Geospatial Multistate Archive and Preservation Partnership

in Partnership with

The Library of Congress

National Digital Information Infrastructure and Preservation Program (NDIIPP)

Interim Report: 2007-2009

March 2010

Prepared by

North Carolina Center for Geographic Information and Analysis

North Carolina State Archives

In partnership with

Kentucky Department for Libraries and Archives
Kentucky Division of Geographic Information
Kentucky State University
North Carolina State University Libraries
Utah Automated Geographic Reference Center
Utah State Archives

Table of Contents

Executive Summary

State governments have long understood the value of using geospatial information in decision making processes and planning efforts. State agencies have embraced the use of GIS information to analyze real world problems, to display and describe the physical world in a digital graphical format, and to provide more efficient and effective services to their citizens. State governments are also beginning to recognize the value of having access to older geospatial data as a resource to explore societal, environmental, and economic change over time. Compelling business drivers such as tracking changes in population, land, or vegetation over time, providing a cultural record of place; or the cost of having to recreate datasets that were not preserved are spurring users to seek out and use superseded geospatial content.

State GIS and archives organizations are making efforts to respond to this information need; however, they are facing serious obstacles. Traditionally, it has not been a priority for data creators to preserve superseded geospatial information or their resultant products. Older data is often overwritten or lost when more current information is received or as data is updated. As such, geospatial data is very much susceptible to either temporary or permanent loss. In addition, limited resources, diminishing budgets, and in some cases a lack of understanding by key decision-makers about the benefits of preserving geospatial data can stifle efforts to implement a formal preservation plan.

The Geospatial Multistate Archive and Preservation Partnership (GeoMAPP) was formed in 2007 to address the challenges associated with identifying, preserving and providing long-term access to temporally significant digital geospatial content in state and local governments and dynamic data that is "at-risk" of being lost when updates are made. The project is one of four initial state government partnerships funded by the Library of Congress' National Digital Information Infrastructure and Preservation Program (NDIIPP), and includes representatives from the geospatial and archives staffs of Kentucky, North Carolina and Utah. From November 2007 to December 2009, the three state partners worked together to investigate approaches for the preservation of and accessibility to superseded geospatial data, while concurrently engaging GIS data creators and archives leaders from local and state government within each state and nationally to raise awareness about geoarchives issues and solicit feedback.

During this initial phase of the project, the GeoMAPP team explored digital preservation issues in a number of topic areas, including: business planning, data inventory and metadata, appraisal and access, content transfer and ingest, and industry outreach.

The partnership established working groups that included both GIS and archives staff to address each of these areas. This model allowed project participants to contribute in areas that mirrored their expertise within their own state. While the partners took a unique "state-centric" approach to investigating the different topic areas, each was mindful of sharing and discussing individual findings and applying them to the collective questions the group was addressing. Accordingly, partners worked diligently to share their experiences, learn from each other and form project-wide generalized recommendations, best practices and standards.

Key observations from GeoMAPP's efforts to date:

- **Collaboration is a key component to establishing a unified approach to preservation**. Frequent formal or informal interactions between data creators, data custodians, and archives staff gives those involved the opportunity to build familiarity with each discipline's jargon and workflows, share experiences, and learn about positive and negative data management experiences. A high level of collaboration helps to prevent the duplication of efforts and adds value when implementing policies and systems and creating generalized recommendations, best practices and standards.
- **Create business case documentation to describe the value of temporal geospatial data and justify preservation investments.** The preservation of geospatial content will only prove valuable to legislators and financial decision makers when they understand that providing sustainable policy and funding support for preservation activities is vital and reaps financial benefits. This can be accomplished by developing a compelling business case that adequately captures both the tangible and intangible benefits of preserving temporal geospatial data, and identifies the risks of inaction.
- **Investigate the existing preservation landscape**. Surveys and data inventories are essential tools when first starting out. Surveys targeting GIS data producers as well as GIS and archival division leadership help to identify the current state of geospatial preservation within state and local government, and can also act as a vehicle for outreach. Inventorying holdings tells you what you have and where it is stored, both critical components for appraisal.
- **Make it official – create GIS specific records retention schedules to help ensure that data is being managed and preserved appropriately**. The archives can be proactive in its collaboration with data creators by providing them tangible guidance in the form of a retention schedule. A well-conceived retention schedule helps data creators identify permanent geospatial datasets as public records and provides guidelines on how to keep these data accessible for long-term future use. A formal records retention schedule compels data producers to think about what information they produce, which data need to be preserved and how to make these data useful to others.
- **Descriptive detail is a wise preservation investment.** Making sure that your geospatial data has descriptive metadata associated with it, assigning a logical file name to a dataset, and being aware of the data's format not only simplifies the ingest of the data, but assures future access and use.
- **Diligence in spreading the word about what you're doing can give others the tools and techniques they need to get started**. Whether it is developing a web presence, "hitting the road" to talk with local governments and regional professional organizations, or attending local and national conferences, outreach efforts can go a long way in sharing information that others may find valuable and can inform and improve your internal practices.

The initial partnership built a solid foundation by fostering relationships between archivists and GIS practitioners and by identifying a number of initial challenges with inventorying, appraising, transferring and ingesting geospatial data and creating unique approaches to begin to address these issues. Based on the initial success of the GeoMAPP project, the Library of Congress awarded additional grant funding to the GeoMAPP team to extend its investigation. GeoMAPP's research and outreach aims will continue in 2010 with at least two new full time partners and ten informational partners joining North Carolina, Kentucky, and Utah in the GeoMAPP 2010 effort.

Project Overview

Introducing GeoMAPP

What happens to superseded versions of dynamic critical state and local government geospatial data when updates are made? How do you identify what data need to be preserved? How does an archival repository appraise, ingest, preserve and provide access to this complex digital data for the long term? How does a state build and grow a program to address these preservation challenges in the face of financial and staffing cutbacks?

In November 2007 under the auspices of the Library of Congress' National Digital Information Infrastructure and Preservation Program (NDIIPP), state government archives and GIS practitioners from Kentucky, North Carolina, and Utah chartered a partnership to investigate these questions and other issues relating to the preservation of geospatial content.

This effort, which became the Geospatial Multistate Archive and Preservation Partnership (GeoMAPP), began with the aims of:

1. Identifying geospatial content within each state that is temporally valuable or is "at-risk" of being lost when updates are made;
2. Analyzing and providing recommendations on workflows in each state that affect the ability to preserve digital geospatial data;
3. Exploring the challenges of building collaborative relationships across organizational units within each state and across state lines;
4. Investigating technical challenges related to the inventory, appraisal, ingest, storage and preservation processes to ensure the long-term viability and accessibility of valuable digital geospatial data;
5. Researching business planning materials and practices that could be used to justify the creation, expansion or maintenance of a sustainable geoarchive;
6. Engaging relevant industry members from both the geospatial and archives communities to learn about products that could benefit the geoarchiving process and potentially encourage product changes that could benefit future archiving efforts;
7. Conducting outreach with geospatial data creators as well as archives and geospatial leaders, providing demonstrable models, practices and tools that can be shared with other state, local and regional government entities.

From the project's inception until the conclusion of its initial phase in at the end of 2009, GeoMAPP partners worked across state boundaries to research answers to these complex obstacles. The challenges were investigated and discussed during collaborative teams meetings and through the efforts of six subject area specific working groups formed to investigate issues relating to: appraisal and access, business case development, communications and outreach, content lifecycle and data transfer, industry outreach, data inventory and metadata.

Each partner introspectively evaluated its own processes in order to build tailored solutions to address the challenge of geospatial data preservation. These solutions often relied on the findings of the other partners and leveraged existing processes and workflows within each state to ease implementation. Drawing from

these individual state findings and collaborative project tasks, GeoMAPP's aim was to identify common solutions and consolidated findings that could be shared with other states and localities to help address the challenges of designing, implementing and sustaining processes and systems to help preserve geospatial data for future use and analysis.

