Digital Preservation of Geospatial Data

Julie Sweetkind-Singer, Mary Lynette Larsgaard, and Tracey Erwin

Abstract

The selection, acquisition, and management of digital data are now part and parcel of the work librarians handle on a day-to-day basis. While much thought goes into this work, little consideration may be given to the long-term preservation of the collected data. Digital data cannot be retained for the future in the same way paper-based materials have traditionally been handled. Specific issues arise when archiving digital data and especially geospatial data. This article will discuss some of those issues, including data versioning, file size, proprietary data formats, copyright, and the complexity of file formats. Collection development topics, including what to collect and why, will also be explored. The work underlying this article is being done as part of an award from the Library of Congress's National Digital Information Infrastructure and Preservation Program (NDIIPP).

Introduction

Digital geospatial data is now routinely found in libraries that carry cartographic data, geologic information, social science datasets, and other materials in support of disciplines using Geographic Information Systems (GIS) in their research and work. Over the course of years, the data have been received on floppy disks, CD-ROMS, DVDs, and hard drives or are available for free or for a fee over the Internet. In the paper world, ensuring longevity of items means creating ideal conditions in which to store collections. Materials will last longer if kept in a cool space without much light and correct humidity and handled as seldom as possible.

LIBRARY TRENDS, Vol. 55, No. 2, Fall 2006 ("Geographic Information Systems and Libraries," edited by Jaime Stoltenberg and Abraham Parrish), pp. 304–314

The same is not true for digital data. As Clay Shirky (of New York University's Interactive Telecommunications Program) pointed out in July 2005 at the bi-annual meeting of the National Digital Information Infrastructure and Preservation Program (NDIIPP), digital materials must be touched and manipulated on a regular basis if they are to survive. Leaving digital data alone will certainly cause it to be lost, and the time frame may be surprisingly short. Technology is changing at such a rapid pace that it can now be a challenge to find a machine that will read floppy discs, much less the obsolete program on which the data was supposed to run. Web sites can be and are removed at a moment's notice. This is especially frustrating for the federal depository libraries that formerly received paper copies of government information now available only in digital formats. Clearly, librarians must begin thinking about long-term preservation of their digital collection, from what to collect to ensuring that it is preserved with the same thoughtfulness and care that is given to hardcopy materials.

The Library of Congress and the NDIIPP Awards

In December 2000 Congress appropriated nearly $100 million dollars in funds to underwrite the cost of studying the issues related to the long-term preservation of digital data. The program was to be administered by the Library of Congress and was named the National Digital Information and Infrastructure Preservation Program (Library of Congress, 2006a). Conference Report H. Rept. 106–1033 stated that

> The overall plan should set forth a strategy for the Library of Congress, in collaboration with other Federal and non-Federal entities, to identify a national network of libraries and other organizations with responsibilities for collecting digital materials that will provide access to and maintain those materials. . . . In addition to developing this strategy, the plan shall set forth, in concert with the Copyright Office, the policies, protocols, and strategies for the long-term preservation of such materials, including the technological infrastructure required at the Library of Congress. (Library of Congress, 2006b)

The goal of the program was to create a network of committed partners willing to work on the policies, protocols, and architectures needed to build a series of archives to house digital materials.

The first round of major funding was announced in September 2004 with eight projects receiving a total of $13.8 million dollars in funding over a three-year period. Two of these projects focused specifically on geospatial data. The North Carolina State University Libraries partnered with the North Carolina Center for Geographic Information and Analysis to create a model for archiving the local and state government output of digital geospatial resources, including digitized maps. The project is designed to be a demonstration project for other states. The second contract was given jointly to the University of California at Santa Barbara

(UCSB) and Stanford University to underwrite the creation of the National Geospatial Digital Archive (NGDA). The NGDA's goal is to design repository infrastructures at each university and to collect materials across a broad spectrum of geographic formats. The team will work to expand the network of organizations committed to preserving geospatial content (Library of Congress, 2004).

