Steven P. Morris, James Tuttle, and Robert Farrell. (2006, May 23, 2006). Preservation of State and Local Government Digital Geospatial Data: The North Carolina Geospatial Data Archiving Project Paper presented at Archiving 2006 Ottawa, Canada.

Paper: Preservation of State and Local Government Digital Geospatial Data: The North Carolina Geospatial Data Archiving Project

Steven P. Morris; North Carolina State University Libraries; Raleigh, North Carolina

James Tuttle; North Carolina State University Libraries; Raleigh, North Carolina

Robert Farrell; North Carolina State University Libraries; Raleigh, North Carolina

Abstract

The North Carolina Geospatial Data Archiving Project (NCGDAP) is a three-year joint effort of the North Carolina State University Libraries and the North Carolina Center for Geographic Information and Analysis focused on collection and preservation of digital geospatial data resources from state and local government agencies. NCGDAP is being undertaken in partnership with the Library of Congress under the National Digital Information Infrastructure and Preservation Program (NDIIPP). "Digital geospatial data" consists of digital information that identifies the geographic location and characteristics of natural or constructed features and boundaries on the earth. Such data resources include geographic information systems (GIS) data sets, digitized maps, remote sensing data resources such as digital aerial photography, and tabular data that are tied to specific locations. State and local data resources, which are in general of greater detail and more current than data available from federal agencies, are generally not addressed by data archiving efforts at the federal level.

Introduction – Risks to Digital Geospatial Data

"Digital geospatial data" consists of digital information that identifies the geographic location and characteristics of natural or constructed features and boundaries on the earth. Such data resources include geographic information systems (GIS) data sets, digitized maps, remote sensing data resources such as digital aerial photography, and tabular data that are tied to specific locations.

These complex data objects do not suffer well from neglect; long-term preservation will involve migration of data to supported data formats, media refresh, and retention of critical documentation.[1] Emerging data-streaming technologies further threaten archive development as it becomes easier to get and use data without creating a local copy—the secondary archive having been in part a by-product of providing data access.

The Challenge of State and Local Geospatial Data Resources

Geospatial data resources are created by a wide range of state and local agencies for use in applications such as tax assessment, transportation planning, hazard analysis, health planning, political redistricting, homeland security, and utilities management. State and local data resources are, in general, of greater detail and more current than data available from federal agencies. Since production points for these resources are so diffuse—92 of 100 North Carolina counties have GIS, as do many cities—they are generally not addressed by data archiving efforts at the federal level.

Although many of the targeted data resources are updated on a frequent basis—daily or weekly—data dissemination practices focus almost solely on providing access to current data. While snapshots of older versions of data may be stored in agency archives, access is almost as a rule not available and there is, in general, no commitment to long-term preservation of the data or to time series creation.

Domain-Specific Preservation Challenges

While digital geospatial data inherits preservation challenges that apply to digital resources in general, this content area also presents a number of domain-specific challenges to the preservation process.

Unique Data Formats

Digital geospatial data comes in two primary types, vector and raster. While the preservation challenges of raster (image) data are being tackled in many content domains, the challenges of vector data preservation are left primarily to the geospatial community.[2] In the case of vector data—also known as point/line/polygon data—there are is no satisfactory, open format to support long-term maintenance of content. The Spatial Data Transfer Standard (SDTS) has been put to use with some federal content but widespread support has not materialized. Geography Markup Language (GML)—a specification of the Open Geospatial Consortium (OGC)—provides a vendor neutral means of encoding vector data, but GML is not so much a format as it is a means to define something like a format in the way of a specific GML application schema that adheres to a specific GML profile. The emerging GML Simple Features Profile offers some hope for a widely-supported GML-based solution for longer-term maintenance of vector data.[3]

The emergence of spatial databases has further complicated the preservation of digital geospatial data. Spatial databases may consist of multiple individual datasets or "data layers," while also storing components such as behaviors, relationships, classification schemes, data models, or annotations that are external to or in addition to the datasets themselves. The whole of the spatial database is greater than the sum of the parts, as database

components that build on the individual data layers add value. These complex databases can be difficult to manage over time due to the complexity of data models, uncertainty over long-term support of proprietary database models, and reliance on specific database back ends for data storage.

Cartographic Representation and Project Files

The counterpart to the old archival map is not so much the GIS dataset as it is a meaningful collection of selected datasets linked with the appropriate symbolization, classification schemes, data models, and annotation. Unfortunately this collection is typically stored in a proprietary project file for which there is no preservation-safe alternative. Exporting the project file to a simple image format would capture the data view but lose the underlying data intelligence.

