Curation and Preservation of Complex Data:
The North Carolina Geospatial Data Archiving Project

Steven P. Morris; North Carolina State University Libraries; Campus Box 7111, Raleigh, North Carolina; steven_morris@ncsu.edu

James Tuttle; North Carolina State University Libraries; Campus Box 7111, Raleigh, North Carolina, 27695; jim_tuttle@ncsu.edu

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. These complex data objects do not suffer well from neglect, and long-term preservation will involve some combination of format migration and retention of critical documentation. While the main focus of NCGDAP is on organizational issues related to the engagement of spatial data infrastructure in the process of data archiving--with the demonstration repository seen more as a catalyst for discussion rather than an end in itself--this paper focuses more narrowly on the technical challenges associated with developing an ingest workflow and archive development process. New preservation challenges associated with emergent content forms are also presented.

North Carolina Geospatial Data Archiving Project - Background

The North Carolina Geospatial Data Archiving Project (NCGDAP), a partnership effort involving the North Carolina State University Libraries and the North Carolina Center for Geographic Information & Analysis, is addressing the challenge of preserving state and local government digital geospatial data.[1] This effort, which is being undertaken in cooperation with the Library of Congress data as part of the National Digital Information Infrastructure and Preservation Program (NDIIPP), aims to inform development of a national digital preservation infrastructure through a "learning by doing" approach focused on identifying, acquiring, and preserving content.[2] The project is taking place within the context of the NC OneMap initiative and its framework of partnerships with state, local, and federal agencies.[3] As a component of the National Map, the NC OneMap framework provides an opportunity to engage content through traditional distribution channels as well as through emerging web services based modes of access.

At the state and local government level geospatial data resources are created by a wide range of agencies for use in applications such as tax assessment, transportation planning, hazard analysis, health planning, political redistricting, and utilities management. These data resources are, in general, of greater detail and more current than data available from federal agencies, yet production points for these resources are diffuse—96 of 100 North Carolina counties have GIS, as do many cities—posing many challenges to the archive development

process. Many of the targeted data resources are updated on a frequent basis—daily or weekly in some cases—yet data dissemination practices, for the most part, focus on providing access to current data.

Characteristics of 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, tabular data that are tied to specific locations, and various ancillary data resources. Since many of these data resources exist in complex forms which do not suffer well from neglect, long-term preservation will often involve migration of data to more sustainable data formats.[4] Emerging web services- or API-based technologies pose further challenges to the archive development process. As it becomes easier to get and use data without creating a local copy, there may be less incentive to create the secondary archives that have been in part a by-product of providing data access.

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.[5] In the case of vector data there are is no satisfactory, open format to support long-term maintenance of content. 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, data models, or annotations that are external to or in addition to the datasets themselves.

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 annotations. Unfortunately this added information is typically stored in proprietary project files or in spatial databases, either of which is difficult to preserve by virtue of complexity. Exporting the project file to an image format such as TIFF would capture the data view but lose the underlying data intelligence. Increasingly these finished cartographic products are being published in PDF format, inheriting any preservation challenges associated with complex PDF documents. PDF variants such as GeoPDF and PDF/E may introduce additional preservation challenges.

Semantic Issues

Heterogeneous approaches to dataset naming, attribute naming, and attribute classification schemes create both short- and long-term barriers to understanding and use of content. While good metadata can make it possible to interpret these components, such metadata is unfortunately often absent or may not include the data dictionaries associated with names and codes found in the data. Framework data content standards, such as those

developed for some data layers in the state of Maine, provide some hope for improved consistency in the content and structure of geospatial data.[6]

Time-Versioned Content

Many vector data resources are continuously or at least periodically updated. Unfortunately this data is often simply overwritten or otherwise modified with no digital record of the historic version maintained. In order to document current practices for obtaining archival snapshots of county and municipal geospatial vector data layers, NCGDAP conducted a frequency of capture survey targeting local agencies in North Carolina. With 58% of targeted agencies responding to the survey, roughly two-thirds of respondents indicated that they create and retain periodic snapshots of vector data. Retention activity was more common in counties with larger populations, and practice with regard to capture frequency, preservation format, and storage environment varied greatly.

Time-versioned content presents some unique challenges to repository development in terms of managing item relationships. Logical item associations are particularly difficult to abstract out of specific repository data models.