The NGDA Project

The NGDA project has both research and development components. Research topics include considerations for long-term preservation; collection development, including prioritization and scope; architectural and economic models; rights issues; and best practices. The two libraries are developing prototype archives for housing the data and jointly creating a geospatial format registry to describe the data being stored. During the second year of the grant the two archives will be federated using the Alexandria Digital Library (ADL) software interface (see Figure 1).

Technical Architectures

The two repositories are being built using similar technologies while at the same time meeting the specific needs of each institution. Both architectures contain standards-based interfaces, clearly defined metadata formats, an underlying format registry, a goal of end-to-end automation of the systems, and exploration into open source front ends. UCSB has developed a repository specifically to house geospatial information, with tools and templates designed around common data structures. Stanford is building a repository to hold all of its digital content no matter what its nature; the goal is to determine if a general digital repository can adequately handle the complexities of geospatial data formats using standard metadata and a content transfer manifest, which include provisions for geospatial information. As of the end of December 2005, both repositories were complete through their first stages and had ingested geospatial data.

Format Registries

Technically, geospatial data is more complex than standard digital formats. This must be accounted for when archiving the data. In order to preserve a data format, information about that format must be known. The archive has to have an automated way to understand the file it has received and to verify that it is what it purports to be. This format information is typically stored in a registry, which records detailed metadata about the types of files. For example, format information for a GeoTIFF would include specifications for the correct TIFF standard and explanations of any accompanying files, such as those containing projection information. The format registry can be as complex as a custom-made database or as simple as a Web page or text document.

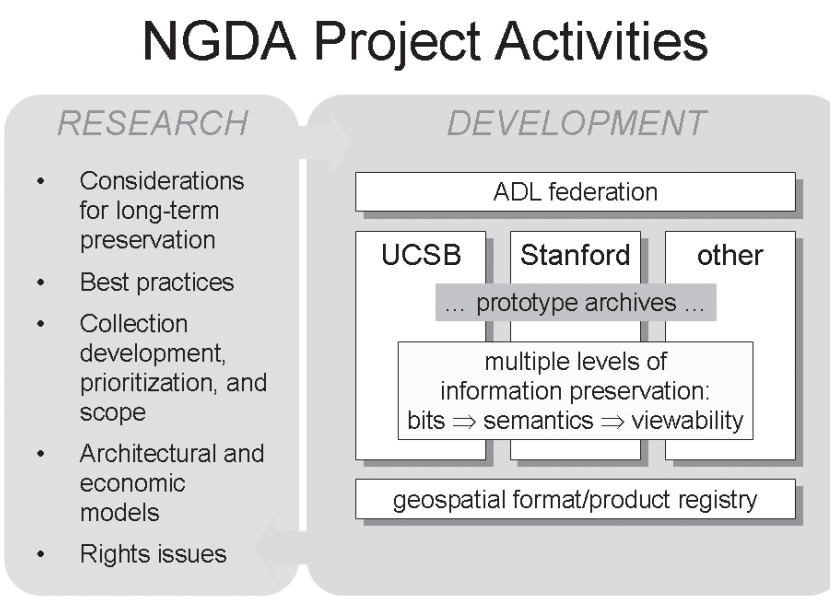

Figure 1. The overarching project activities for the three-year life of the NGDA contract with the Library of Congress

The Library of Congress, along with many other organizations, has spent a great deal of time describing formats and housing that information in format registries. Caroline Arms and Carl Fleischhauer compiled the format registry for the Library of Congress in order to help determine the sustainability of any given format throughout its content life cycle, to gain an understanding about which formats are more sustainable than others, and to develop strategies for sustaining the content they receive. The content categories studied include still image, sound, text, and moving image. They did not populate a format registry for geospatial data formats (Arms & Fleischhauer, 2005).