Semantic Issues

Inconsistent dataset naming, attribute naming, and attribute coding create both short- and long-term barriers to understanding and use of content. Data producers are discovering that naming and coding inconsistencies complicate the process of data sharing even in the context of present day use. Good metadata can make it possible to interpret these components, but unfortunately the data dictionaries associated with names and codes often do not accompany a data set.

Time-Versioned Content

At the local level many vector data resources are continuously or at least periodically updated, presenting three distinct challenges to the archiving process. First, the updated data in many cases is simply over-written or otherwise modified with no digital knowledge of the historic version maintained. Second, even if a data provider captures historic information, the absence of a standard identifier scheme makes it difficult to relate data versions outside of a local data collection context. Third, an optimal capture frequency may be difficult to determine for any particular type of data given the significant variation in update frequencies among data producers.

Geospatial Metadata

In the Unites States, the current geospatial metadata standard is the Federal Geographic Data Committee (FGDC) Content Standard for Digital Geospatial Metadata, commonly referred to as FGDC metadata.[4] In the near future, the current standard will be supplanted by the emerging North American Profile of the ISO 19139 metadata implementation specification for geographic information.[5] In terms of government data, FGDC metadata is mandated at the federal level, common at the state level, and sporadic at the local level. The archive may create or remediate certain FGDC sections, but only the data producers are adequately positioned to populate data quality and lineage information.

The North Carolina Geospatial Data Archiving Project – Origins and Context

NCSU's first efforts to acquire and preserve state and local geospatial data began in 2000 as rising user demand for state and local data coincided with a growing sense of long-term risk to this newly emerging content. The challenge of engaging and archiving content from most of the state's 100 counties as well as many municipalities helped to cultivate an understanding of the need for an infrastructure-based approach to archive development. Such an approach would build from existing geospatial data infrastructures that are evolving under the auspices of the National Spatial Data Infrastructure (NSDI), Federal Geographic Data Committee, and Geospatial One-Stop (GOS). These infrastructures—which incorporate local, state, and federal government agencies as well as the private sector—are already focused on such issues as data standards, best practices, data sharing agreements, metadata production and harvesting, catalog development, and services integration. Unfortunately, archiving and preservation have not yet become an area of focus in these efforts.

The primary manifestation of spatial data infrastructure in North Carolina is NC OneMap, 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 agency data holdings through an on-line data inventory.[6] Included in the NC OneMap vision statement is the assertion that "Historic and temporal data will be maintained and available."[7] While primarily focused on access and content standardization, NC OneMap has offered a means by which to engage the 100 counties and many municipalities of the state in the process of creating a digital preservation infrastructure.

In 2004 the NCSU Libraries and the NC Center for Geographic Information & Analysis entered into an agreement with the Library of Congress to pursue preservation of state and local digital geospatial data as part of the National Digital Information Infrastructure and Preservation Program (NDIIPP).[8] The North Carolina Geospatial Data Archiving Project (NCGDAP) will help inform development of a national digital preservation infrastructure through a "learning by doing" approach focused on identifying, acquiring, and preserving content within the context of the NC OneMap initiative and its framework of partnerships with state, local, and federal agencies.[9] As a component of the National Map, NC OneMap provides an opportunity to engage content through traditional distribution channels as well as through emerging web services based modes of access.

Project Work Plan Overview

Although this three-year project is focused solely on the state of North Carolina, it is expected to serve as a demonstration project for data archiving and time series development elsewhere. The objectives of the project include:

- Identification of available resources through existing statewide data inventory processes;
- 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; and
- Development of a model for data archiving and time series development.

NCGDAP Solutions in Progress: Ingest Workflow

The NCGDAP workflow incorporates a number of tools and processing steps that are commonly found in repository ingest workflows. JHOVE is used for format validation, although geospatial formats are, with the exception of GeoTIFF, not yet supported. MD5 checksums are used to validate content transfers at various points in the ingest process and ClamAV antiviral software is used for detecting infected files. The Nice Opaque Identifier (NOID) module is used to generate unique identifiers for ingest items.

Mechanisms are used to support the distillation of collections and complex objects into discrete repository ingest items within the workflow. A *transfer set seed file* is used to record and pass along metadata, typically administrative, which applies to an ingest *transfer set* as a whole. A software-generated *file manifest* accretes technical metadata elements during the ingest process and supports the process of grouping incoming files into *items* for the purpose of repository ingest.

Ingest Workflow: Domain-Specific Solutions

Geospatial data is characterized by a number of complicating factors that require careful redress within the ingest workflow.

