Geospatial Web Services and Geoarchiving: New Opportunities and Challenges in Geographic Information Services

Steven P. Morris

Abstract

Over the course of the past fifteen years the role of Geographic Information Systems (GIS) has changed significantly. Initially the role of the map library was confined to that of building and providing access to collections of hard copy maps and imagery. Later, digital data, whether on CD-ROMs or network based, was added as a new type of resource within that collection and service model. By the late 1990s some academic libraries began to take on a Web map server role, providing interactive Web mapping access to collections of digital geospatial data. In the new era of distributed, interoperable map services, libraries will have an opportunity to explore new roles as portals to streaming content available in the form of geospatial Web services. At the same time, the increasingly ephemeral nature of digital geospatial content will make even more critical the need to address the long-term digital preservation challenges that are facing geospatial content.

This article focuses on two major geographic information issues facing academic libraries as well as libraries in general. First, what role should libraries play in the development and utilization of emerging geospatial Web services? Second, how should libraries address the challenge of long-term preservation of digital geospatial data in light of a shift to distribution methods that make the content ever more ephemeral?

Introduction

Over the course of the past fifteen years the role of Geographic Information Systems (GIS) has changed significantly. Initially the role of the map library was confined to that of building and providing access to collections

LIBRARY TRENDS, Vol. 55, No. 2, Fall 2006 ("Geographic Information Systems and Libraries," edited by Jaime Stoltenberg and Abraham Parrish), pp. 285–303

of hard copy maps and imagery. Later, digital data, whether on CD-ROMs or network based, was added as a new type of resource within that collection and service model (Journal of Academic Librarianship, 1995, 1997). By the late 1990s some academic libraries began to take on a Web map server role, providing interactive Web mapping access to collections of digital geospatial data. In the new era of distributed, interoperable map services, libraries will have an opportunity to explore new roles as portals to streaming content available in the form of geospatial Web services. At the same time, the increasingly ephemeral nature of digital geospatial content will make even more critical the need to address the long-term digital preservation challenges that are facing geospatial content.

This article will focus on two major geographic information issues facing academic libraries as well as libraries in general. First, what role should libraries play in the development and utilization of emerging geospatial Web services? Second, how should libraries address the challenge of long-term preservation of digital geospatial data in light of a shift to distribution methods that make the content ever more ephemeral? Specific experiences with engaging geospatial Web services and with instituting preservation-focused action responses will be drawn from the North Carolina State University (NCSU) Libraries data services program and the North Carolina Geospatial Data Archiving Project, a cooperative effort with the Library of Congress and the NC OneMap Initiative.

Brief Overview of Digital Geospatial Data Services in Academic Libraries

There are many components of academic library GIS services. At the core is the data collection, but accompanying the data is a mix of services that vary from campus to campus. A brief summary of typical service components follows.

Data Collections

Libraries acquire, license, catalog, make discoverable, archive, and carry out value-added processing on digital geospatial data. While, in the United States at least, much data is available in the public domain, the data is not always organized or readily accessible in such a way as to allow the user to easily sort through the wide range of data options available, and effort is required to make such freely available data discoverable. Furthermore, in order to improve data availability it is sometimes necessary to acquire and license additional commercial or fee-based government data for use. In some cases libraries also engage in large-scale value-added work—retiling, projecting, or otherwise converting and reorganizing data resources into a more convenient form for the libraries' target audience.

Data Discovery Tools and Support

Libraries support the discovery, selection, and use of geospatial data. While the most common form of promoting access to data collections has been the development of Web documentation for data collections, in some cases searchable databases of geospatial metadata are also made available. Data resources may also be included in the library's catalog, but the catalog is not usually the most effective vehicle for exposing or searching for digital geospatial data.

Technical Support

The line between providing reference support for finding and selecting data and providing actual technical support for using the data is a blurry one, and it has become more common for academic libraries to play a prominent role in providing technical support to GIS users. At NCSU, for example, the library holds one of four "right to call" spots for the campus Environmental Systems Research Institute (ESRI) site license and provides technical support as needed to campus users. Libraries also play varying roles in supporting campus software licenses, facilitating distribution of software, and troubleshooting installations.

Workshops and Training

As an extension of reference and technical support many libraries offer workshops on a variety of topics such as introductory GIS, data discovery, or use of specific software tools. The mix of workshops offered often reflects the sort of reference and technical support demands placed on the library. Increasingly, in-library workshops have now been complemented by and even supplanted by online training resources. At NCSU, for example, the library supports over 600 registrations per year for the ESRI Virtual Campus online courses.

Marketing and Outreach

Another academic library function, which goes hand-in-hand with workshops, is marketing and outreach—promoting geospatial resources and services to the campus community. GIS activity typically initially takes root in one or just a few core departments where there is a high level of activity and support. Meanwhile, latent demand exists in a broad range of academic disciplines where awareness of geospatial tools and resources is lacking, or where there is a perceived barrier to entry in terms of lack of access to tools, data, training, and support. Libraries, as a neutral space focused on customer service, are well positioned to cultivate new GIS users by promoting the use of geospatial tools and content and by providing ready access to software, data, training, and support. At NCSU the number of academic departments engaged in GIS grew from fewer than ten to thirty-five within just a few years as a result of combined campus and library efforts to develop a campus GIS infrastructure.

