GICC Archival and Long Term Access Ad Hoc Committee
Final Report
November 19, 2008

Background

The Archival and Long Term Access *ad hoc* Committee was established by the GICC in November, 2007 subsequent to the adoption of the GICC Data Sharing Guidelines. Recommendation #8 in the Data Sharing Guidelines report included an action on Archive and Long Term Access to geospatial data. Specifically, the report recommended that "Data producers evaluate and publish their long term access, retention, and archival strategies for historic data."

At the GICC meeting where the Data Sharing Report was presented, it was noted that the strong interest in archival and long term access guidelines indicated that more clarification and guidance was needed, and therefore, the *ad hoc* Committee was chartered.

The committee was seated in February. Anne Payne, Council Member from Wake County GIS, was named Chair. A roster of all members and staff is included on page 8. The first meeting of the committee was held on February 29, 2008 at the USGS offices in Raleigh. A list of all meetings held is included on page 8.

A description of related ongoing projects in North Carolina is included in Appendix B and an overview of the Archivists' Perspective on preservation of digital data is included in Appendix C.

General Description of the Issue

While key feature data layers such as land records, street centerlines, jurisdictional boundaries, and zoning are constantly changing, current data management practices commonly involve overwriting of older versions of data which are then no longer available. If retained, the data could serve several business purposes, including historical/cultural interests, support of legal proceedings, enforcement of environmental regulations, and aid in analysis of trends, as examples. Retention and preservation requirements and schedules, if they exist, are not considered nor included in up front data life cycle planning, budgeting, nor in work flow development, by local and state agencies. To the extent that data snapshots are retained, the archived data does not suffer well from neglect; long-term preservation will involve migration of data to supported data formats, media refresh, and retention of critical documentation. (See Appendix A – Issues Description for more detail)

Guiding Principles and Scope of Work

At an early meeting, the committee agreed to use the following to guide the development of its recommendations:

- Solutions should be doable, affordable and easy to adopt. Recommended practices should not place an undue additional workload on state and local GIS professionals. Retention strategies should be easy to accomplish as part of the agencies' normal workflow.

- An organized and structured approach for life cycle creation, management and sharing of geospatial content brings order and efficiencies to the retention and archival process.

- Recommended technical approaches should be designed to minimize the risk of loss of data over time.

- Archiving practices should be consistent with all other GICC-approved standards and recommendations. (Examples: Content Standards for Metadata; Data Sharing Recommendations).

- Recommendations should be consistent with electronic records guidelines, policies and requirements published by the NC Department of Cultural Resources – Archives and Records Section. (NCDCR).

- Existing retention policies and schedules of local and state agencies should be considered in the development of recommendations.

- Existing infrastructure should be employed as much as possible (example - NC One Map).

- Recommendations should address the following issues:

  - What content should be preserved?

  - How often should we create data archives?

  - Where are the archived data stored and will they be accessible?

  - What data formats, compression formats, and media? Should joined attribute data be included?

  - Who should be responsible for creation and long-term storage of archived data?

Scope of Report
The recommendations contained here are intended as a beginning guide for initiating data retention and archiving practices; not a comprehensive set of standards.

Initial guidelines may not equal the rigor of NCDCR policies and recommendations, but will reflect the spirit and intent of those guidelines. Recommendations are not to be used as a substitute for the record disposition requirements included in existing agency records retention schedules, but can be used to update those schedules.

Recommendations will not include the handling of older archived data, such as data already saved to tape or CD, but will provide a "day forward" approach. (However, if older data are retrieved, they should be preserved according to the "Best Practices" below.)

Backups/Disaster Recovery/Business Continuity vs. Archiving
BACKUPS periodically (nightly/ weekly) capture active datasets and are intended to provide a means to restore changing records that have been deleted or destroyed. The purpose of a back-up is to manage short term risk and address disaster recovery. Typically these snapshots are only retained for a few days or weeks before being overwritten by newer snapshots.