Format Recognition and Complex Object Bundling

Complex, multi-file objects are the norm for many types of geospatial data. The individual dataset may be composed of files of different mime-types, and formats can often have optional file components. As a result, the ingest process must include an item grouping process, combining automated and manual approaches. Further complicating the ingest process is the existence of ancillary information such as license files, documentation, data models, scripts, and data dictionaries. Associating these entities with data items for repository ingest can be a challenging process that requires human intervention. Individual ancillary data files may apply to multiple separate datasets, requiring that the ancillary files be replicated across items prior to ingest.

Object Conversions

Much data is received in formats that are not archive-friendly by virtue of their complexity or impending obsolescence. Therefore, formats requiring immediate conversion in response to an immediate sense of risk will be identified early in the ingest process. These formats include the spatial database, topological vector data, and certain raster formats. Datasets of this sort will be archived in their native format as well as in a more archive friendly format.

Spatial databases provide a good example of the challenges faced in ingest workflow. A spatial database may be archived as is, but long term access to such content is not reliable given the complexity of the systems and the closed, proprietary nature of commercial implementations. In some cases XML export from these databases presents a more reliable preservation approach, but long-term software support of the XML files is open to question. Individual datasets may be extracted for retention in more stable forms, but database elements which span across datasets are lost in the process. In the NCGDAP workflow, a mix of approaches is taken, with binary databases, XML exports, and dataset extracts all being retained or cultivated. Extracted datasets become items unto themselves and receive ingest and metadata preparation attention at a level of detail that is higher than that given to the parent database.

NCGDAP Solutions in Progress: Metadata Workflow

NCGDAP will not just passively receive metadata. The metadata received will undergo integrity checking, inform administrative and technical metadata, and serve as a basis for feedback to the metadata production process in the geospatial community.

Metadata Quality Issues: Structural

FGDC metadata structural problems relate to the absence of an encoding standard associated with the current FGDC standard, which is focused on metadata content. This situation will be rectified with the finalization of the North American Profile of ISO 19139. The absence of consistent metadata structure and encoding necessitates a number of workflow steps related to normalization. The NCGDAP workflow currently involves normalization to the ESRI Profile of the FGDC standard.[10] The ESRI Profile is used due to its support for synchronization and due to its inclusion of additional technical and administrative metadata elements that inform the preservation process.

Metadata Quality Issues: Content

A number of metadata content quality issues need to be addressed in the workflow. In the case of existing metadata, it is necessary to synchronize the metadata with the dataset. Lack of concurrency between data and associated metadata is common. Resulting factual errors which need to be addressed include: incorrect format information resulting from producer shift in data environments; incorrect geodetic datum information resulting from an unrecorded shift in datum; and changing data access or contact information. Other metadata quality issues that need to be addressed include ambiguous data layer names and inadequate theme keyword assignment.

In cases where FGDC metadata is absent, minimal metadata is software-generated using available tools, while additional metadata elements are populated on the basis of data documentation, data inventories, and data producer information.

The Complete Metadata Package

Data producers and data consumers expect to deal with FGDC metadata. Archives will desire additional technical and administrative metadata elements, beyond those provided by FGDC, to support the preservation process. In the NCGDAP workflow selected FGDC elements are extracted, disambiguated, and in some cases improved in order to serve as discrete preservation metadata elements outside of the FGDC context. In the current iteration of the workflow METS records will be used as wrappers to combine FGDC metadata with additional technical and administrative metadata elements. These elements will be extracted from the FGDC metadata as well as from transfer set metadata and ingest workflow technical operations. In a future iteration of the workflow it is possible that this superset of preservation metadata will be repackaged as PREMIS metadata elements within the METS record.

Repository Software Considerations

NCGDAP is initially making use of DSpace repository software. A major consideration is that of cost, leveraging an existing organizational investment in DSpace for other projects. A second consideration is a desire to assess the interplay between geospatial data and mainstream digital repository software environments. Institutions which are already pursuing more general repository programs are increasingly interested in folding geospatial data into those efforts. It remains an open question whether or not domain-specific repositories are the best or only reasonable approach to handling this type of content.

An initial mapping of content and metadata to DSpace ingest objects and associated Qualified Dublin Core metadata is seen as just the first spoke in what is expected to be a multi-repository process. The core set of metadata will be maintained in a structure that is independent of the separate repository environments.