Evolution of Technical Approaches to Delivering Geospatial Data

The manner in which libraries have provided access to geospatial information has changed significantly in recent years, with analog map and image offerings increasingly being supplemented by or replaced by digital resources. At NCSU campus-wide networked access to data was initiated in 1993, with data made available both for download and for use online from GIS workstations in a networked environment. By 2000 one began to see more libraries offering Web mapping services, making the GIS content available to a much broader audience, including those who otherwise lacked the skills, software, and data access ordinarily needed to utilize GIS content.

The Early Library Experience with Web Mapping

While the Web mapping approach was initially fruitful—and still is in some contexts—these library-based map servers have increasingly risked becoming liabilities to the extent that volatile state and local content is included. State and local agency data producers are typically better positioned to manage data updates, and the number of available state and local map servers has risen steadily since 2000. In North Carolina, for example, the number of county map servers increased from 15 in 2000 to 77 out of 100 counties in 2005 (NCSU Libraries, 2006a). User demand for county and city data is high because it is larger scale, more detailed, more current, and more accurate than state and federal alternatives. Furthermore, many resources, such as cadastral data, zoning, and building footprints, tend to be available only at the local level. Meeting real user demand for data has increasingly required that local content be made available, yet the rate of update of that data has made it increasingly unfeasible to integrate and successfully update such content within library-based Web mapping services. The existence of stale data hosted on library servers, coupled with concern some data producers have about liability issues, have made the library Web map servers an increasingly untenable option. At NCSU Web map services, which began in 1997, were ceased in 2001 in deference to emerging state and local map services.

Data Interoperability and Emerging Geospatial Web Services

By the year 2000 producer-operated map servers were proliferating, but these emerging federal, state, and local map servers remained data islands that could be viewed only in isolation from one another. There was no way to zoom in and see federal, state, and local content together for a particular location. There was also no easy way to view adjacent county or municipal services in a side-by-side manner. Social, environmental, and economic processes did not stop at county borders, but local map services did.

Around the same time, however, the various initiatives of the Open Geospatial Consortium (OGC) began to bear fruit and some key initial

steps toward data interoperability were made. Data interoperability is necessary to integrate disparate data resources; allow sharing of content; allow interoperability between resources in different formats, commercial software environments, and coordinate systems; and facilitate service chaining (Reichardt, 2005). A key initial OGC specification, the Web Map Server (WMS) 1.0 specification, was adopted in 2000 (OGC, 2004), and activities related to the Web Mapping Testbed led to a subsequent explosion in the development of WMS services (Doyle, 2000). Initiatives such as the National Map, at the national level, and NC OneMap, at the state level, helped to further the integration of federal, state, and—increasingly—local map services in a flexible interoperable environment. In North Carolina, for example, by virtue of extensive outreach carried out by the NC Center for Geographic Information and Analysis and the U.S. Geological Survey with their partners, the number of state and local WMS services in the state grew from two in 2002 to seventy-four in February 2006 (NCSU Libraries, 2006a; NCGICC, 2006a). As a standalone system as well as a component of the National Map, NC OneMap provides services in the context of statewide needs while also feeding content directly into the National Map system (NCGICC, 2003).

The rapid growth in availability of geospatial Web services has been followed by the development of new services focused on geospatial Web service discovery and integration. Initial work in ESRI's Geography Network, available from 2000, was followed by development of the National Map Catalog and later Geospatial One-Stop. At the same time, commercial geospatial Web services also began to proliferate, with offerings such as ArcWeb Services from ESRI and other commercial services from firms such as TopoZone. While such commercial mapping services initially took the form of noninteroperable Web mapping services, it has increasingly been the case that these offerings are interoperable services based on OGC interoperability specifications, Simple Object Access Protocol (SOAP), or Application Programming Interfaces (API) that support application integration.

The Attraction of Geospatial Web Services

Geospatial Web services are potentially attractive to libraries and their users for a number of reasons:

- The services are available in a time and location independent manner.
- Access to extremely large datasets is possible even over low bandwidths.
- The most current data is readily available and data update does not require local maintenance action by libraries or other intermediate information providers.

- Differences in native formats and coordinate systems can become less of a barrier to use.
- Access to data can be more efficiently offered for regions where demand does not merit static data purchase.

Map services can be used in a broad range of situations, from complex projects involving application fusion to rather basic one-off uses. For example, one of the common uses of paper maps seen in the map library is that of tracking down the locations and coordinates of specific places on USGS topographic quad sheets. Since the late 1990s there have been a variety of commercial and public domain servers that allow users to examine topographic maps online, identify coordinates, and make annotations. There is no question that examining a topographic map by holding the large-format analog copy is preferable, from an ergonomic or aesthetic perspective, to looking at a smaller map area on the computer screen, or that many map analysis tasks can be more effectively carried out using large-format analog maps and images. Yet when one factors in issues such as convenience of access, travel time to the library, expanded resource availability, and other factors, the Web-based option becomes an attractive alternative for many map uses.