Geospatial Metadata

In the Unites Stated, the geospatial metadata standard since 1994 has been the Federal Geographic Data Committee (FGDC) Content Standard for Digital Geospatial Metadata, commonly referred to as FGDC metadata.[7] In the near future, the current standard will be supplanted by the emerging North American Profile of the ISO 19115 metadata implementation specification for geographic information, using the ISO19139 XML schema implementation.[8] While FGDC metadata is mandated at the federal level and is often available at the state level, detailed metadata is rarer at the local level. In cases where metadata is absent an archive may able to populate some FGDC record sections, yet only the data producers have the information needed to populate sections such as data quality and lineage. To the extent that metadata exists, the records are often asynchronous with the data or simply incorrect. In addition, since the original FGDC standard was a content standard for which no standard encoding was defined; existing metadata commonly requires some degree of structural normalization in order for the metadata to be interoperable with a repository.

Ancillary Data

Geospatial datasets, in addition to often being multi-file in nature, are frequently accompanied by ancillary data files which must be retained in order to properly use or understand the data. Ancillary files include metadata records, data dictionaries, additional data documentation, legend files, data licenses, disclaimers, and associated images. In many cases an individual ancillary file will be shared by many or all datasets in a given collection, creating the requirement that the ancillary files either be replicated for bundling with individual datasets or referenced separately through a persistent access mechanism. Given the absence of a standard content packaging scheme such as METS, MPEG 21 DIDL, or XFDU in the geospatial industry, it falls to the repository to create item associations through a combination of automated mechanisms and human intervention.

Repository Ingest Approach

In developing an ingest workflow some key decisions needed to be made with regard to degree of ingest automation, degree of dependence on the initial repository software environment, and degree of trust in received metadata.

Balancing automation with human intervention

NCGDAP has no control over the manner in which received data is organized, and designing automated ingest processes for heterogeneous and irregular content is very difficult. Hands-on data processing remains important for some parts of the ingest process, raising ingest costs and increasing the possibility of introducing human error. However, as experience is accumulated it becomes possible to automate additional portions of the ingest process. In the longer term, partnership efforts focused on developing routine data transfer processes are hoped to increase the ability of the archive to automate ingest automation and lower ingest costs.

Repository-agnostic architecture

In order to reduce long-term reliance on a particular digital repository software platform the project has attempted to abstract repository ingest processes, developing more generalized ingest workflows in hopes of making the data more adaptable to other repository environments. This adaptability will be necessary to support repository software upgrades, migration to other repository software, and inter-archive exchange. Ultimately, however, it can be difficult to prevent the repository organizational model from imprinting on the data.

Metadata remediation

Supporting long-term access to the data requires both retention of original metadata--as an archival object--and remediation of that metadata to increase the likelihood of successful use of that data. Remediation steps include error correction, functional improvements to key access fields, structural normalization, and synchronization of the metadata record with the dataset at hand.

Data Receipt Processes

The repository ingest workflow involves a multi-stage process that starts with receipt of heterogeneous data collections from a variety of data producers and ends with content that has been distilled into discrete data items that can be handled by digital repository software. It should be acknowledged that this is only one approach to data preservation, with *in situ* management of complex data within native software environments being another approach. For example, the Maine GeoArchives project has conducted investigations into issues related to maintaining spatial database content over time.[9]

Data Acquisition

Geospatial data is received by the project either as a data download or as a set of files delivered on optical or magnetic media. To ensure file integrity MD5 checksums are generated and recorded as early in the process as possible. Data transferred to local systems from media are checksum validated as part of the copy procedure.

An initial metadata “seed file”, including collection information, is generated on transfer of all data. Collection-level information such as acquisition date, acquisition circumstances, and transaction-specific rights information are captured in the seed file, which then informs later metadata processes as individual technical and administrative metadata elements propagate to the item level. The seed file metadata is entered into a form that encapsulates it in well-formed XML. Using a simple, internally-developed rights expression scheme, rights are registered in both human-readable and machine-actionable forms to support the possibility of fine-grained access control in the future.

Dataset Reorganization

Given the lack of control over of the manner in which data is delivered to the archive, some degree of remediation of data organization is required prior to ingest. Data files must also be extracted by archive file formats such as tar, zip, and the ESRI e00 format. Extraction of archive formats requires some manual intervention due to the unpredictable nature and hierarchy of the contents. In addition, particularly complicated formats, such as Geodatabases, are disaggregated into constituent elements in order to form discrete repository items.

Dataset Validation and Threat Analysis

All data are scanned for viruses and executability as identified by magic numbers using the UNIX 'file' utility, and infected and executable files are deleted and reported to the contributing agency. Harvard's JHOVE utility is used to validate formats that it recognizes. Support for geospatial formats is currently limited, but validation of the ESRI Shapefile format is a planned feature of JHOVE2.[10] In the case of digital orthophoto collections, image extent grids are developed in order to identify any gaps in transfer sets.