Emerging Issues

The proliferation of web services based on OGC specifications such as Web Map Service (WMS) and Web Feature Service (WFS) raises the possibility of automated harvesting of content. In the case of WMS, this activity might focus on construction of atlas-like collections of static images. In the case of WFS, the underlying data might be gathered as GML. As these web services are increasingly used as the basis for decision-making, documenting the basis for decisions will become more challenging. The OGC Web Map Context (WMC) specification provides a means to create a "spatial bookmark" that saves application and service state, but the issue of saving data state is not addressed.

The emergence of new mainstream web mapping environments such as Google Maps, Google Earth, Yahoo Maps, and MSN Virtual Earth is posing new technical and rights challenges to the preservation process. Dynamic map applications, or "web mashups," integrate data and services from multiple points or origin. At the same time, the related emergence of map service caching and tiling schemes has created some possible opportunities in the area of harvesting static, tiled content from caches in order to feed repository development.

Conclusion

Digital geospatial data resources are subject to various elements of exceptional risk owing to their complex and ephemeral nature. State and local geospatial data is particularly at risk given frequency of update and near absence of centralizing archiving efforts. The North Carolina Geospatial Data Archiving Project is, in collaboration with the Library of Congress, building a demonstration preservation experience in which the archive being developed is seen not as an end in itself but rather as a catalyst for discussion among the various elements of spatial data infrastructure. That discussion, which includes libraries and archives, is centered not just on preservation processes and best practices but also on roles and responsibilities of the various players in what constitutes spatial data infrastructure. In terms of the technical processes, distilling complex geospatial content into discrete ingest items presents a significant challenge. NCGDAP will be providing feedback to the geospatial producer community about content quality and metadata quality issues in hopes of improving the consistency of content and metadata acquired and in hopes of further routinizing the process of archive development.

References

[1] Bleakely, Denice R., "Long-Term Spatial Data Preservation and Archiving: What are the Issues?" Sand Report, SAND 2002-0107. Sandia National Laboratories. (2002). Available from: http://www.prod.sandia.gov/cgi-bin/techlib/access-control.pl/2002/020107.pdf [accessed 07 March 2006].

[2] Zaslavsky, Ilya, "Archiving Spatial Data: Research Issues." San Diego Supercomputer Center Technical Report TR-2001-6. (2001). Available from: http://www.sdsc.edu/TR/TR-2001-06.doc.pdf [accessed 07 March 2006].

[3] Open Geospatial Consortium, "GML Simple Features Profile." Available from: http://portal.opengeospatial.org/files/?artifact_id=11266 [accessed 10 March 2006].

[4] Federal Geographic Data Committee, Content Standard for Digital Geospatial Metadata (FGDC CSDGM). Available from: http://www.fgdc.gov/metadata [accessed 10 March 2006].

[5] Federal Geographic Data Committee, FGDC/ISO Metadata Standard Harmonization. Available from: http://fgdc.er.usgs.gov/metadata/whatsnew/fgdciso.html. [accessed 10 March 2006].

[6] NC OneMap. Available from: http://cgia.cgia.state.nc.us/nconemap/ [accessed 10 March 2006].

[7] NC OneMap Vision Statement. Available from: http://cgia.cgia.state.nc.us/nconemap/documents/visiondoc.pdf [accessed 10 March 2006].

[8] Library of Congress, National Digital Information Infrastructure and Preservation Program (NDIIPP). Available from: http://www.digitalpreservation.gov/ [accessed 10 March 2006].

[9] North Carolina Geospatial Data Archiving Project (NCGDAP). Available from: http://www.lib.ncsu.edu/ncgdap/ [accessed 10 March 2006].

[10] Environmental Systems Research Institute, "ESRI Profile of the Content Standard for Digital Geospatial Metadata." Available from: http://www.esri.com/metadata/esriprof80.html [accessed 10 March 2006].

Author Biography

Steve Morris is Head of Digital Library Initiatives at North Carolina State University Libraries. He is principal investigator in the NCGDAP cooperative project with Library of Congress under the NDIIPP partnership program. Mr. Morris has an M.A. in Geography from California State University Chico and a Masters degree in Library and Information Science from University of California, Berkeley.

Jim Tuttle is the Geospatial Data Librarian and is a Libraries Fellow at North Carolina State University Libraries. Jim holds a B.A. in Anthropology and Master of Science in Library and Information Science from the University of Illinois at Urbana-Champaign. His current responsibilities include designing and implementing archive ingest workflow and metadata automation processes.

Rob Farrell is Geospatial Initiatives Librarian for North Carolina State University Libraries. His primary functions for NCGDAP revolve around ingest and metadata workflow. He holds a B.S. in Statistics from North Carolina State University and an M.A. in Geography from the University of North Carolina at Charlotte.