Drawbacks of Geospatial Web Services

There are of course many drawbacks to utilizing geospatial Web services or relying on them as a core information resource within geographic information services, including the following:

- Application performance when using Web services will frequently not match that which can be achieved using locally loaded data.
- Uptime reliability can be a problem, lead to service chain failures, and threaten project work.
- Some services are of a demonstration nature and can disappear without notice.
- While the content underlying Web services might be updated with some frequency, some applications may have a need to rely upon static, snapshot content for consistency in results and analysis.
- Screen-generated maps are aesthetically and ergonomically no match for large-format analog maps and images.

Geospatial Web services are clearly more useful in some situations than others, depending upon application and user requirements. These services are probably most useful when

- the user needs the most current data;
- the data is subject to frequent change;
- the user needs to make use of extremely large datasets, perhaps over lower bandwidth connections;

- the user wishes to preview the data prior to acquiring it;
- the user just needs the data for background use;
- the data needs to be integrated into remote or portable devices;
- the data is not otherwise available or cannot be efficiently acquired and stored for local use.

Integrating Geospatial Web Services into the Library Environment

Awareness of and promotion of these emerging Web services are still rather low both on the part of end-users and on the part of academic libraries. Integrating and managing access to services presents some problems that are very different from those associated with locally hosted content, including the following:

- Geospatial Web services have been difficult to discover and select from.
- In the case of commercial services, sustainable licensing models that work on a campus scale have yet to be worked out to satisfaction (problems include allowing for the volume of requests related to simple operations such as pan zoom, the ability to restrict access to authorized users, and anticipating an unknown volume of requests).
- Linking data resources with services that act upon them has been a sticky issue, with metadata standards and practices not adequately addressing the linkage of data resources with services that act upon them.
- Rights issues and approved use are in many cases ambiguous, with Web services in something of a "Wild West stage" (for example, it is not clear whether it is acceptable to extract data from ArcIMS services through ArcGIS connections; this is technically possible but not typically an intended use of the service).
- Integrating Web services into the physical browsing environment of the map collection in order to stimulate awareness of these new resources is tricky given the transient nature of such services.

With regard to the issue of physical browsing, while libraries have become increasingly if not overwhelmingly digital, the map room still provides a rich and effective browsing environment. While volatile resources such as Web services do not lend themselves easily to hard-coded representation on shelving or in map cases, emerging mobile device technologies might, in time, make it more feasible to integrate discovery and use of these resources within the context of the physical browsing environment.

Possible Library Roles vis-à-vis Geospatial Web Services

So what might academic libraries do to promote and facilitate access to geospatial Web services? Some possible roles might include facilitating discovery of services; producing new map services to fill the gaps in service availability; building new map portal services on top of existing map ser-

vices; licensing commercial Web services for use; and utilizing Web services consumption data to inform collection development planning.

Facilitating Discovery and Selection Libraries can support user discovery and selection of resources by incorporating such services into catalogs, GIS data collections, and the physical map room browsing environment. Just as libraries provide support in user selection of maps or datasets, support can also be offered in selecting from among competing service options. The notion of the reference interview, as it applies to geospatial data, can be extended to geospatial Web services.

The more traditional geospatial data-focused reference interview will tend to focus on content issues, a *subset* of which might include the following questions:

- Data extent: Does the data cover the study area as required?
- Thematic content: Does that street dataset have street centerlines or curbs and gutters?
- Attribute availability: Are there street addresses? Are they complete across the entire dataset? Is the format friendly to geocoding processes?
- Currency: How recently was the data produced? What real world time period does it represent? How concurrent is it with other data to be used in the project?
- Format: Is the data in a vector format that the project's software can support or at least convert without unacceptable data loss? In the case of imagery, has a level of compression been used that entails unacceptable data loss?
- Openness of licensing: Can the data be copied off of the CD-ROM or server? Can maps created from the data be used in publication? Can the data be used in a Web mapping application? Can a value-added derivative of the data be redistributed?
- Ease of access: Can the data be downloaded right now? In the case of very large datasets, is it possible to connect directly to the resources and use the data across the network? Is it possible to extract data for extremely large areas, or must one make numerous much smaller extractions to assemble data for the study area?
- Coordinate system, datums, etc.: Will it be necessary to re-project the data? Will a datum conversion be necessary? Is this information even recorded in the metadata?

In the Web service context, some of the content facets, such as format, can become less important, while some additional service or "functional" metadata come into play. These facets might include the following:

- Type of service: Image service, feature service, geocoding service, etc.
- Access protocol: ArcIMS image service, ArcIMS feature service, WMS, WFS (Web Feature Service), SOAP, and other methods such as the

Google Maps API. Is the service exposed through a protocol that is compatible with the user's technical environment?
- Reliability and uptime: Will downtime impact project work or service chaining? Is this a demonstration service that is liable to disappear at an inconvenient moment?
- Licensing or pricing scheme: How will trivial transactions such as pan and zoom count against overall service consumption costs? Can licensing effectively be extended to multiple, concurrent users within a constrained domain of authorized and authenticated users?
- In the case of image services, what image formats are offered (GIF, JPEG, PNG, etc.)?