Dataset Inventory

Following initial remediation, data sets are cataloged in a machine-readable manifest used to track changes and validate integrity. The path, size and checksum from the set manifest are compared to the data set after each processing stage to flag unintended changes to the data set. Intended changes, such as format migrations, are recorded in the provenance metadata for each file. The manifest accompanies the dataset through the ingest process and is used to ensure validity throughout the pre-ingest workflow as well as post-ingest.

Format Processing

Geospatial vector data file formats are highly complex and extremely sensitive to both format migration and software display environment, and backwards compatibility between versions of the same software can be problematic. The absence of widely adopted, open vector formats ensures that a preponderance of vector data exists in proprietary formats. Due to the complexity of the content, migration between formats can lead to unacceptable data distortion and data loss.

Format Conversions

Vector data provided in formats with less market share than the market leader are migrated while the data is also preserved in the original format. Image data provided in obscure formats or in proprietary compressed formats are converted to TIFF.

Compound Format Handling

In the case of compound formats such as the ESRI Geodatabase or desktop GIS project file formats, the sum of the parts (mostly individual datasets) is not equal to the whole, but these complex data structures are less likely to weather long-term preservation and so, as a hedge against loss, are disaggregated into individual repository items to be redundantly archived. Unfortunately, compound format migration is a one-to-many operation and the relationships between migrated files as maintained by the compound file structure are completely lost in these extractions. Individual shapefiles and raster datasets are extracted, but the original database is retained both in its native binary format and in its XML export format.

Metadata Processing

Data is commonly produced and distributed at the local level without metadata. To the extent that arriving data has any metadata at all, the received metadata will often require remediation.

Creating Metadata Where None Exists

In cases where metadata is absent minimal metadata is generated for ESRI formats using agency-specific templates applied using ArcGIS software. In addition to elements that can be auto-extracted using the ArcGIS toolset, additional metadata elements are derived from agency websites and statewide data inventories. Provenance statements are added to explicate the role of the archive in the data lifecycle.

Metadata Remediation

Existing metadata typically must be remediated from the point of view of content, structure, and concurrence with the data. A coherent dataset title, organization name, date, subject keywords, and bounding coordinates form the key access points for geospatial datasets in data discovery systems and--at minimum--it is important that these fields are populated with functional content. To facilitate data discovery and identification dataset names are disambiguated to the publishing agency and ISO 19115 subject keywords are assigned to enhance discoverability. To insure concurrence of the data with the metadata the ArcGIS toolset is used to synchronize the metadata, while bounding coordinates are also extracted in an automated fashion. In the case of orthophotography the date field is used to denote flight year. In the case of vector data the date is assumed to represent the year of vector data production.

Repository Ingest Process

The processed geospatial metadata together with the seed file metadata and the technical metadata generated during pre-ingest processing all form a superset of metadata which can be used to inform creation of Submission Information Packages for various systems as well as to populate a metadata database used to separately track project content.

The Metadata "Hub"

In order to facilitate transformations to alternate archive Submission Information Packages and in order to flexibly acquire content through different pre-ingest processes; the project has adapted the "hub-and-spoke" model employed by the University of Illinois ECHO DEPository NDIIPP project.[11] The premise behind this metadata transformation model is that each transformation is a spoke connected to a central hub in which metadata exists in a special profile from which it may be transformed into other schemas. An independent metadata database stores metadata independent of the DSpace archive and may be used to interoperate with other discovery systems, including the local data discovery environment. The external metadata database also provides a simple means to generate reports concerning the contents of the archive.

It is expected that a spoke will later be added to generate a METS record for each item. The METS record would not be actively used in the current archive, yet it might prove useful for inter-archive exchange.

Submission Information Packaging

Some initial SIP organization is achieved with a custom-built ArcGIS toolbar as formats are migrated. The remaining items that may be machine organized are grouped using a Python script employing a file association logic. Common geospatial formats may consist of as many

as four required files and any number of optional files. Additionally, there may be ancillary files that require manual association to items.

Repository-specific SIP transformations do not occur until item formations have been completed. The hub-and-spoke metadata transformation process currently creates a DSpace Qualified Dublin Core record for each item, however it is anticipated that the hub profile metadata could be used to generate SIP's for other environments. This SIP might also be used as a DIP to create a SIP for other environments.