Preservations Efforts Paving the Way for GeoMAPP

NDIIPP

In December 2000, the United States Congress authorized the Library of Congress to develop and execute a congressionally approved plan for NDIIPP. An initial $100 million congressional appropriation was made to establish the program, with the goal of building a network of committed partners throughout the country to develop preservation architecture with defined roles and responsibilities.[1] To address this goal, the Library developed *a Preserving Our Digital Heritage: Plan for the National Digital Information Infrastructure and Preservation Program*,[2] a document that explains how the plan was developed, who the Library worked with to develop the plan and the key components of the digital preservation infrastructure. The plan was approved by Congress in December 2002.

NCGDAP

Early in the program NDIIPP realized that born digital geospatial data was a critical component of the overall digital preservation strategy. Launched in the fall of 2004, the North Carolina Geospatial Data Archiving Project (NCGDAP)[3] was one of NDIIPP's initial grant projects and acted as a catalyst for discussion about the issues surrounding the preservation of state and local government geospatial content. NCGDAP featured collaboration between North Carolina State University Libraries and the North Carolina Center for Geographic Information and Analysis (CGIA) in partnership with NDIIPP. From 2004 to 2009, NCGDAP primarily focused on the collection and preservation of digital geospatial data content harvested from state and local government agencies in North Carolina.

Key NCGDAP objectives included:

1. Identification of available resources through the NC OneMap data inventory;
2. Acquisition of "at risk" geospatial data, including static data such as digital orthophotos as well time series data such as local land records and zoning data;
3. Development of digital repository architecture for geospatial data, using open source software tools;
4. Enhancement of existing geospatial metadata with additional preservation metadata;
5. Investigation of automated identification and capture of data resources from remote servers using emerging Open Geospatial Consortium (OGC) specifications;
6. Development of a model for data archiving and time series development; and
7. Outreach to the North Carolina GIS community about the preservation of geospatial data.

[1] National Digital Information Infrastructure and Preservation Program Information Bulletin, http://www.digitalpreservation.gov/library/program_back.html
[2] The complete text of the "Plan for the National Digital Information Infrastructure and Preservation Program" is available at http://www.digitalpreservation.gov/library/resources/pubs/index.html
[3] For more info about NCGDAP see: http://www.lib.ncsu.edu/ncgdap/

In addition to the lessons learned from the project's investigation of technical preservation challenges, one of the lasting impacts from NCGDAP has been the establishment of a dialog with data producers about the value of preserving geospatial data that is at risk of being overwritten or lost. NCGDAP's outreach included encouraging local government and state agency geospatial data creators to enter and manage information about their data holdings by registering and participating in the GIS Inventory. [4] The GIS Inventory has proven to be an invaluable source for information about data created within North Carolina and became a key starting point for the archives appraisal process for the state as part of GeoMAPP. NCGDAP's initial engagement with the GIS community within North Carolina and with national geospatial and archives bodies, not only provided a platform to communicate the issues of geospatial preservation, but also identified the need to continue and expand the scope of research and outreach efforts.

The NCGDAP project team also conducted surveys in 2006[5] and 2008[6] targeting municipal and county government GIS practitioners as a measure of outreach and to get a sense of preservation practices in local government. NCGDAP efforts ended in 2009, however the project identified several key preservation issues that continue to be explored and laid the groundwork for items to be examined by the GeoMAPP team such as business planning, records scheduling and transferring diverse content between states.

Preserving State Government Information Initiative

As the initial NDIIPP projects were ramping up in 2005, the Library of Congress sponsored a series of workshops involving all 50 states and three territories to discuss the issues surrounding the preservation of state government digital information. These workshops served as an opportunity for the Library to gather information and explore potential opportunities for engagement between NDIIPP and the states. The report that resulted from the workshops, *Preservation of State Government Digital Information: Issues and Opportunities,*[7] not only provided a detailed view of the formidable challenges facing the states but also identified collaborative opportunities.