Service discovery is available through the National Map catalog, Geospatial One-Stop (GOS), and regional services such as NC OneMap, but exhaustive, comprehensive access is still not available. The National Map Catalog covers a subset of the services available in GOS, and GOS covers a subset of all available services. The National Map Catalog exposes an API for application developers, raising the possibility of drawing from these service metadata collections to develop specific local catalogs (USGS, 2005). Other more extensive service catalogs are being developed, including the Naval Research Laboratory (NRL) GIDB Portal, which lists nearly 1,400 map servers and over 300,000 individual data layers (Naval Research Laboratory, 2006a), and Mapdex, which lists over 1,700 servers (Mapdex, 2006). The NRL is working on a searchable catalog system that will be compliant with the OGC Catalog Services Specification and will provide the capability to browse, search, and query using any OGC Catalog client application (Naval Research Laboratory, 2006b).

Providing Map Services Another possible library role lies in the area of helping to fill the holes in map service availability by, for example, serving up WMS layers that are not otherwise available. Rather than risk providing stale data that are better provided by the data producers, libraries might focus on serving out specific strategic content that users and other services could choose to consume. NCSU Libraries, for example, is deploying census data map services that will be integrated with the NC OneMap environment, helping to plug a hole in data availability within the statewide framework. Libraries, by virtue of their mission, might be more predisposed than other organizations to serve out lower-demand older or archival content that is not served up by data producers, who may tend to focus on the highest-demand, most current data.

Map Portals and Cascading Map Services Libraries may also have a role to play in deployment of the next-generation version of the old map server: setting up map servers that draw from and build on top of multiple existing map services, thereby creating single map interfaces. The USGS National Map at the national level and NC OneMap at the state level are two promi-

nent examples of cascading map services. In general, one of the things libraries try to do is build windows to the world of information where the window is orientated in a way that best suits the library's client base, often resulting in a particular geographic focus. In the case of map services, this notion might be translated into building specialized views that integrate existing map services.

In practice there are many complicated issues involved with setting up cascading map services: services adhere to different versions of OGC standards, use different symbolization, apply different scale restrictions, and name their data in different ways (for example, land parcels versus cadastral or property boundaries). Also, metadata that is needed to properly integrate resources may be missing, and rights issues concerning services are often ambiguous.

Building an effective cascading map service often becomes an exercise in community building that the technical interoperability specifications do not themselves address. Service builders must work closely with data providers to standardize service characteristics such as symbolization, classification schemes, scale thresholds, and layer naming. The relevant community must agree to and promote a set of practices that go beyond whatever requirements the actual standards or specifications might impose, as has been illustrated in the NC OneMap experience of developing a statewide integrated set of services (NCGICC, 2006c). The reality is that federal and state agencies participating in spatial data infrastructure are usually better positioned to carry out the community-building process.

Licensing Commercial Web Services Another opportunity for libraries lies in the area of licensing fee-based services for use by patrons. Such services may offer more than just content, with functions such as geocoding and routing being offered by emerging commercial services. Key challenges lie in the area of working out effective licensing models and in integrating campus identification and authorization schemes with these commercial products.

Using Web Services Consumption as a Measure Demand Another possible use of geospatial Web services is in the measurement of data demand associated with a library's user market. Development of digital geospatial data collections that fit the spatial demand footprint of the library's audience can be a challenging task. Funds are limited, and only so much data covering so much territory can be acquired and managed. To the extent that content exists, user demand can be measured based on data downloads by region, but if data holdings do not exist for given areas then demand cannot easily be assessed. It might be possible to carry out more rigorous market analysis if, for example, libraries were able to obtain zoom-in density maps from aggregated data reflecting their institution's traffic on national portals such as Geospatial One-Stop and the National Map.

Web Mashups, Geo-Hacking, and the New Geospatial Frontier

Geospatial Web services in the form of Web map servers, OGC services, and SOAP services have grown rapidly in the past five years, but these services have for the most part only been exposed to a confined market of geospatial data users. Geospatial Web services are entering a new phase of wider use with the availability of new, more mainstream services such as Google Maps, Google Earth, MSN Virtual Earth, and Yahoo Maps. APIs have made it possible for third-party developers and the general public to build applications on top of these offerings, which are more accessible than traditional geospatial industry offerings. These new services have experienced rapid growth in use since their inception in early 2005, with a vast new audience of "geo hackers" without traditional GIS backgrounds beginning to work with geospatial content and creating "web mashups" or "map mashups," which integrate content from multiple, distributed environments using AJAX (Asynchronous Javascript and XML) and other technologies ("Mashing the Web," 2005). The explosion during 2005 of creative activity on top of these services is likely to be just the beginning of a revolution in how geospatial content is used and republished.