After searching unsuccessfully for other groups that had created a format registry for geospatial data, the NGDA team decided to build its own. Work is ongoing to describe the data elements necessary for preservation on four formats: digital orthographic quarter quadrangles (DOQQs), digital raster graphics (DRGs), Environmental Systems Research Institute (ESRI) "shapefiles," and Landsat imagery. (Other document formats will be analyzed as content selection for the archives progresses.) These file formats have differing levels of complexity. All of them contain multiple files that must travel together in order to make the format usable now and in the

future. For example, ESRI shapefiles are in a proprietary data format and are used in proprietary software. According to the specifications published by ESRI, only the .shp, .dbf, and .shx files are mentioned as part of the shapefile itself. But, with each shapefile there may also exist numerous other files, such as .sbn, .ain., and .prj files. Public documentation to reference these files and their role in the playability of the file itself is not available. Correspondence with ESRI staff was necessary to ascertain whether or not they considered the last three files necessary to preserve in the archive along with the published specification files. Building the format registry is labor intensive as it is necessary to trace the dependencies of files (a GeoTIFF must also include the correct TIFF specification, for example), and one must locally collect as much documentation as possible about each format. However, the set of format specifications should have to be created only once and then updated as necessary.

Rights Management and Contracts

Information regarding the rights governing the ownership, use, and copyright status of the data is associated with each file included in the repositories. A great deal of domestic geospatial data is produced by the U.S. government, which allows for wide use of its output due to the fact that most of it is in the public domain. But even government data may have copyright stipulations attached to it if it has been distributed through a third-party vendor contracted through a Cooperative Research and Development Agreement (CRADA) or has had value added to it by a commercial firm.

Important datasets, such as California's SPOT Image coverage and the base data on ESRI's Data & Maps CD-ROMS, are governed by strict licensing and use agreements. These datasets provide high-quality base map layers for GIS work and, especially with the yearly release of ESRI Data & Maps, provide longitudinal data that allow for the study of change over time. As the NGDA project moves forward, these agreements may make it impossible to federate data if the other potential repositories in the NGDA federation do not also have the legal right to hold the data. During the second year of the grant, the NGDA staff will begin a dialog with commercial data and imagery producers to assess their preservation strategies, awareness, and willingness to work with preservation archives.

In order to codify the rights and responsibilities of the repositories, each depositor will sign a contract licensing their content for preservation in the NGDA repositories. The goal is to create a single contract that can be used, and modified if necessary, by both Stanford and UCSB. The contract (in draft form as of this writing) governs the use, display, delivery, and preservation of materials in the NGDA. It clearly states who owns the copyright to the materials and ensures that those depositing materials in the archive have a right to do so. It further clarifies that the copyright stays with the original depositor and that the archives are not responsible in cases

of copyright infringement. It explains what may be distributed from the archives—the metadata, the data, or both—and to whom. It details how the repositories' rights and responsibilities will be carried out, including the need to use best practices and standards for preservation. The archives agree to take measures to prevent unauthorized access to the data, to permit only authorized users to access the content, to credit the copyright holders, and to use the utmost care in the preservation of the content. The contract explicitly allows the archives to manage the data to maximize its chances of survival over the long term.

In addition to legal protection for both parties afforded by the contract, a well-thought-out contract explicating the roles of each party builds an important element of trust that will encourage content creators to deposit their content in our repositories. The contract embodies one of the aspects of the trust-building activities recommended in the Research Libraries Group/Online Computer Library Center report, "Trusted Digital Repositories: Attributes and Responsibilities" (Research Libraries Group, 2002).

In order to further investigate how copyright law affects archiving of digital data, the Library of Congress has convened the Section 108 Study Group. Section 108 of the Copyright Act, created in 1976 and amended in 1998, governs the use of copyrighted materials held in libraries and archives. It is believed that even with the 108 revisions, the law is designed to meet the needs of the analog world, not the complex issues and needs of the digital one. This group has been charged with reviewing existing copyright laws as they pertain to libraries and archives, and specifically as they apply to digital media. The group will advise the Librarian of Congress in May 2006 on their findings and make recommendations based upon the needs of the content producers as well as those wishing to archive and access their output (Library of Congress, n.d.).