ARCHIVING data, on the other hand involves the long term collection and maintenance of data snapshots retained permanently that can be utilized to help manage long term risk (i.e. regulatory/ legal requirements) while allowing ongoing access to authentic historical data for the purposes of analysis or cultural preservation.

General Best Practices for Geospatial Preservation
The single most important thing that GIS producers should do to assist in the archival process is to organize and document their data holdings and databases.

Archiving Schedule
The archiving schedule for geospatial data layers should be based on frequency of update and will be based on business drivers identified by individual agencies; minimum and preferred frequencies and recommendations are included in the frequency section below.

Records retention schedules, as agreed upon by the agency and DCR, will vary in content and the frequency with which to transfer data. While some GIS data are constantly created, other data are created less frequently. Schedules reflect the business practices of units, the purpose for which the data has been created, and the short-term and long-term value of the data. At a minimum, annual snapshots of archival GIS data should be retained. As a safeguard, agencies should transfer GIS data deemed archival in the records retention schedule to the State Archives annually. GIS data producers should see www.ah.dcr.state.nc.us/records and consult their unit's records retention schedule.

Inventory
Participation in the NC OneMap Inventory is strongly recommended. By participating in the Inventory, via the Ramona tool, data producers provide data availability information, contact names, minimal metadata, information on rights, technical environment, and in some cases future development/maintenance plans. Historical or superseded data should be documented and included as part of the Inventory process.

Storage Medium
On-line network disk drive storage of archival data is strongly recommended. This storage method holds several preservation advantages if good IT practices are in place in the organization. Such advantages include: 1) data are available online anytime; 2) regular, consistent backups are made of the data; 3) off-site storage provides a secondary, secure source; and 4) media currency (due to upgrades) is ensured. Geospatial data creators should engage their IT organizations in planning for archival of geospatial data.

If necessary, when choosing an off-line storage media for archiving, storage media formats that are open standards (CD-ROM,DVD-ROM, LTO tape, etc.) that can be read by multiple vendor's devices give the most assurance for later archival data recovery.

If on-line storage of uncompressed archival raster data poses a capacity problem for an agency, resources of other organizations (such as state, regional, or federal agencies) should be considered as an alternative primary repository. Those same resources should be considered to serve as a secondary or tertiary repository.

NCDCR recommends a minimum of 3 copies as an archival best practice: a preservation master, an access record, and a backup. NCDCR also recommends that you always work from a copy when migrating or making changes to mitigate the risk of data loss.

Using a data conversion or other data delivery vendor as a permanent data repository is not recommended due to possible unreliability of the business stability of a private company.

There are service providers that support business continuity needs; however, those providers do not necessarily support true archival processes. The committee recommends that each agency engage their IT unit in the development of an internal archives storage medium strategy.

Formats
The following criteria were used in developing format recommendations: The format: 1) should be publicly documented, 2) should not be proprietary nor have any intellectual property restrictions associated with it, and 3) should be supportable in its non-translated form by existing and readily available software tools.

Vector Data

Shape files currently provide the most open, widely accessible, broadly achievable archival format for government agencies. Compression of data is not recommended. An agency whose vendor of choice is not ESRI should also maintain archival data in their native vendor format as well as in shape files. Tabular data from dbase tables, spreadsheets, and other external data sources that have linkages to the vector spatial data layers also need to be captured in any archival methodology.

Shape files are the strongly recommended archival format. However, if an agency chooses to archive ESRI coverages for retention of annotation or for other reasons, they should be archived in the e00 export form, due to the fragile nature of multi-file coverages, especially when being moved.

Note: For vector datasets the current most commonly used formats are proprietary data formats such as geodatabases, ESRI Personal and File. While these data storage technologies are commonplace, the data format has not been publicly documented to this point. While the vector data layers can be exported to ESRI shapefile snapshots, the internal constraints to attributes and linkages between data tables cannot be saved in this way. Although not addressed specifically in this report, the issue of archival of geodatabases and their associated intelligence and relationships must be addressed by the geospatial community.

