INSPEL 36(2002)3, pp. 207-215

GIS AND DATA SHARING IN LIBRARIES: CONSIDERATIONS FOR DIGITAL LIBRARIES*

By Nancy C. Hyland

Introduction

The Cornell University Geospatial Information Repository (CUGIR) is a Web-based repository providing searching, browsing and download access to geospatial data and metadata for New York State. Subjects such as landforms and topography, soils, hydrology, environmental hazards, agricultural activities, and wildlife and natural resource management are appropriate for inclusion in CUGIR. Most data in CUGIR come from government sources.

Cornell University's Mann Library has provided support for Geographic Information Systems since 1994. Demand for government GIS data increased exponentially and it was often difficult for users and librarians to obtain the data. CUGIR was created to fill that need. Since its start in the fall of 1998 at Mann, more than 100,000 geographic datasets have been downloaded. The CUGIR team continues to work with government agencies to provide data to researchers, planners and other government agencies.

What is GIS?

Geographic Information Systems (GIS) consist of hardware, software, and data that can be combined to create a relational database to be used for the retrieval and analysis of any information with a spatial component. Although GIS is sometimes thought of simply as a map-making tool, it is the capacity of a GIS to store and link to information behind a point on a map that makes it so powerful. Many emergency response systems, for example, employ a GIS. When a call is made to the emergency system, a dispatcher, simply by knowing the number from which the call was placed, knows the location of the caller, the fastest travel route to get there, whether any hazardous materials are within a two-block radius of the location, and which emergency workers are closest to the scene. Data can be used from many different sources and organizations in these systems.

---

* Paper presented at the 68th General Conference and Council of the International Federation of Library Associations and Institutions, August, 16 – 24, 2002 in Glasgow

A GIS can also assist in the analysis and understanding of demographic, environmental and agricultural issues. Many government agencies make planning and environmental decisions based upon analyses that include the use of geographic information systems. One might, for example, relate information about the water run-off from agricultural areas near wetlands boundaries in order to tell which areas need stricter pesticide regulations. Governing bodies often encourage citizens to get involved through participation in town hearings and meetings with GIS analysts. These analysts create maps using GIS based upon the variables unique to the particular planning decision. Maps provide a powerful way for users to understand complex sets of data quickly.

Data are created by taking measurements on the ground with a global positioning system survey, by remote sensing using aerial photography or satellite images, and by digitizing existing maps. Once digital base-maps are created, they can then be linked with any numeric or attribute data that has a spatial component. These digital base maps are costly and time-consuming to produce. Not surprisingly countries with more highly developed infrastructures have more highly developed GIS infrastructures and have, therefore, produced more data. However, core datasets, such as digital elevation models, hydrography, and topographic data, exist for nearly the entire world. Without data, a GIS is of no use at all. Data are becoming more widely available but can still be difficult for end-users to locate.

Because of the nature of GIS data, governments have not distributed the data as widely as other paper and text government documents. New distribution models have had to evolve. The United States Federal Geographic Spatial Data Committee (FGDC) has advocated for better distribution of GIS data within the United States and internationally with its National Spatial Data Infrastructure (NSDI). It is a series of data-sharing nodes linked together with a common metadata standard and search interface.[1] The NSDI now consists of over 250 national and international digital nodes. By setting data and metadata standards in the NSDI, the FGDC has made significant contributions towards the globalization of GIS data sharing. Since libraries are already committed to preserving, organizing and providing access to information, they are ideal places for such nodes to reside.

GIS in Libraries

If a library is considering adding GIS support to its services, the first consideration should be the level of staff commitment it can make. There are three general possible modes of GIS services: "bare-bones" GIS services, a physical GIS

---

[1] National Research Council (1994). Promoting the National Spatial Data Infrastructure Through Partnerships. Washington, D. C., National Academy Press.

collection, and the digital library (or clearinghouse node). The following recommendations are based upon Mann Library's eight years of GIS service and support.