The geospatial content available in these environments is still limited, with only a very, very small slice of all available geospatial content exposed for use with these systems; however, holes are being poked and then widened through the walls that separate the new commercial Web mapping realm from the much more content-rich traditional GIS realm as developers create tools that integrate WMS or WFS services with Google Maps (Flood, 2005; Mulka, 2005) or convert traditional geospatial data to Keyhole Markup language (KML) for integration with Google Earth (Martin, 2005). The new mainstream mapping space has a very large audience, and yet only a relatively small proportion of available data is exposed to these environments. Meanwhile, in the traditional geospatial industry space there is a relatively small audience and a very large amount of data available. As these two information spaces begin to connect and merge, a number of new opportunities are likely to emerge for libraries.

One very immediate impact of Google Maps, Google Earth, and the like will be the creation of a much larger audience and market for geospatial information resources. While those doing Web mashups are often commercial information technology developers, they are also often members of the general public developing maps for their churches, schools, or community groups. Mainstream developers crossing over to geospatial systems, while initially naïve on the topic of data quality, are developing a more sophisticated understanding of the qualitative differences between data alternatives and are seeking guidance from others in the selection of data sources for integration. One opportunity for libraries will be in the area

of exposing archived content to the Web mashup environment for before-and-after and time-related uses, as the emerging services currently focus, for the most part, on delivery of only the most current data.

Digital Preservation Challenges in the Web Services Era

While the emergence of geospatial Web services has opened up a number of opportunities for libraries, a significant threat is also posed. History has shown that it is quite often secondary archives that preserve content over long periods of time rather than the original content producers. For example, libraries typically preserve books rather than publishers. Until recently, in order to provide efficient access to content it has been necessary to physically acquire the data in order to make it available to users. As a result, data archives have often evolved as a somewhat accidental by-product of the process of providing access. With the emergence of Web services it will be much easier to point to the data source and avoid handling data and storage media altogether. This is convenient for the user and eases the burden on the library, but who then archives and preserves the data? If preservation of digitally born resources was already a problem before the advent of Web services, the shift to new distribution efforts will require an even more focused and intentional effort on the part of libraries to preserve data.

Many GIS professionals will readily admit that retention of older content is often very low on the list of priorities. "Kill and fill" is often the operating archival strategy. In the early years of GIS it may have made more sense to ignore the temporal component of geospatial data resources: there was little older content so time series analysis was out of the question anyway, barring massive and expensive vector digitizing of old maps. Most GIS projects are focused on problems that require the use of the most current data. Issues of convenience also undermine demand for older content: the fact that students, during their formative training, will tend to build class projects around available data perhaps reinforces the inclination to focus on more current content and topics.

Yet there is increasing evidence of a rise in demand for older content and of interest in doing associated temporal analysis. GIS has been in use for decades now, so users—especially younger users—are starting to expect that older content will exist. More projects are focusing on time series components—looking backwards at land use change and looking forward at business trends, for example. As GIS becomes more of a core enterprise resource at the local levels, the stakes are raised vis-à-vis accountability for the disposition of taxpayer-funded data development work.

Early Geoarchiving Efforts at NCSU

Geoarchiving is one term that has been used to describe the problem of preserving digital geospatial content (Maine GeoArchives, 2004). In 2000

the coincidence of emerging local agency data, rising user demand for that data, and a growing sense of long-term risk to data sparked an NCSU project targeting county and city data for acquisition and archiving. One learning outcome of that project was a deeper understanding of the complexity of the process of identifying data resource availability across many counties and municipalities. Another learning outcome was an awareness that more efficient and effective data management processes were needed. It was surprisingly easy to turn on the "fire hose" that sent torrents of data into the library collection, but the "plumbing" to deal with all of the content that could be acquired needed to be developed.

The Need for an Infrastructure-Based Approach to Preservation

This early archiving effort made it clear that a statewide infrastructure-based approach was required, one that would build from existing geospatial data infrastructures that were evolving under the auspices of the National Spatial Data Infrastructure (NSDI), Federal Geographic Data Committee, and Geospatial One-Stop. Two key developments in 2003 helped push NCSU Libraries' preservation effort to the next level: the NC OneMap Initiative and the National Digital Information Infrastructure and Preservation Program (NDIIPP), with its Cooperative Partnership Program.

NC OneMap Initiative In February 2003 the NC OneMap initiative was announced (NCGICC, 2003). NC OneMap is a combined state, federal, and local initiative that is focused on allowing users to view geographic data seamlessly across North Carolina; search for and download data for use on their own GIS; view and query metadata; and determine who has what data through an online data inventory (NCGICC, 2006b). Included in the NC OneMap vision statement was the assertion that "Historic and temporal data will be maintained and available" (NCGICC, 2003). While primarily focused on access and content standardization, NC OneMap has offered a scalable framework by which the 100 counties and many municipalities in the state might be engaged in the problem of preservation.