Raster Data

GeoTIFF (or TIFF with world file) is the recommended retention format for raster data; it is open, available in the public domain, is non-proprietary, and is used and supported by a wide range of users and data providers. Compression of data for archiving (MrSID) is not recommended, due to the risk of lossiness; also, there is no Open Source implementation of the MrSID format. MrSID files should be maintained as an "access" copy of historic data, if needed, but should not serve as the official archival copy.

Consistent use of file naming schemes in the community facilitates the development of an efficient archival process. Data should be named in such a way to make the characteristics of the data easily discernable from the name. Such information should include (at a minimum):

- Keyword for jurisdiction or geographic extent
- Keyword for theme content or layer name
- Date created

Example: WAKE_PARCELS_2008_01

Files that are packaged together for access should be named following the naming convention outlined above. Files that should remain together should be packaged in the same folder. Avoid packaging other files that do not belong or that are extraneous.

Compression or archive files (e.g. zip, gzip, and tar), although convenient for packaging and distribution, add another potential point of failure in data retrieval. Also, zipped files pose a problem for the receiver of those files if the receiver wishes to unzip them automatically. If the files are identical in name to files currently on their system, the old files will be overwritten. If compressed/archive files are reserved for convenience or because of space considerations, an uncompressed version of the data should be archived periodically, if less frequently than the compressed data. The uncompressed data could be the "dark" archive (off-line, not readily accessible and protected).

Metadata Issues
Producers should develop and maintain fully compliant FGDC standard metadata. Archived data should always include metadata; the metadata should include the software version used to produce the data and document the naming convention used for archival data. Including metadata in an archive enhances the value of the archived data in legal situations.

Distribution Availability
Agencies should make an effort to make archived data readily available to other agencies and the public. If public data download of archival data via the web is not feasible or desirable, an effort should be made to make the public aware that the archive exists and provide a methodology for distribution. NC OneMap should be used as a conduit for data access to archival data just as it is to current data.

Periodic review of policies, data integrity
Jurisdictions should conduct an annual archival/retention policy review including content, format media, frequency and all other aspects of retention and archiving practices. Mechanisms that ensure data integrity over time such as checksums or digital fingerprinting should be employed.

Each agency should assign a responsible party/position for assuring long term accessibility and long term preservation. The agency responsible for geospatial data collection should communicate/coordinate geospatial preservation policies and procedures with the person within the agency who is officially assigned the responsibility of records preservation (city/county clerk, chief records officer)

Publicize Agency retention and archival schedules/practices
Organizations should publicize geospatial records retention schedules and archival practices on-line.

Frequency of Capture for Local and State Agencies
Preservation practices should be based on any applicable required retention schedule and on enterprise or external retrieval needs as documented by business needs of the agency.

A suggested retention schedule* for local government is shown in the table below:

Layer Name	Preferred Frequency	Minimum Frequency	Include Attributes?
Parcel	Quarterly	Annually	Yes
Street Centerlines	Quarterly	Annually	Yes
Corporate Limits	Quarterly (or as modified)	Annually	Yes
Extraterritorial Jurisdictions	Annually	Annually	Yes
Zoning	Annually	Annually	Yes
Address Points	Annually	Annually	Yes
Orthophotography	When Created	When Created	No
Utilities	Annually	Annually	Yes
Emergency/E-911 Themes	Annually	Annually	Yes

*This is not intended to be an exhaustive list – other data should be added to meet the agency's individual needs

State agencies create and maintain many individual geospatial data layers that are unique to each agency. It is difficult to provide a comprehensive list that addresses the individual frequency of capture scenarios and, therefore, it is recommended that state agencies adopt the practices described in this document and work with the Archives and Records Section to evaluate their records retention schedules and update them for geospatial content.

Other Key Recommendations
The committee also finds that several key actions should take place with the support of the GICC and its membership:

    a. Update the records retention schedule for NC OneMap

    b. Develop the capacity and expertise at the Archives and Records Section to guide development of agency records retention schedules for geospatial content.

    c. Develop a plan and implement steps at the Archives and Records Section to handle ingest of geospatial content as prescribed in emerging schedules. This plan should evaluate various geospatial content harvesting and ingest scenarios, including the leveraging of the NC OneMap clearinghouse and data

work flow as a content transfer point between data stewards and the Archives and Records Section.