Collection Development

When the Library of Congress announced the Digital Preservation Program in August 2003, they enunciated the following three goals:

> "The continuing selection, collection, and organization of the most historically significant cultural materials and of important information resources, regardless of evolving formats,
>
> The long-term storage, preservation, and authenticity of those collections, and
>
> Persistent, rights-protected access for the public to the digital heritage of the American people." (Library of Congress, 2003)

Nature of Risk A required outcome of the project is to focus on materials that are deemed to be "at-risk" of disappearing or have no analog counterpart. While the Library of Congress did note that they considered historical and cultural materials or information "that document[s] key social and political developments necessary to understand contemporary

events" (Library of Congress, 2003) to be preservation worthy, they did not specifically define what it meant to have materials be "at risk." This is not surprising given the broad range of information across all disciplines in need of preservation.

Digital geospatial data may be deemed to be at risk because of many factors. The sheer magnitude of geospatial data being created and in existence makes it nearly impossible to collect it all for the future without significant efforts toward collaborative collecting models. MODIS data, used to study global dynamics and processes on the Earth, are being captured in thirty-six spectral bands from the MODIS satellites at a rate of a terabyte a day; over two petabytes of MODIS data are now stored at NASA. It is highly unlikely that a university, even the largest, would want to archive the whole MODIS output. On a more localized level, the problem of data storage is still significant. The state of California as represented in the Digital Orthophoto Quarter-Quadrangles includes approximately 13,200 scenes requiring roughly 670 gigabytes of storage space. In order to ensure viability of this dataset into the future, geographic redundancy is necessary in addition to the information being stored on different types of storage systems to lessen the chance of loss or corruption of data. This means the large datasets cannot be stored in a single location, creating the need for numerous, large, robust preservation environments.

In addition to the volume of data being produced, geospatial data are often updated and changed, creating the need to save different versions of the same information. How often the versions are collected will have to be decided on a case-by-case basis. For example, the National Elevation Dataset is updated on a bi-monthly basis by the United States Geological Survey (USGS) as higher-resolution or higher-quality data become available. Even a single data layer of a city GIS that is used by many different departments may be updated as often as several times a day. The different versions may be considered to be at risk because of the possibility that each iteration may need to be preserved (for example, for legal reasons, such as to prove when a change in a city's infrastructure was made). A strong argument can be made that each version need not be preserved in order to get a valid snapshot of the data environment.

Government geospatial data may well be considered at risk given the sensitive nature of some of the information, the decentralization of the computing environment, the lack of distribution of digital content that used to come to libraries as part of the Federal Depository Library Program, and the ease with which content can be removed from a government Web site. According to OMB Watch, the Bureau of Transportation Statistics (BTS) removed all GIS data, maps, and resources from their Web site after September 11, 2001. These data were later restored after the decision was made that their release did not pose a threat to national security (OMB Watch, 2005). Pipeline mapping data was removed from the Department of

Transportation's (DOT) Web site around the same time and has not been released to the public again. The DOT notes on their Web site that the data is now restricted "to pipeline operators and Local, State, and Federal government officials ONLY" (PHMSA, 2005).

Geospatial data is also potentially at risk for long-term preservation when it is produced by a small group or a single person. The ease with which content is now created and displayed has caused an explosion of small producers of high-quality geospatial content. Digital preservation requires a good deal of planning and expertise. It may also be prohibitively expensive to undertake. Simply making a backup copy of these data does not ensure that sufficient metadata has been captured to understand the environment in which the data was created, guard against failure of the storage mechanism, allow for geographic distribution, or solve the problem of file format migration over time. It is hoped that through the work done by this group and others the ability for small groups and individuals to archive safely their data for the long term will increase.