NDIIPP In August 2003 the Library of Congress put out a call for proposals in connection with a new congressionally funded initiative focused on preservation of digitally born content: the National Digital Information Infrastructure and Preservation Program. In this first funding round of the program, entitled "Building a Network of Partners: Collaborative Collection Development," the Library of Congress sought to engage with a diverse set of partners in a "dual effort to identify, get, and sustain significant material while also collaborating with the Library and the other partners to advance digital preservation methods and best practices" (Library of Congress, 2003). The eight selected projects address a range of content types, including Web pages, numeric social sciences data, business records, and cultural heritage resources (Library of Congress, 2006). One of the NDIIPP cooperative projects is the NC Geospatial Data Archiving Project

(NCGDAP), a partnership between NCSU Libraries, the NC Center for Geographic Information and Analysis, and NC OneMap (NCSU Libraries, 2006b).

North Carolina Geospatial Data Archiving Project

NCGDAP is focused on collection and preservation of digital geospatial data resources from state and local government agencies in North Carolina. The objectives of NCGDAP include

- identification of available resources through the NC OneMap data inventory;
- acquisition of at-risk geospatial data, including static data such as digital orthophotos as well time series data such as local land records and assessment data;
- development of a digital repository architecture for geospatial data using open source software tools such as Dspace;
- enhancement of existing geospatial metadata with additional preservation metadata using Metadata Encoding and Transmission Standard (METS) records as wrappers;
- investigation of automated identification and capture of data resources using emerging Open Geospatial Consortium specifications for client interaction with data on remote servers;
- development of a model for data archiving and time series development.

The project is operating under a three-year timeline from late 2004 to late 2007. Since the project is set within the context of an emerging Web services framework—NC OneMap and the National Map—the project is especially focused on responding to evolving data distribution methods and engaging emerging geospatial Web services in the archive development process.

Geoarchiving Challenges

Although the Web services aspects of the preservation problem are the focus of discussion here, a few salient issues related to the challenge of long-term preservation of digital geospatial data should be highlighted.

Geospatial Data Formats The absence of reliable, open vector formats is a stumbling block to preservation. SDTS (Spatial Data Transfer Standard), while open, has proven problematic and is not in wide use. The initial plan of the NCGDAP project involves retention of the data objects in the format received, while also exporting the content into a safer commercial vector format and buying time until a reliable, open alternative emerges. It is considered preferable to retain the content in a widely understood and supported commercial format rather than to rely solely on a migration of the content to an open format that may not be widely supported and con-

version to which may involve subjecting the content to some unfortunate transformations and data loss.

One thread of investigation involves the use of Geography Markup Language (GML) in an archival capacity. The challenge with this approach is that GML is not really a format per se but rather a means to define something akin to formats in the form of GML profiles and GML application schemas (Lake, 2005). The emerging Simple Features Profile for GML provides a potential solution in the form of a widely supported GML profile that is more sustainable over time, though quality and functionality tradeoffs against industry-specific GML application schemas will be a consideration (OGC, 2005a). NCGDAP will be participating in a broader effort by the National Archives and Records Administration (NARA) and the Federal Geographic Data Committee (FGDC) Historical Data Working Group to investigate the role of GML in preservation (FGDC, 2006).

Another area of investigation concerns mining of data inventories, possibly using the emerging RAMONA system being developed by the National States Geographic Information Council (Indiana GIC, 2006), to detect format "doppler signals" using a format's loss of market share as a possible indicator of format risk.

Geospatial Databases Another problem is the widespread emergence of complex spatial databases, either of the commercial variety or in more open varieties such as PostGIS-based systems. A spatial database stores geographic features and attributes as objects hosted inside a relational database management system. Multiple data layers may be stored in a single database, which may also host elements such as topology, relationships, behaviors, and annotations that are not exportable to conventional vector file formats. Within the project domain, the ESRI Geodatabase format is a prominent example of this approach to data management. Until recently spatial databases were relatively rare in the project domain, but local agencies—especially municipalities—are increasingly turning to the ESRI Geodatabase format in particular to manage geospatial data (NCGICC, 2004).

Preserving Cartographic Representation The true counterpart to the old, preserved map is not the current GIS dataset but rather the cartographic representation that builds on that data. The representation is the result of a collection of intellectual choices and application of current methods with regard to symbolization, classification, data modeling, and annotation. One goal of capturing cartographic representation will be to preserve data in the form that decision makers and others encountered and interpreted it. Another goal, in the case of image capture approaches, would be to provide a stable, preservation-friendly—though "dumbed down"—alternative in the case of long-term failure in the vector data preservation process. The derived image might also serve as a content preview, helping future researchers decide whether to commit time and resources to do whatever "digital archeology" (Ross & Gow, 1999) might be necessary to resurrect

the underlying content. Any preservation of cartographic representation should, ideally, occur in addition to preserving the underlying data.

In the Web services context, one issue to consider is that decisions will increasingly be made on the basis of ephemeral maps created online, making it difficult to document the basis for decisions. The OGC Web Map Context specification addresses the issue of saving the application state in order to re-create maps but does not address the issue of saving data state (OGC, 2005b).