Conclusions

Geospatial archiving and retention is a community wide problem. It is important to note that all jurisdictions have to handle public records with a retention plan and geospatial data are a public record. Agencies should work with Records officers and IT on retention schedules. Business case development is a key to justify the additional investment of resources.

The geospatial community is at the early stages of tackling this problem and these recommendations are intended to be initial, practical steps. The committee acknowledges that there are outstanding issues related to geospatial data preservation that are not addressed (or are only mentioned briefly) in this report. It is recognized that this is an ongoing effort and that additional guidelines will be needed.

Work by the GICC, the Archives and Records Section, and CGIA must continue. Some of the additional work will be included as parts of the other geospatial archive and preservation projects that are going on within North Carolina, such as the North Carolina Geospatial Data Archival Project, and the Geospatial Multistate Archive and Preservation Partnership, both described in Appendix B.

Committee Members

Name	*Organization*
Anne Payne, Chair	Wake County
Kathryn Clifton	City of Salisbury
Kelly Eubank	NC Dept of Cultural Resources
John Gallimore	Davie County
Tracey Glover	City of Fayetteville
Amy Keyworth	NC DENR – Water Quality
Bill Lefurgy	Library of Congress
Butch Lazorchak	Library of Congress
Scott Miller	Western Piedmont Council of Governments
Tom Morgan	NC Secretary of State – Land Records
Steve Morris	NCSU
Zsolt Nagy	NC CGIA
Doug Newcomb	US Fish and Wildlife Service
Thomas Parrish	NC Dept of Cultural Resources
Joe Sewash	NC CGIA
Ed Southern	NC Dept of Cultural Resources
Rebecca Troutman	NC Assoc. of County Commissioners

Committee Meetings

February 29, 2008
March 26, 2008
May 2, 2008
May 19, 2008
June 19, 2008
July 30, 2008

Appendices

A. Detailed Issues Description

B. On-Going Partnerships on Digital Preservation in NC

C. The Management and Preservation of Digital Media: An Overview from the Archivist's Perspective

APPENDIX A
Issues Description

Risks to Geospatial Data

There is a chain of possible failure events that can impede permanent access to data:

- To the extent that such data is saved, it may be stored in such a way that it is not recoverable.
- If the data is discoverable, policies may not have addressed the issue of what sort of access should be provided to older versions of data.
- If the data is accessible, there is a possibility that the storage media will no longer be readable.
- If the media is readable, the data files themselves may be corrupt.
- If the files are not corrupt, it is possible that the files will be in a format that is no longer supported by current software.
- If the format is useable, it is possible that the documentation needed to use and understand the contents of the data will not exist.

Unlike vector data, digital orthophotography is not typically at risk of overwrite, yet data from older flights are known to have become less discoverable and less accessible over time.

Archiving Challenges

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

- *Complex or Proprietary Data Formats* - Future support of data formats is in question. Due to the complexity of the content, migration between formats can lead to unacceptable data distortion and data loss.

- *Spatial Database Complexity* - Spatial databases, most notably the ESRI Geodatabase, are increasingly used for data management. These databases may consist of multiple individual datasets or "data layers," while also storing components such as behaviors, relationships, classification schemes, data models, or annotations. These complex databases may be difficult to manage over time due to the complexity of the data and uncertainty over long-term support of proprietary database models.

- *Fragility of Cartographic Representation* - The true counterpart to the old, preserved map is not the current GIS dataset but rather the cartographic representation that builds on that data. The representation, which is an important component of documented decision-making processes, is the result of a collection of intellectual choices and application of current methods with regard to symbolization, classification, data modeling, and annotation. This representation is often stored in proprietary project file, such as an .mxd, or in a complex PDF document in which the underlying data linkages have been severed.

- *Semantic Issues* – Widely ranging 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.

- *Frequency of Capture* - Update cycles for data resources are quite variable, ranging from "daily' updates for very transactional themes (parcels, addresses) to annual updates (imagery). The community at-large needs better guidance on desirable frequency of capture schedules for retention and archival purposes.