Collection Development Policies Collection development policies play a critical role in map libraries and have been important for many decades. The University of California/Stanford Map Libraries Group (UCSMLG) is still using the Research Library Group (RLG) conspectus portion for maps and geospatial data. The cooperative agreement is updated every five years and clearly spells out policies related to collaborative purchasing, collecting commitment levels for cartographic types of data and regions, and interlibrary loan. This agreement and the list of collecting responsibilities assigned to each university by call number have proved to be useful to this day (UC/Stanford Map Libraries Group, 2006).

Collection development policies typically do not include directives for long-term archiving of the collection itself. It has proven useful for us to review the work being done in the archiving community. While research libraries do, in general, keep their materials for a long period of time, they also weed with impunity for reasons of cost, space, and lack of use. An archive has made a commitment to keeping the material with the idea of turning it over to another trusted archive when they cannot or do not want to steward it any longer. In archives provenance is an integral component of responsible stewardship. Provenance details who or what group created and/or managed the records and traces the history of ownership of the records. This is critical information for a geospatial archive as well, and it must be included in the metadata. A good primer on archival practices is available from the University of Albany's M. E. Grenader Department of Special Collections and Archives (Parker et al., 2005).

Another area where archival practices influence long-term preservation is multiple file dependencies. Archival practices have codified the process of accessioning items in a specific order. This is necessary to preserve the contents as they were originally received and/or arranged. This is impor-

tant for a digital archive as well. Preserving such dependencies becomes critical when one thinks about long-term preservation of geospatial data incorporated on a Web site. The Arizona Model, being developed by the Arizona State Library, Archives, and Public Records, is using the framework of archival records management for the curation of collections of Web documents. They note that archiving Web documents by order translates into the correct management of the directories and subdirectories, which are called *series* and *subseries* in archival parlance. They argue that only through judicious use of archiving practices can large amounts of data be captured with a relatively small amount of human input. The system created must be able to scale and cannot do so if curators must select items one by one (Pearce-Moses & Kaczmarek, 2005).

The collection development policy for the geospatial archives being built by the NGDA will be a hybrid between a library policy and an archival one. It will include standard sections of a collection development statement that outline the user community; the geographic scope; the methods, scales, and frequency by which the materials are collected and updated; and the types of materials included. In addition, the policy statement will include descriptions of the type and quality of metadata that need to be included for ingestion into the repositories. Widely used file formats and types will be explained on a general level with the expectation that these will need to be updated over time. The Cornell University Geospatial Information Repository (CUGIR) has posted its collection development policy on its Web site, and it is a good example of this hybrid format (CUGIR WorkGroup of Mann Library, 2006).

The NGDA librarians will also produce specific collection development guidelines for their respective institutions. We hope that over time there will be many partner repositories in the federated network with broad collecting responsibilities. The individual nodes will focus on the needs of their primary audience, revising the policy to reflect individual institutional priorities. It is expected that areas of collecting interest will fall roughly along the same lines that were used when accessioning print materials. For example, the UCSB Map and Imagery Laboratory has a long history of collecting aerial and satellite imagery, while the Stanford Map Collections has focused heavily on geologic mapping and data. It is imperative that multiple collecting bodies be engaged in the process of selection and retention. There is just too much geospatial information being produced for a few libraries or institutions to preserve it all.

Conclusion

The National Geospatial Digital Archive team has completed the first of three years of their contract with the Library of Congress. Much more will be learned over the next two years from our research, the research of

other NDIIPP grants, as well as the work being done by others around the world in this field.

Year two goals include investigating to what degree, if any, commercial geospatial data producers are concerned with archiving and whether there is an interest in partnering with academic institutions; gaining a better understanding of existing mandates for archiving government-produced geospatial data; continuing to grapple with complex legal issues surrounding archiving; and ongoing technical development of the repositories themselves.

References

Arms, C., & Fleischhauer, C. (2005). *Sustainability of digital formats planning for the Library of Congress collections*. Retrieved January 24, 2006, from http://www.digitalpreservation.gov/formats/.

CUGIR WorkGroup of Mann Library. (2006). *About CUGIR: CUGIR collection development policy*. Retrieved January 31, 2006, from http://cugir.mannlib.cornell.edu/about.jsp.