Geospatial Data Archiving Challenges: Looking Forward

Web Services and Mapping APIs

As web services-based systems are increasingly used in decision-making processes, documenting the basis for decisions will become more challenging. The proliferation of web services based on Open Geospatial Consortium (OGC) specifications such as Web Map Service (WMS) and Web Feature Service (WFS) raises the possibility of automated harvesting of content.[12] 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 (Geography Markup Language), yet WFS has not been widely deployed, and GML itself poses preservation challenges. The emergence of new mainstream web mapping environments such as Google Maps, Windows Live Local, Yahoo Maps, Google Earth and MSN Virtual Earth is posing new technical and rights challenges to the preservation process, as dynamic map applications and "web mashups" integrate data and services from multiple points or origin. The related emergence of map service caching and tiling schemes is creating some possible opportunities in the area of harvesting static, tiled content from caches in order to feed repository development. Open source community efforts to define a Tile Map Service specification might lead to opportunities to capture temporal context in the web services, mapping API, and decision support system environments.[13] Some of these issues may become subjects of focus in the newly formed OGC Data Preservation Working Group.[14]

Place-based Data

The increasing prominence of place-based—as opposed to spatial—data is creating new challenges and opportunities for archive development and temporal data management. Examples of place-based data include county tax assessors building images, state department of transportation videologs, oblique imagery produced for state and local agencies, and street view imagery such as that once produced by A9.com and now produced for the MSN Virtual Earth environment. While place-based content might once have been considered ancillary to the spatial data, the geographically comprehensive nature of these newer resources makes them central in new mass market applications in the area of mobile computing, Location-Based Services (LBS), and social networking. The lateral views that these images present become particularly valuable in a historical context because these views are more descriptive of place and building function than are the overhead views provided by traditional geospatial imagery such as digital orthophotography.

Structural and Three-Dimensional Data

The convergence of geospatial with three-dimensional and structural information, as evidenced by the emergence of Virtual Earth 3D and Google SketchUp, and by the proliferation of oblique imagery data products, promises to create even more complex challenges in archiving of digital geospatial data. Related specifications development activities associated with PDF/E and within the context of 3D Information Management (3DIM) Working Group in the OGC bear watching.

Conclusion

Digital geospatial data resources are subject to various elements of exceptional risk owing to their complex and ephemeral nature. 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. More direct integration of archives with emergent content exchange networks may help to streamline ingest processes and lower overall archive development costs. As a by-product of the repository ingest effort, 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 building more efficiency into the process of archive development.

Acknowledgments

The work of the North Carolina Geospatial Data Archiving Project is funded by the Library of Congress under the National Digital Information Infrastructure and Preservation Program.

---

[1] North Carolina Geospatial Data Archiving Project (NCGDAP). Available from: http://www.lib.ncsu.edu/ncgdap/ [accessed 15 March 2007].

[2] Library of Congress, National Digital Information Infrastructure and Preservation Program (NDIIPP). Available from: http://www.digitalpreservation.gov/ [accessed 15 March 2007].

[3] NC OneMap. Available from: http://www.nconemap.net/ [accessed 15 March 2007].

[4] Bleakely, Denise 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/accesscontrol.pl/2002/020107.pdf [accessed 15 March 2007].

[5] 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 15 March 2007].

[6] State of Maine: Standards For Digital Parcel Files, Presented to the Maine Geolibrary Board, by the Digital Parcels Standards Committee January 19, 2005. Available from: http://apollo.ogis.state.me.us/standards/ [accessed 15 March 2007].

[7] Federal Geographic Data Committee, Content Standard for Digital Geospatial Metadata (FGDC CSDGM). Available from: http://www.fgdc.gov/metadata [accessed 15 March 2007].

[8] Federal Geographic Data Committee, FGDC/ISO Metadata Standard Harmonization. Available from: http://www.fgdc.gov/metadata/geospatial-metadata-standards [accessed 15 March 2007].

[9] Maine State Archives. Geoarchives, a Collaborative Project between the Maine State Archives and GeoLibrary Board. Available from: http://www.maine.gov/sos/arc/GeoArchives/geoarch.html [accessed 15 March 2007].

[10] JHOVE2: A Next-Generation Architecture for Format-Aware Digital Object Preservation Processing. Available from: http://hul.harvard.edu/jhove/JHOVE2-proposal.doc html [accessed 15 March 2007].

[11] Rani, Shweta, et al. Technical Architecture Overview: Tools for Acquisition, Packaging and Ingest of Web Objects into Multiple Repositories [poster]. JCDL'06, June 11-15, 2006, Chapel Hill. Abstract available from: http://www.ndiipp.uiuc.edu/pdfs/jcdl_2006_poster_pp217_b.pdf [accessed 15 March 2007].

[12] Morris, Steven P. "Geospatial Web Services and Geoarchiving: New Opportunities and Challenges in Geographic Information Services," Library Trends, Vol. 55, No. 2, Fall 2006.

[13] Open Source Geospatial Foundation.Tile Map Service Specification. Available from: http://wiki.osgeo.org/index.php/Tile_Map_Service_Specification [accessed 15 March 2007].

[14] Open Geospatial Consortium. Data Preservation WG. Available from: http://www.opengeospatial.org/projects/groups/preservwg [accessed 15 March 2007].