- *Metadata Unavailability or Inconsistency* – Inadequate metadata impedes discovery and use. Even if metadata exists, the metadata information is often asynchronous with the data (e.g., the metadata may not have been updated to reflect format or datum change). Existing metadata commonly requires some degree of structural normalization in order for the metadata to be incorporated in central catalogs or repositories.

- *Ancillary Data and Data Bundling* - Geospatial data is characterized by complex, multi-file formats. Ancillary files, which need to be bundled with the core dataset files, include metadata records, data dictionaries, additional data documentation, legend files, data licenses, disclaimers, and associated images.

- *Ephemerality of Data in a Web Services Environment* – As the geospatial industry increasingly shifts to web services-based access, data is becoming more ephemeral. Data does not get transferred when information can be gleaned or used through portals and viewers. It becomes more difficult to document the basis for decisions when operating in a web services environment.

Business Case

The value of historic geospatial data in accessible form has not yet been clearly articulated. Although data preservation is an emerging area of interest, it is a low priority topic. Data users are accustomed to working primarily with current data and, for lack of availability, have not yet discovered meaningful scenarios to utilize historic geospatial records and incorporate analysis into business activities. This scenario is perpetuated by data producers who are overly focused on current data, and as a result, overwrite data frequently. In addition, addressing digitally borne records as the "only" rendition of a

record has not yet become incorporated in agency administrative proceedings and planning. (Ex: Imaging analog documents commands higher attention)   The business case has not yet been made for the investment justifications necessary to successfully carry out a digital geo-preservation plan. Emerging business uses for older data include documentary support to legal proceedings, detection of permit violations, and analysis of shoreline change, land use change, and growth of impervious surfaces. Opportunities to engage the community and industry in building out the business case are desirable.

Access

While snapshots of older versions of data may be stored in agency archives, access is almost as a rule not available. Users are more likely to discover the value of historic data for business purposes if data can be relied upon and accessible on-line. It is also believed that the data that tend to survive are those that are maintained and accessible in an on-line service, since ongoing use of the data will spur user demand in the event that the data later becomes inaccessible.

APPENDIX B
On-Going Partnerships on Digital Preservation in NC

The work of the Archival and Long Term Access Committee was carried out in cooperation and collaboration with other geospatial archive and preservation projects that were going on within North Carolina at the same time. These projects are described briefly below.

North Carolina Geospatial Data Archival Project
The North Carolina Geospatial Data Archival Project is a partnership between the NCSU Libraries and NCCGIA, with Library of Congress under the National Digital Information Infrastructure and Preservation Program (NDIIPP). It is one of 8 initial NDIIPP partnerships. A few key elements of the project:

- A focus on state and local geospatial content in North Carolina;
- It is tied to the NC OneMap initiative, which provides for seamless access to data, metadata, and inventories;
- An objective is to engage and leverage NC OneMap collaboration and existing state/federal geospatial data infrastructures in preservation
- Local Government Frequency of capture surveys (2006; 2008)
- Involve NC State Archives agency in the extension work 2008-2009
- Goals:
    - Capture at-risk data for repository
    - Explore technical and organizational challenges for repository
    - Improve temporal data management practices of data producers
    - More efficient means of acquiring and preserving data for archival;
    - Progress towards best practices, including for long term access
    - Outreach and socialization of the problem

Geospatial Multistate Archive and Preservation Partnership
This project is a partnership among Archives, Libraries, and GIS state government agencies from Utah, Kentucky and North Carolina.   Project is co-lead by NC Archives and CGIA, under the National Digital Information Infrastructure and Preservation Program (NDIIPP). NCSU Libraries is a key partner in the project.

