and CTDATA websites has facilitated regular collection development and satisfied user community needs through standardized access procedures over the Internet.

REFERENCES

Buckland, M., 1992. Redesigning Library Services: A Manifesto. http://sunsite.berkeley.edu/Library/Redesigning/challenge.html

Burrough, P. and R. McDonnell, 1998. *Principles of Geographical Information Systems*. Oxford: Oxford University Press.

Clarke, K., 1995. *Analytical and Computer Cartography*, 2nd Edition. Englewood Cliffs: Prentice Hall.

Marble, D., 1987. The computer and cartography. *The American Cartographer*, 14(2): 101-103.

McGlamery, P., 1989. The librarian's dilemma: A map librarian's access to machine-readable information. *Cartographic Perspectives*, n2:7-13.

Morehouse, S., 1985. ARC/INFO: A geo-relational model for spatial information. *Proceedings of Auto-Carto 7*, 388-397.

Waugh, T. and R. Healey, 1987. The GEOVIEW design. A relational data base approach to geographical data handling. *International Journal of Geographical Information Systems*, 1(2):101-118.

Worboys, M., 1995. *GIS: A Computing Perspective*. London: Taylor and Francis.