Next Generation Archives: The NC Geospatial Data Archiving Project

Jeff Essic
North Carolina State University Libraries

Zsolt Nagy
North Carolina Center for Geographic Information and Analysis
NC Geospatial Data Archiving Project (NCGDAP)

- Three year partnership between university library (NCSU) and state agency (NCCGIA), with Library of Congress under the National Digital Information Infrastructure and Preservation Program (NDIIPP)

- One of 8 initial NDIIPP collection building partnerships

- Focus on state and local geospatial content in North Carolina (*state demonstration*)

- Tied to NC OneMap initiative, which provides for seamless access to data, metadata, and inventories
**NCGDAP Specifics**

- **Funding:**
  - $520,000 for 2005-2007
  - $500,000 for 18 month extension

- **Staff:**
  - 1.5 FTE at NCSU
  - Approx. same at NCCGIA

- **Website:**  [http://www.lib.ncsu.edu/ncgdap](http://www.lib.ncsu.edu/ncgdap)
## Selected Geospatial Data Archive Projects

<table>
<thead>
<tr>
<th>Project</th>
<th>Organizations</th>
<th>Funding</th>
</tr>
</thead>
<tbody>
<tr>
<td>Persistent Archives Testbed</td>
<td>San Diego Supercomputer Center, NARA</td>
<td>NARA</td>
</tr>
<tr>
<td>VanMap</td>
<td>San Diego Supercomputer Center</td>
<td>Inter-PARES</td>
</tr>
<tr>
<td>Geospatial Repository for Academic Deposit &amp; Extraction</td>
<td>EDINA</td>
<td>JISC</td>
</tr>
<tr>
<td>Geospatial Electronic Records</td>
<td>CI ESI N</td>
<td>NHPRC</td>
</tr>
<tr>
<td>various</td>
<td>Carleton University</td>
<td>various</td>
</tr>
<tr>
<td>National Geospatial Digital Archive</td>
<td>UC Santa Barbara</td>
<td>NDIIPP</td>
</tr>
<tr>
<td>Maine GeoArchives</td>
<td>State of Maine</td>
<td>NHPRC</td>
</tr>
</tbody>
</table>
Tracking data, map servers, and web services since 2000

Earliest use: Links to local data contacts and downloads

Now: Ranked 3rd in traffic among entry points to entire library website

Community help in site maintenance
County Map and Data Services in NC

100 Counties in North Carolina

![Graph showing the number of counties from 2000 to 2008.](chart.png)

- **Map Server**
- **Data Download**
- **WMS**
Value in Older Data: Cultural Heritage

Future uses of data are difficult to anticipate (as with Sanborn Maps)
Geospatial Data: Compelling Issues

- Dynamic content
  - Constantly updated information
  - Data versioning

- Digital object complexity
  - Spatially enabled databases
  - Complicated, multi-component formats
  - Proprietary formats
**Digital Preservation Points of Failure**

- Data is not saved, or …
- can’t be found, or …
- media is obsolete, or …
- media is corrupt, or …
- format is obsolete, or …
- file is corrupt, or …
- meaning is lost
Risks to Geospatial Data

- Producer focus on current data
  - Data overwrite as common practice
- Future support of data formats in question
  - No open, supported format for vector data
- Shift to web services-based access
  - Data becoming more ephemeral
- Inadequate or nonexistent metadata
  - Impedes discovery and use
- Increasing use of spatial databases for data management
  - The whole is greater than the sum of the parts
Preservation Business Case

- Land use change analysis
- Site location analysis
- Real estate trends analysis
- Disaster response
- Resolution of legal challenges
- Impervious surface change mapping
Business Case: Identifying Land Use Changes

Use case:
Land use and impervious surface change analysis
Geospatial Data Preservation Challenges

- Data Capture
  - Backups are common, but not long-term archives
  - Producer focus is on current data
  - Shift to web services-based access