The first requires only the time commitment of a public services librarian who can help patrons navigate various data nodes and has a familiarity with free GIS software resources, such as ESRI's ArcExplorer that is available at http://www.esri.com. Any public services librarian with an interest and aptitude should be able to provide a minimum-level GIS service after about twenty hours of work in either a hands-on workshop or through a self-paced tutorial, as well as through learning the types of data available at such sites as:

- The GIS Data Depot http://www.gisdatadepot.com/
- http://www.gisnet.com/notebook/GIS_Resources.htm
- US based data http://www.cast.uark.edu/local/hunt/index.html
- http://www.fgdc.gov/nsdi/nsdi.html
- http://www.gsdi.org/main2.html
- http://www.spatial.maine.edu/~onsrud/GSDI.htm
- http://www.gisdevelopment.net/tutorials/ - http://www.csc.noaa.gov/products/sccoasts/html/gistut.htm

At this minimum level, it is not possible to guide patrons through the entire process of creating a fully-integrated GIS. Rather, the library would serve simply as an intermediary helping the user to find the data and providing only the most basic of support for web-based mapping systems.

Supporting a physical GIS collection requires more monetary and staff resources. In addition to the hardware and software costs, considerable staff time needs to be allotted to cataloging compact disks, and providing public service support and software training, as well as the purchasing of the data disks themselves. A number of articles have appeared in the library literature with information on creating GIS services at this level. I encourage those who are interested in starting a physical GIS data collection to read Dean Jue's "Implementing GIS in the Public Library Arena"[2] and Jurg Buhler's "Digital Maps in Map Collections."[3]

At Mann Library, most of our GIS efforts focus on the third type of library GIS service: a remotely accessible digital collection, or a geolibrary. Providing access to GIS data in a digital library requires staff participation from across the library's

---

[2] Jue, D. K. (1996). Implementing GIS in the Public Library Arena. 1995 Clinic on Library applications of Data Processing, Urbana, IL, University of Illinois.

[3] Buhler, J. (1999). "Digital Maps in Map Collections - Presenting New Electronic Media." LIBER Quarterly, the journal of the European research libraries 9(2): 228-234.

functional departments. CUGIR has been an NSDI node since 1998. At Mann, librarians from technical services, collection development, information technology and public services all participate in this endeavor. The technical services librarian supports metadata services within CUGIR. There are a number of metadata used in a GIS service. At present, data are described with FGDC metadata standards. In the summer of 2002, the FGDC standard will merge with the International Organisation for Standardisation (ISO). CUGIR is in the process of converting our current FGDC metadata to ISO. Our metadata is also converted into xml, sgml, html, DublinCore and MARC to allow for the broadest possible access to the data.[4] In addition another type of metadata has emerged called "service metadata." They describe the capabilities of experimental web-mapping servers that will be able to share data seamlessly. Our information technology team member supports the server on which CUGIR is housed, is designing a new relational database for better access, and provides all programming needed for the web and Z39.50 interface. Collection development assists in refining the preservation and collection policies. Finally, the public services librarian is responsible for end-user support and is the primary contact and negotiator with data partners.

Data Partners and Negotiation

We started CUGIR in partnerships with the New York State Department of Environmental Conservation (NYSDEC) and the Soil Information Systems Laboratory (SISL). Most of the data available at CUGIR are either directly from or derived from government sources. Other local government data have also been added in partnership with local planning offices. We have found these partnerships to be very rewarding, however, there are a number of issues that ought to be negotiated from the beginning in forging partnerships with government agencies.

What's in it for you; what's in it for them?

As with any negotiation, it helps to spell out the benefits each party would receive from the partnership. In a sense, CUGIR functions as a remote intranet for the offices. The GIS data are very large and can consist of hundreds of different files. Some employees within the agencies reported that they previously had to keep data on their hard drives or contact a colleague in another building who happened to have the data on CD. Since the data became available on CUGIR, access to their own data is simpler and faster. Furthermore, most government offices who produce GIS data spend a significant amount of time creating and enhancing the

---

4 Chandler, A. and E. Westbrooks (2002). "Distributing Non-MARC Metadata: The CUGIR Metadata Sharing Project." Library Collections, Acquisitions & Technical Services Forthcoming

data. They are ill equipped or simply do not have the time to respond to public requests to provide the data on CD. Allowing the library to provide access eliminates the need for end-user support related to access.