A few key elements of the project:

- Demonstrate and Document Best Practices for:
    - State-to-state transfer of geospatial content using spatial data infrastructure
    - Replication of content
    - Strategies to enable long term access and preservation of geospatial content
    - Engagement of States and national organizations such as NAGARA, NASCIO, NSGIC, COSA

- Key Issues
  - Content Selection (which layers needed)
  - Inventory of geospatial content
  - Integrating geospatial content into archival systems (process and frequency)
  - Managing/tracking content once archived
  - Flow of content between states
  - Role of Metadata in the above processes

APPENDIX C
The Management and Preservation of Digital Media: An Overview from the Archivist's Perspective

Digital records have taken over many of the functions that paper records once served. Like their older counterparts, digital records contain evidence of government responsibilities, citizen rights, public and private economic activities and financial transactions/obligations, scientific projects, and historical events and trends. The volume, complexity, and pace of the advanced technology, however, require the careful and consistent management of digital records if accountability and the preservation of digital records are to be assured. The integrity and accessibility of digital records also rest upon planning, documentation, and committed custodianship throughout their life cycle. In brief, to be available today, tomorrow, and into the next century, digital records must have both proper management and long-term (and in some cases, permanent) preservation. For digital records that are deemed permanent or archival, their durability needs to approach that of microfilm.

Best Practices for Archiving Electronic Records:

- Maintain at least three to four copies of the record. One copy should be designated as the preservation master; one copy should be designated as the access record; and one record should be designated as back-up. Having four copies allows margin should one copy fail. At least one of the duplicate copies should be stored off-site to ensure the information is preserved should an unforeseen disaster occur.

- Provide bit-level preservation storage of the original record. If the preservation strategy includes migration of data, keep original bits for future solutions.

- Work from a copy of the material when migrating or making changes as information may be lost during migration.

- Metadata, a digital fingerprint, and the data must be maintained and bundled together in order to preserve the integrity and admissibility of the data.

Best Practices—Policies and Procedures:

- Create and update policies and procedures defining proper development, maintenance, and use of the system. They should be available in electronic and hard copy print formats. These policies and procedures should include the metadata file required to interpret the records as well as technical components and characteristics necessary for reading, processing, accessing, using, and processing of records.

- Develop a digital risk management plan that includes regularly scheduled migration of archival digital objects to new media, formats, and technologies.

Best Practices—Integrity of Data:

- Metadata must be collected about the record and maintained with the record, either embedded in it, or stored separately. Descriptive metadata is used for the indexing, discovery, and identification of a digital resource. At a minimum, descriptive metadata should include creator, date, collector, and description. Land and property transactions should include the grantor/grantee names, PIN number, attributes if necessary, and description. Administrative metadata is information that is needed for the management of the digital object, which includes information regarding ownership, transfer information, access and display, and rights management. Preservation metadata that needs to be collected includes the file format, record type, e.g. tax map or correspondence, the operating system, software configurations, the rights/security, and versioning information.

- Security measures—Digital Fingerprinting

  o Information can be lost during transmission, migration, or when media break down or are corrupted. To ensure that the data does not and has not changed, you should perform a digital fingerprint procedure [e.g. digital certificates, Cyclical Redundancy Checksums or CRCs, and cryptographic hashing algorithms such as a Secure Hashing Algorithm (SHA)]. A digital fingerprint is unique to each document and verifies the integrity (unaltered state) of the document. When auditing the information or storage media, reproducing the digital fingerprint can determine if data has been lost. If you employ digital fingerprinting, retain the method by which it was applied so it can be recreated and compared to the original fingerprint.

  o Integrity of the record: If you elect to employ/allow digital fingerprints, have a migration strategy in place and a method to verify the fingerprint in the future so that it can be preserved and upwardly migrated. As part of the migration strategy, a digital fingerprint should be created at the beginning and at the end of the migration to ensure that the numbers produced from the algorithm are the same. If the two "fingerprints" match, then no error occurred during the transmission or migration.

- Security measures—Authority Rights. If special authority is needed to access the information, indicate who has that authority and for what data type (e.g. document or photograph).

- For admissibility of records, the content, context, and structure should be preserved.

Best Practices—Eye to the Future:

Practitioners of a trusted digital repository should take measures to keep abreast of and adapt to changing industry standards and technologies to ensure the survivability of the system. This includes preparing for the impending obsolescence of current formats and media.