- Inadequate or Nonexistent Metadata
  - Consistent NC survey stats: Only 40% of data producers create and maintain metadata
Challenge: Vector Data Formats

- No widely-supported, open vector formats for geospatial data
  - Spatial Data Transfer Standard (SDTS) not widely supported
  - Geography Markup Language (GML) – diversity of application schemas and profiles a challenge for “permanent access”

- Spatial Databases
  - The whole is more than the sum of the parts, and the whole is very difficult to preserve
  - Can export individual data layers for curation, but relationships and other context are lost
Challenge: Other Data Types

- Cartographic Representation
  - Software Project Files, PDFs, GeoPDFs, WMS images

- Web 2.0 content
  - Street views, Mashups

- Oblique Imagery

- 3D Models
Other Challenges

- Rights management
- Data versioning
- Digital Object Complexity
- Semantic issues
- Content Packaging
- Large scale content transfer
- Integrating older analog materials
- More …
Different Ways to Approach Preservation

- **Technical solutions**: How do we preserve acquired content over the long term?

- **Cultural/Organizational solutions**: How do we make the data more preservable—and more prone to be preserved—from point of production?

*Current use and data sharing requirements – not archiving needs – are most likely to drive improved preservability of content and improvement of metadata*
Repository of Temporal Data Snapshots

- Question: Frequency of Capture?
- Content Exchange – Getting Data in Motion
- Repository Development
Frequency of Capture

Issue: How frequently should county and municipal vector data layers be captured in archives?

Parcels, centerlines, jurisdictions, zoning, …
Frequency of Capture Surveys

- How often should continually changing vector datasets be captured?
- Tap into data custodian understanding of production patterns and uses
- Tap into local innovation
- Learn about local business drivers for data archiving
  - 2006 and 2008 surveys of NC cities and counties
  - 2008 survey of archival practice in state agencies in NC
  - Planned survey of data users in NC
FOC 2006 Survey Results: Overview

- 58% response, two-thirds of whom create and retain periodic snapshots
- Long-term retention more common in counties with larger populations
- Storage environments vary, with servers and CD-ROMs most common
- Wide variation in frequencies of capture.
- Offsite storage (or both onsite and offsite) is used by nearly half of the respondents
- Popularity of historic images has resulted in scanning and geo-referencing of hardcopy aerial photos among one-third of the respondents
Content Exchange Infrastructure

- High volume of state/federal requests for local data
- Solving the present-day problems of data sharing is a pre-requisite to solving the problem of long-term access
- Nov. 2007: NC Geographic Information Coordinating Council (GICC) approved “Ten Recommendations in Support of Geospatial Data Sharing”

http://www.ncgicc.org/
Most costly part of archive development is identifying, negotiating acquisition, and then transferring data.

**Getting the Data in Motion**

- **Important Objectives**
  - Minimize Direct Contact
  - Provide Metadata
  - Clarify Rights
  - Routinize Transfers

- **Leverage other business uses that put data in motion:**
  - Continuity of operations
  - Highway Planning
  - Floodplain Mapping
  - Census
Getting the Data in Motion

Orthophoto Data Distribution System – “sneakernet”
Transfer of large quantities of imagery

Street Centerline Data Distribution System
Efficient transfer of data from 100 counties, with metadata and clarified rights
http://www.ncstreetmap.com

NC GIS Inventory
• Efficient data identification
• Adding preservation elements

Local Government Responses
- Cities
- Counties

NC OneMap Data Discovery and Download
Data can be downloaded from the NC OneMap FTP site at no cost. Vector and raster data are available. If you do not find what you are looking for, below is the list of available data. You can download using an FTP client. To do so, log in with the username "anonymous" and password "<your email address>". The data is in a ZIP file.