Time-Versioned Content Many of the vector data layers to be acquired are subject to frequent update. County cadastral (land parcel) datasets, for example, are typically updated on a daily or weekly basis. Such time-versioned content, if preserved, can form the basis of time series analyses such as land use change analysis.

Version-handling over time, however, can be quite difficult to manage within the archive. And experience in the content domain has shown that some resources of only a few years of age have already lived in two or three repository environments, so any single repository cannot be expected to have all of the versions.

Content Packaging One of the points of frustration in working with geospatial content in a library context has been the absence of a packaging or bundling scheme for data. Geospatial data is characterized by complex multifile formats that need to be tied together, bundling data with associated metadata and ancillary documentation. Content packaging mechanisms may be used to bundle different versions of the dataset (by format, coordinate system, tiling scheme, etc.), to attach rights information and licensing, to supplement FGDC metadata with additional technical and administrative metadata, and to link objects with services that act upon them. The NCGDAP project will experiment with the use of the Metadata Encoding and Transmission Standard, a technology that has emerged in the library community, as a data bundling scheme. Other packaging schemes, such as the MPEG 21 Digital Item Declaration Language, are being considered in connection with the OGC Geo Digital Rights Management (GeoDRM) initiative (OGC, 2006).

Other Geoarchiving Challenges Other preservation challenges include securing and adequately defining archival and use rights for content; preserving semantic information associated with datasets; providing long-term support of coordinate systems and datums; and maintaining the independence of the preserved content from any particular repository software environment.

Putting Web Services to Work in Geoarchiving

While the shift toward Web services–based distribution of geospatial data may pose a threat to long-term preservation of content, it is also possible that those same geospatial Web services might in the future aid in the

onerous process of developing archives on the basis of widely distributed sources. Taking the example of North Carolina alone, there are 100 counties and over 140 municipalities. Nearly all North Carolina counties have GIS systems, as do many municipalities. Keeping track of data availability across this many agencies is not a trivial problem. Even more problematic is the task of routinely harvesting content from such a diversity of agencies.

In this context Web services become interesting from the point of view of automating inventory creation and automating extract and transfer of content. One of the difficult selling points for digital preservation has been the level of effort that must be applied to solve a problem that is very low on the list of priorities of data producers. If the process of archive development can be automated using Web services, then the barrier to participation in the preservation process might be lowered considerably.

Unfortunately, currently deployed services based on OGC specifications are not really fashioned to the needs of archive development processes. In terms of data transfer, WMS involves transfer of "dumb" images with the data intelligence removed. Web Feature Service (WFS), which involves transfer of the actual data as GML, is perhaps not really optimized for full-scale transfer of entire datasets or databases (OGC, 2005c). Furthermore, WFS is not yet widely deployed. What is lacking, so far, is a sort of rsync-like layer in the spatial data infrastructure that allows for efficient, full-scale replication of data resources while also being informed by data update processes, rights arrangements, and metadata. In cases where delta files—or change files—are used as a means of transferring database changes across the network, archival processes will need to handle conflation of the delta files with the archived database and certify that no delta files have been missed.

Conclusion

Geospatial Web services, which may be image services, feature services, geocoding services, or offer other functionality, are clearly on the rise. These new, dynamic resources are more useful for some applications than others, where access to static resources will continue to be more suitable. These services are notably difficult to discover, creating opportunities for libraries in the area of facilitating discovery of and access to them. The rise of more mainstream map services such as Google Maps, through its API, appears to be leading toward a rapid growth in the use of geospatial data and services by a broader audience.

At the same time, digital geospatial data is becoming increasingly ephemeral. The challenges in preserving static geospatial data are already daunting as we face the issue of preserving proprietary formats and spatial databases, capturing time series snapshots, and preserving cartographic representation. The advent of geospatial Web services raises additional challenges to data preservation, as static files are replaced by dynamic, changing services. At the same time, new Web services technologies may offer some possibility

of making the process of archive development more efficient through the use of automated approaches.

References

Doyle, A. (2000). *Web mapping testbed.* Retrieved February 24, 2006, from http://www.intl-interfaces.com/wmt2/wms.html.

Federal Geographic Data Committee (FGDC). (2006). *Federal Geographic Data Committee (FGDC) Historical Data Working Group.* Retrieved February 24, 2006, from http://www.fgdc.gov/participation/working-groups-subcommittees/working-groups.

Flood, B. (2005) *Google Maps API and WMS servers.* Retrieved February 24, 2006, from http://www.spatialdatalogic.com/cs/blogs/brian_flood/archive/2005/07/11/39.aspx.

Indiana Geographic Information Council (GIC). (2006). *Beta RAMONA GIS inventory.* Retrieved February 24, 2006, from http://in.gisinventory.net.

Journal of Academic Librarianship. (1995). *21*(4).

Journal of Academic Librarianship. (1997). *23*(6).

Lake, R. (2005). *Application schemas help build the Geo-Web.* GeoWorld. Retrieved February 24, 2006, from http://www.geoplace.com/uploads/FeatureArticle/0502tt.asp.