The nature of GIS data makes it difficult and sometimes impossible to use without adequate metadata. Many GIS professionals find metadata creation to be an onerous task. It not only helps tremendously for the general public who might want to use the data, but may also be the only documentation a worker leaves. Job turnover can translate into a fair amount of duplicated work if metadata are incomplete. Catalogers, on the other hand, are quite adept at describing work and data and take quite easily to the FGDC and other metadata standards. Unlike a traditional catalog record, metadata cannot be created without the input of the data creator. Staff at CUGIR work closely with our data partners and provide training and advice on the creation of metadata. We are also the final metadata editors and ensure compliance with current metadata standards. Our partnerships with government agencies help keep us aware of trends within the GIS community so we are better able to serve our patrons.

Preservation, Versions and Updates

Another issue to consider in negotiations for data is preservation. There was great hope at the advent of digital and online information that preservation worries would be relieved. Once an item is digitized, it is easy to reproduce, and, therefore, there would be less worry about losing a document. Indeed many preservation efforts in the past decade have involved converting documents or other items in analog format into digital. The lifespan of digital documents, however, is far shorter than any counterparts, in part because of media degradation. There is almost a direct inverse correlation between how long it takes to produce a document and the life expectancy of the medium on which is it stored.[5] A clear plan should be made to handle older and outdated versions of the data and needs to be negotiated with the data partners.

Many data are constantly being updated and corrected, which has implications for preservation. When data are updated, it is important to ascertain whether they are an update or a new version. If the data are simply an update, it may only be a correction of previous mistakes. In this case, it may make the most sense simply to destroy the earlier data. Indeed, data producers often want any previous data in a series to be destroyed immediately. One may argue, however, that if a project is

---

[5] Conway, P. (1996). "Preservation in the Digital World." Commission on Preservation & Access, Council on Library and Information Resources: 24pp.

underway or completed, it may be important to get the imperfect data as they are the data of record for a particular time period.

In some cases data are updated because the agency has decided to employ a new projection, coordinate system or datum. These are key geographical concepts employed in GIS. Put simplistically, they are different mathematical models that allow the three dimensions of the earth to appear in two dimensions on screen or on paper. For further explanation see "Datums And Projections: A Brief Guide" at http://biology.usgs.gov/geotech/documents/datum.html.[6] If an end-user is overlaying data from different sources, the three models must match. If, for example, a soil coverage is overlayed with a hydrography coverage and they are in different datums, the soil in the stream bed may appear to be thirty meters from the stream itself.

Once an organization decides to change one of the three models and updates accordingly or has made corrections, they may want previous data destroyed. This may be for legal liability reasons, but we have also discovered that it is often the case that the organizations do not want the end-user to confuse older and current data. If two datasets exist in the digital library with the same name, this can easily happen. Although information about the dates and versions is listed in the metadata, they distrust that users will read it, perhaps justifiably so. In our case, we are creating an archival area. The user will not be able to download the data without being notified of the data's currency. Because we can create an archive, some organizations are willing to allow us to continue access to older data.

It is possible to transfer the older data from one medium to another without making them accessible online. This strategy can work well but needs to be done carefully. If, for example, I were to copy some of the outdated CUGIR data onto CDs, I would not only need to document very carefully what the data are and include all pertinent metadata, but I would also need to record how the data was transferred onto the CD, such as the hardware and software used in the process of copying the data. It would also be necessary to have a detailed plan on when to refresh the data again. Simply putting the CDs in a box and then forgetting about them will mean that the data will eventually be lost. If it is understood by the institution that the data are to be refreshed again in a specified number of years with explicit details recorded, the data are less likely to be lost. Keeping data accessible and online makes data less likely to be forgotten.

---

[6] Brown, K. (1999). "Datums And Projections: A Brief Guide." USGS Center for Biological Informatics. 2002.