NC OneMap Data Download and Viewer
• Public access
• Data visualization

Layer Name (click name for more information):
- Airports
- Animal Operation Permits
- Ambient Water Quality Monitoring Sites
- Beach Access Sites
- Bridges
- Bendite Monitoring Sites
- County Boundaries
- County Boundaries with Shoreline
- Community Colleges
Repository Development

- Downloading or acquiring “low hanging fruit”
- Tapping into current data flows
- Developing our own metadata when necessary
- Converting and preserving vector data in shapefile format
**Repository Status**

- Acquired 6+ TB of data with more on the way

- Disk space being used initially for “data staging”
  - Inventorying

- In the process of ingesting content into DSpace
  - Metadata generation
Data Preservation Like Fruit Desiccation?

- Complex data representations can be made more preservable (yet less useful) through simplification.
  - Conversion of various formats to shp
  - Image outputs (web services, PDF maps, map image files)
- Open GeoPDF standard
  - Analogous to paper maps
  - Combines data, symbology, annotation
  - More data intelligence than simple image
  - PDF content retained in addition to, NOT instead of data
**Engaging Spatial Data Infrastructure**

- **Cultural/Organizational solutions**: How do we make the data more preservable—and more prone to be archived—from point of production?

  - Engage and outreach to the data producer community and SDI
  - Sell the problem to software vendors and standards development
  - Find overlap with more compelling business problems: disaster preparedness, business continuity, road building, etc.
  - Discuss roles at the local, state, and federal level
SDI Role in Data Preservation

- Data inventories support content identification
- Metadata standards support discoverability and use
- Content standards support data interoperability over time and help eliminate semantic confusion
- Data exchange networks:
  - Minimize need to make contact
  - Add technical, administrative, descriptive metadata
  - Establish rights and provenance
NC Spatial Data Infrastructure: NC OneMap

- Next generation mechanism to coordinate and disseminate geographic information in North Carolina and interact with the NSDI.
- NC GICC
- Inventory for all geospatial data holdings – [http://nc.gisinventory.net](http://nc.gisinventory.net)
- Develop content standards for key data themes

One of the defined characteristics of NC OneMap is that “Historic and temporal data will be maintained and available”.

[Map of NC OneMap GIS Inventory]

GIS Inventory
- Registered
- Pending
Archival and Long Term Access Working Group

- Initiated by NC Geographic Information Coordinating Council in 2008 to address growing concerns of state and local agencies about long-term access to data
- Federal, state, regional, and local agency representation
- Key focus
  - Best practices for data snapshots and retention
  - State Archives processes: appraisal, selection, retention schedules, etc.
- Valuable outcome of NCGDAP – multiple parties and levels discussing data archiving on their own.
Archival and Long Term Access Working Group

- Final Report approved by NC GICC in November, 2008

- Best Practices for:
  - Archiving Schedule
  - Inventory
  - Storage Medium
  - Formats
  - Naming
  - Metadata
  - Distribution
  - Periodic Review
  - Data Integrity
  - Publicity

- http://www.ncgicc.org/
How to Recognize a Retention Schedule: Sample Schedule Item from NC OneMap

**ITEM** xxxxx, **STATE OWNED LANDS DATASET FILE**.

File includes datasets created to identify state-owned complexes, to define the exterior boundaries of state-owned complexes in North Carolina such as N.C. Department of Transportation (NC DOT) maintenance yards, state parks, state universities, etc. Datasets were created by the N.C. Department of Administration, State Property Office, in cooperation with the N.C. CGIA, and are responsible for providing new datasets. Datasets may represent different scale sizes and file formats.

**DISPOSITION INSTRUCTIONS**: Retain superseded dataset in office upon the receipt of new dataset. Transfer electronic records upon update to the State Records Center for immediate transfer to the custody of the Archives. Contact the Government Records Branch, Electronic Records Unit prior to transfer of electronic records.
Sample Proposed Local Schedule—County Management Schedule

**ORTHOMAGERY**

**DISPOSITION INSTRUCTION:**

GIS dataset: Permanent. Create a snapshot of dataset annually.