Library of Congress. (2003). *Program announcement to support building a network of partners: Collaborative collection development.* Retrieved February 24, 2006, from http://www.digitalpreservation.gov/partners/pa_081203.pdf.

Library of Congress. (2006). *Digital preservation partnerships.* Retrieved February 24, 2006, from http://www.digitalpreservation.gov/partners/project.html.

Maine GeoArchives. (2004). *GeoArchives project agenda.* Retrieved February 24, 2006 from http://www.maine.gov/geoarch/agendaminutes/11_15_04_Agenda_files/11_15_04_Agenda.htm.

Mapdex. (2006). *Mapdex.* Retrieved February 24, 2006, from http://www.mapdex.org/search.

Martin, K. (2005). *Export to KML 2.1.* Retrieved February 24, 2006, from http://arcscripts.esri.com/details.asp?dbid=14273.

"Mashing the Web." (2005). *Economist 376*(8444), Special section, p. 4.

Mulka, K. (2005). *WMS in Google maps.* Retrieved February 24, 2006, from http://blog.kylemulka.com/.

Naval Research Laboratory. (2006a). *GIDB OpenGIS Web services.* February 24, 2006, from http://columbo.nrlssc.navy.mil/ogcwms/servlet/WMSServlet.

Naval Research Laboratory. (2006b). *Digital mapping, charting and geodesy analysis program team.* Retrieved February 24, 2006, from http://columbo.nrlssc.navy.mil/dmap/idx.jsp.

North Carolina Geographic Information Coordinating Council (NCGICC). (2003). *North Carolina OneMap: Geographic data serving a statewide community.* Retrieved February 24, 2006, from http://www.nconemap.net/documents/visiondoc.pdf.

North Carolina Geographic Information Coordinating Council (NCGICC). (2004). *2003 statewide local government GIS data inventory.* Retrieved February 24, 2006, from http://cgia.cgia.state.nc.us/nconemap/inventory.html.

North Carolina Geographic Information Coordinating Council (NCGICC). (2006a). *Current NC OneMap participants.* Retrieved February 24, 2006, from http://www.nconemap.net.

North Carolina Geographic Information Coordinating Council (NCGICC). (2006b). *NC OneMap.* Retrieved February 24, 2006, from http://www.nconemap.net.

North Carolina Geographic Information Coordinating Council (NCGICC). (2006c). *NC OneMap: Map service, layer naming and symbology standards.* Retrieved February 24, 2006, from http://204.211.135.110/onemap_standards/.

North Carolina State University (NCSU) Libraries. (2006a). *North Carolina county GIS data.* Retrieved February 24, 2006, from http://www.lib.ncsu.edu/gis/counties.html.

North Carolina State University (NCSU) Libraries. (2006b). *North Carolina geospatial data archiving project.* Retrieved February 24, 2006, from http://www.lib.ncsu.edu/ncgdap/.

Open Geospatial Consortium (OGC). (2004). *OpenGIS Web Map Service (WMS) implementation specification 1.3.* Retrieved February 24, 2006, from http://portal.opengeospatial.org/files/?artifact_id=5316.

Open Geospatial Consortium (OGC). (2005a). *GML simple features profile: Request for comments*. Retrieved February 24, 2006, from http://www.opengeospatial.org/specs/?page=requests&request=rfc22.

Open Geospatial Consortium (OGC). (2005b). *OpenGIS Web map context implementation specification 1.1*. Retrieved February 24, 2006, from https://portal.opengeospatial.org/files/?artifact_id=8618.

Open Geospatial Consortium (OGC). (2005c). *Open Web Feature Service (WFS) implementation specification 1.1*. Retrieved February 24, 2006, from https://portal.opengeospatial.org/files/?artifact_id=8339.

Open Geospatial Consortium (OGC). (2006). *Geo Digital Rights Management (GeoDRM) working group*. Retrieved February 24, 2006, from http://www.opengeospatial.org/groups/?iid=129.

Reichardt, M. (2005). *The havoc of non-interoperability*. Retrieved February 24, 2006, from http://portal.opengeospatial.org/files/?artifact_id=5097.

Ross, S., and Gow, A. (1999). *Digital archaeology: Rescuing neglected and damaged data resources. A JISC/NPO study within the Electronic Libraries (eLib) Programme on the Preservation of Electronic Materials*. Retrieved February 24, 2006, from http://eprints.erpanet.org/47/.

U.S. Geological Survey (USGS). (2005). *The National Map catalog service: A guide for application developers, document version 0.3.3, for catalog service versions 1.1.0 and 2.0.0*. Retrieved February 24, 2006, from http://mcmcweb.er.usgs.gov/catalog/docs/catalog_api.pdf.

Steve Morris is Head of Digital Library Initiatives as North Carolina State University (NCSU) Libraries and has worked for the past dozen years in the area of facilitating access to geospatial data. He is principal investigator in a partnership with the Library of Congress focusing on preservation of digital geospatial data. Steve has a master's degrees in both geography and library and information studies.