Library of Congress. (n.d.). *The Section 108 working group*. Retrieved January 24, 2006, from http://www.loc.gov/section108/.

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

Library of Congress. (2004). *Library of Congress announces awards of $13.9 million to begin building a network of partners for digital preservation*. Retrieved January 24, 2006, from http://www.loc.gov/today/pr/2004/04-171.html.

Library of Congress. (2006a). *Digital preservation: National Digital Information Infrastructure and Preservation Program*. Retrieved January 31, 2006, from http://www.digitalpreservation.gov.

Library of Congress. (2006b). *Funding legislation for NDIIPP*. Retrieved January 24, 2006, from http://www.digitalpreservation.gov/about/fund.html.

OMB Watch. (2005). *Access to government information post September 11th*. Retrieved January 31, 2006, from http://www.ombwatch.org/article/articleview/213/1/DOT.

Parker, J., Williams, G., Keough, B., & Schnidler, A. (2005). *M. E. Grenander Department of Special Collections and Archives: Accessioning and processing manual*. Retrieved January 31, 2006, from http://library.albany.edu/speccoll/processing.htm.

Pearce-Moses, R., & Kaczmarek, J. (2005). *An Arizona Model for preservation and access of Web documents*. Retrieved January 31, 2006, from http://www.lib.az.us/DigGovt/azmodel/AzModel.pdf.

PHMSA. (2005). *National Pipeline Mapping System: Data dissemination*. Retrieved January 31, 2006, from http://www.npms.rspa.dot.gov/data/data_dissem.htm.

Research Libraries Group, Inc. (2002). *Trusted digital repositories: Attributes and responsibilities, an RLG-OCLC report*. Retrieved July 11, 2006, from http://www.rlg.org/legacy/longterm/repositories.pdf.

UC/Stanford Map Libraries Group. (2006). *Cooperative agreements*. Retrieved January 31, 2006, from http://library.ucsc.edu/maps/ucsmg/agreemts.html.

Julie Sweetkind-Singer is the Head Librarian at the Branner Earth Sciences Library and Map Collections at Stanford University. Her subject specialization is maps and GIS. She is currently Stanford's project lead on an NDIIPP grant from the Library of Congress. She has worked at Stanford since May 2000. In 1999 she worked jointly with David Rumsey on the Rumsey Map Collection Web site, which displays over 12,000 maps from the eighteenth and nineteenth centuries. She was the President of the Western Association of Map Libraries from June 2004 to July 2005. She was the Vice-President of the California Map Society, Northern Chapter, in 2001 and 2002. She received her M.L.I.S. from San Jose State University and M.B.A. from the University of Colorado at Boulder.

Mary Lynette Larsgaard has been Assistant Head of the Map and Imagery Laboratory, Davidson Library, University of California at Santa Barbara, since 1988. The Map and Imagery Lab has a collection of remote-sensing imagery and maps of approximately 5.5 million items and is the largest of its kind in any university library in North America. Larsgaard has published extensively in the field of geospatial data in libraries, most notably with a widely used text, *Map Librarianship: An Introduction* (now in a third edition, published in 1998 by Libraries Unlimited); she is also a coeditor (with Paige Andrew) of the *Journal of Map and Geography Libraries/Geoscapes*. Her specialties are cataloging/metadata creation and twentieth-century and more recent topographic and geologic maps. In the year 2000 she was promoted to Librarian, Distinguished Step, a promotion given only to librarians who have demonstrated superior competence and are internationally recognized as an authority in an area of library science.

Tracey Erwin is a Geospatial Librarian for the National Geospatial Digital Archive (NGDA), a grant-funded project of the University of California at Santa Barbara and Stanford University. The NGDA is a collaborative initiative funded by the Library of Congress to identify, collect, and preserve geospatial digital materials within a nationwide digital preservation infrastructure. Tracey is a recent graduate of San Jose State University's Library and Information Science Program.