*Either:*

Transfer snapshot to NCOneMap according to established procedures, complying with standards and procedures adopted by the *North Carolina Geographic Information Coordinating Council.* (See *Geospatial Records, page __*)

*Or,

If retained in office permanently, your agency must comply with standards (for metadata, file naming, data sharing, and long term preservation) and procedures adopted by the *North Carolina Geographic Information Coordinating Council.* (See *Geospatial Records, page __*)
Potential Solution

County Orthoimagery

Ortho flights flown by contractor -> Data set Created

County GIS

NC OneMap FTP

NC OneMap

State Archives

DOT

Feds

Flood Mapping

Agriculture

Universities

DENR

Emergency Mgmt

TBD: Historic Orthos
NOAA PROCEDURE FOR SCIENTIFIC RECORDS
APPRaisal AND ARCHive APPROval

NOAA PROCESS STEPS

Identify Records
- Initiation of Request
- Receipt of Request

Appraise Records
- Designate Appraisal Team
- Preliminary Records Appraisal
- Formal Records Appraisal
- External Science Review

NOAA Formal Records Appraisal Questions

- Mission Relevant?
  1. Where in NOAA’s mission?
  2. Environmental or geospatial?
  3. Legal Mandates?

- Uniqueness, Provenance?
  4. Unique? Duplicated elsewhere?
  5. Relationship to other NOAA data?
  6. Authentic, reliable, unaltered, and usable?
  7. Original purpose? New purposes?
  8. Records value (now, future)?

- Nature of the Data?
  9. Volume (bytes), growing or static?
  10. Temporal and Spatial extent?
  11. Data format?
  12. Soley digital or does analog exist?
  13. Physical condition? Deterioration?
  14. Is information retrievable?
  15. Records location? Science center?

- Metadata?
  16. Does it exist?
  17. Conformity to standard format?
**NDIIPP Multi-State Geospatial Project**

- Lead organizations: North Carolina Center for Geographic Information & Analysis (NCCGIA) and State Archives of NC
- Partners:
  - Leading state geospatial organizations of Kentucky and Utah
  - State Archives of Kentucky and Utah
  - NCSU Libraries in catalytic/advisory role
- State-to-state and geo-to-Archives collaboration
- Archives as part of Spatial Data Infrastructure
OGC Data Preservation Working Group

- Formed Dec. 2006
- Engage archival community
- Find points of intersection with other OGC activities:
  - GML for archiving
  - Content packaging
  - Large scale data transfers
  - Time in decision support
Cultural: Changing Industry Thinking

- Is the geospatial industry “temporally-impaired?”
  - Lack of access to older data
  - Lack for tool/model support for temporal analysis
  - Metadata: poor support for changing data
  - Education: building class projects around available data (i.e., not temporal)

- Increased interest now in temporal applications?
  - Increased demand for temporal data?
  - Improved tool support: ArcGIS 9.2 animation tools; Geodatabase History, etc.
  - Emerging commercial market in older data
Conclusions

- “Supporting temporal analysis requirements” gets more attention than “archiving and preservation”
- Leverage existing infrastructure
- Current data sharing needs drive infrastructure improvements that help archiving
- Leverage business needs that are more compelling than preservation (e.g., continuity of operations)
- Facilitate stakeholder ownership of the solutions
- Mine state and local archiving innovations
Slide Presentation:

http://www.lib.ncsu.edu/ncgdap/presentations.html

Steve Morris  
Head, Digital Library Initiatives  
NCSU Libraries  
ph: (919) 515-1361  
Steven_Morris@ncsu.edu

Jeff Essic  
Geospatial Data Services Librarian  
NCSU Libraries  
ph: (919) 515-5698  
Jeff_Essic@ncsu.edu
Getting the Data in Motion

- Harvesting use cases for older data as part of outreach

Factors Driving Capture of Temporal Data

Survey of current archiving practice among NC counties and municipalities