Data Access and Ownership

In addition to issues of preservation, several key questions should be spelled out with data partners regarding access and ownership:

1. **Who owns the data and what does it cost?**
   Once the data is available at the geolibrary, does the clearinghouse own the data or does the government agency? This should be determined even if the data is free to the library. Unforeseen events can make question of ownership crucial to continued access.

2. **How may the data be used?**
   As a general policy all data at CUGIR may be used in any application and may be used in a commercial product. This is spelled out in the metadata and does not have to be applied to an entire clearinghouse. If data is made with use restrictions, we have no technological limitations to doing so, and we can make such decisions on a case-by-case basis.

3. **Who may download the data?**
   Many geospatial clearinghouses have mechanisms in place to restrict who may access some or all of their data. At CUGIR, there are no purposeful restrictions on downloaded data. Some government agencies decided not to add their data to CUGIR because of this. This should be made clear at the beginning of the negotiation.

At CUGIR we deliberately wanted all data to be accessible without restrictions. We were, therefore, not able to add some specific geospatial data to CUGIR, but it did allow us to provide a fast and easy-to-use interface. It seemed to us that CUGIR was providing access to data that had no significant risk and wide distribution of the data was for the public good. In February of 2002, the New York State Office of Public Security requested that CUGIR shut down completely until a review of the data was conducted assessing any risk to national security. Any reference to bridges or airports could be considered a threat. CUGIR conducted a risk assessment of the data available and kept the clearinghouse open despite the request.[7] We did inform our data partners of our decision and one partner did request that we remove their data until they could conduct their own risk assessment and we complied with this request.

[7] Knezo, G. J. (2002). "Possible Impacts of Major Counter Terrorism Security Actions on Research, Development and Higher Education." Congressional Research Service.

Risk Assessment

We did comply with part of the request issued to us in February and reviewed all data available at CUGIR. Two members of the Mann Library staff conducted a risk assessment of all CUGIR data. They formulated two main criteria for the assessment: risk level and availability. For risk level, they looked at each dataset thinking of any possible use criminals and terrorists could possibly do with the information. The second criterion dealt with the availability of the same data from other sources. It was determined that distribution was an important factor because even if one might consider the location of an airport to be a significant security risk, if that information is available in thousands of other places, it cannot be classified as a security risk in a practical way. In the rankings, data were given rankings of 1 through 5, with a 1 being the lowest level of risk and 5 indicating the highest.[8] The analysis showed that our only data that posed any kind of risk are widely distributed in other formats. The method for our analysis was based on preservation risk assessments that Mann Library has conducted on numeric data that we provide in cooperation with the United States Department of Agriculture.[9] Our review was shared with the agency that had removed all of their data. At the time of this writing, nearly all data are restored to CUGIR, with the exception of three series. Such risk assessments are recommended for geolibrary data before a problem arises. Every data will have unique characteristics that should be evaluated individually.

Another risk factor involved with geospatial data concerns civil-legal liability. With the amount of information and high degree of detail, mistakes are nearly inevitable on the part of the data producers. Many agencies are loath to provide public access for fear that an error may result in a lawsuit. All data within CUGIR have a liability disclaimer stating among other things, that "the burden for determining fitness for use lies entirely with the user." Providing such disclaimers should be standard practice in any digital library.

Conclusion

The nature of GIS data requires some flexibility to be built into the digital library. Carefully planned partnerships with data producers and working to accommodate

---

8 Martindale, J. (2002). National Security and Access to GIS Data via the Internet: Cornell University Geospatial Information Repository (CUGIR). ESRI Education User Conference, San Diego, CA, Forthcoming.

9 Lawrence, G., W. Kehoe, et al. (1999). Risk Management of Digital Information: A File Format Investigation: A Report Prepared for the Council of Library and Information Resources. Ithaca, NY, Cornell University: 1-19.

their requests make the process run as smoothly as possible. Libraries are ideally suited for providing long-term, user-friendly access to the data. Partnering with government organizations provides benefits both for the library and for the organization. With careful planning, we at Mann Library hope to continue to provide a quality service at CUGIR for years to come.

Nancy C. Hyland
Coordinator, Information Services
Albert R. Mann Library
Cornell University
Ithaca, NY 14853-4301
USA
nch9@cornell.edu