NDIIPP prepared a call for proposals for state government partners that built on the initial set of NDIIPP investments in establishing a network of preservation partners. The call resulted in the Preserving State Government Information initiative, four partnerships of state government entities addressing the preservation of a variety of state and local government information. Following in the footsteps of NCGDAP's successful exploration of geoarchiving, in November 2007 the states of Kentucky and Utah joined North Carolina under an effort originally titled "the Multi-State Demonstration Project for Preservation of State Government Digital Information" the project later to be named GeoMAPP.

[4] http://www.gisinventory.net/
[5] http://www.nconemap.net/portals/7/documents/2006_LocalGovt_Geoarchives_Survey_Results.pdf
[6] http://www.nconemap.net/portals/7/documents/LocalGovt_GeoArchives_Survey_Results.pdf
[7] Library of Congress, "Preservation of State Government Digital Information: Issues and Opportunities," http://www.digitalpreservation.gov/library/resources/pubs/docs/preserving_state_gov_info2005.pdf [8] NC OneMap http://www.nconemap.net/.

GeoMAPP: A Collaborative Partnership at Work

One of the unique elements of the GeoMAPP partnership has been the distributed nature in which work has been completed with involvement from each of the project partners. While the partners took a unique "state-centric" approach to investigating geoarchiving, each was mindful of sharing and discussing their individual findings and applying them to the collective questions that the group was addressing. Collaboration in a multistate consortium is atypical of how problems are customarily addressed in state government, where tight staffing constraints often limit organizations to be narrowly focused on managing existing processes and addressing issues only when production challenges occur. Partners worked diligently to share their experiences, learn from each other and form project-wide generalized recommendations, best practices and standards.

The GeoMAPP partnership includes the following agencies:

- North Carolina Center for Geographic Information and Analysis (Principal Investigator)
- North Carolina State Archives- Government Records Branch (Co- Principal Investigator)
- North Carolina State University Libraries

- Kentucky Department for Libraries and Archives
- Kentucky Division of Geographic Information
- Kentucky State University

- Utah State Archives and Records Service
- Utah Automated Geographic Reference Center

GeoMAPP's pairing of archives and GIS staff from each state has enabled each of the state partners to establish or enhance the relationship between these organizations and to jointly investigate the challenges of preserving geospatial content. The "getting to know each other" process has featured building a familiarization of each discipline's terms and jargon while providing formal cross training between groups on both archival and GIS tools and technologies. By understanding each other's language and learning workflows and responsibilities, the state teams are better prepared to tackle the challenge of preserving geospatial content.

The following section provides background on each of the partner state's GIS and electronic records programs and organizations and addresses geoarchiving activities prior to the partnership's inception:

North Carolina

North Carolina is the principal investigator (PI) and lead state for the GeoMAPP effort. The North Carolina team pairs staff from the North Carolina State Archives Government Records Branch and the North Carolina Center for Geographic Information and Analysis (CGIA). The North Carolina State Archives is part of the North Carolina Department of Cultural Resources which has archival responsibility for records created by state and local government agencies in North Carolina. CGIA manages NC OneMAP, [8] North Carolina's geospatial data repository, and is responsible for project management, coordination, and contracts administration for GeoMAPP. CGIA began the project organizationally aligned with the state Department of Environment and Natural Resources, but in late 2009 was transitioned to the Office of Information Technology Services. North Carolina State University (NCSU) Libraries assisted GeoMAPP in a technical advisory role, sharing lessons learned from their experiences with NCGDAP and involvement with national geospatial organizations such as the Open Geospatial Consortium (OGC). [9]

Electronic Records Program Background

At the beginning of the GeoMAPP project, the archives had 2.5 staff dedicated to collecting and managing electronic records including:

- E-mail from the Superintendant of Public Schools and the Governor of North Carolina
- State Agency website archives (since 2005)
- Audio files from the State Senate
- Files from the state Office of Information Technology Services

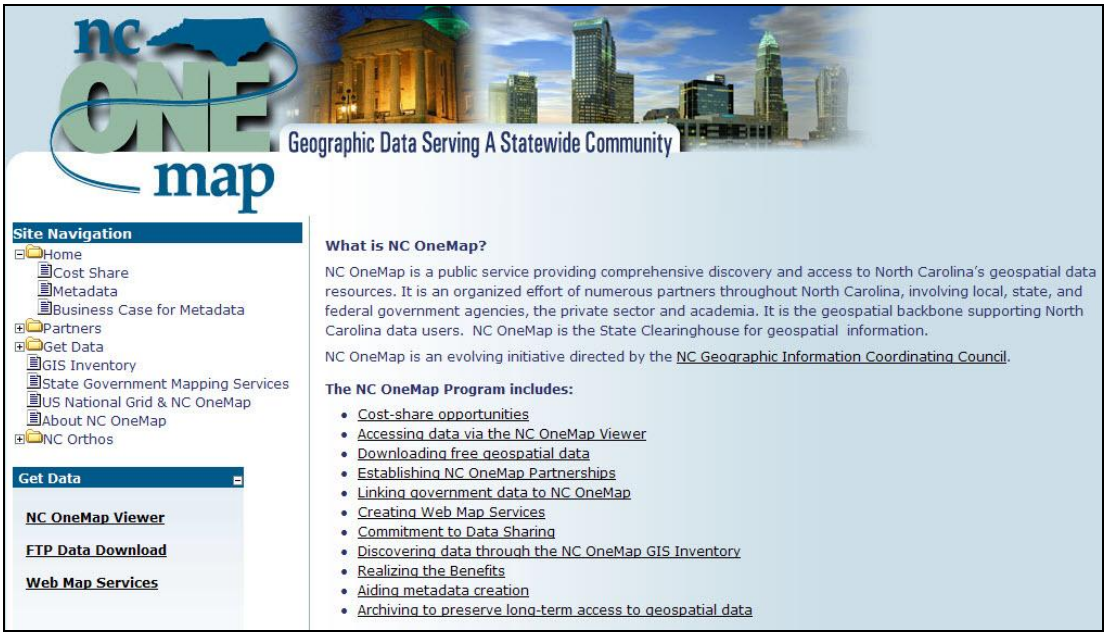
These data were typically stored on CDs or DVDs. Despite losing a staff analyst in late 2008 and the section's branch head in early 2009, the electronic records program has continued to grow. In January 2009, the archives received over 200,000 files (90 GB) from the outgoing Governor's administration and collected 50,000 e-mails, while also continuing to capture websites, accessioning senate audio files and actively participating in the exploration of ingesting and preserving geospatial content. Additionally, in late 2009 items such as archiving of state government-wide e-mail and the capture of state government maintained Web 2.0 tools such as Twitter© and Facebook© have also arisen as archival challenges for the state.

North Carolina's Geospatial Architecture

North Carolina's spatial data clearinghouse, known as NC OneMap, provides freely accessible data created by state, local and federal agencies via download in the ESRI shapefile format. Raster data is available in MrSID, JPEG, and IMG formats. In 2009, the site provided File Transfer Protocol (FTP) download access to over 110 vector and 125 raster geospatial datasets. NC OneMap's datasets feature

[8] NC OneMap http://www.nconemap.net/.
[9] For more info about the OGC, see: http://www.opengeospatial.org/.

Federal Geographic Data Committee (FGDC) compliant metadata records.[10] If a metadata record is not included when data is submitted for posting, staff will create a new metadata record with input from the data creator. The OneMap team will also enhance or refine existing metadata records transferred with datasets when they are missing critical information with input from the data creator. Before data is posted it is also opened and checked to assess file validity, dataset projection and geographic extent.

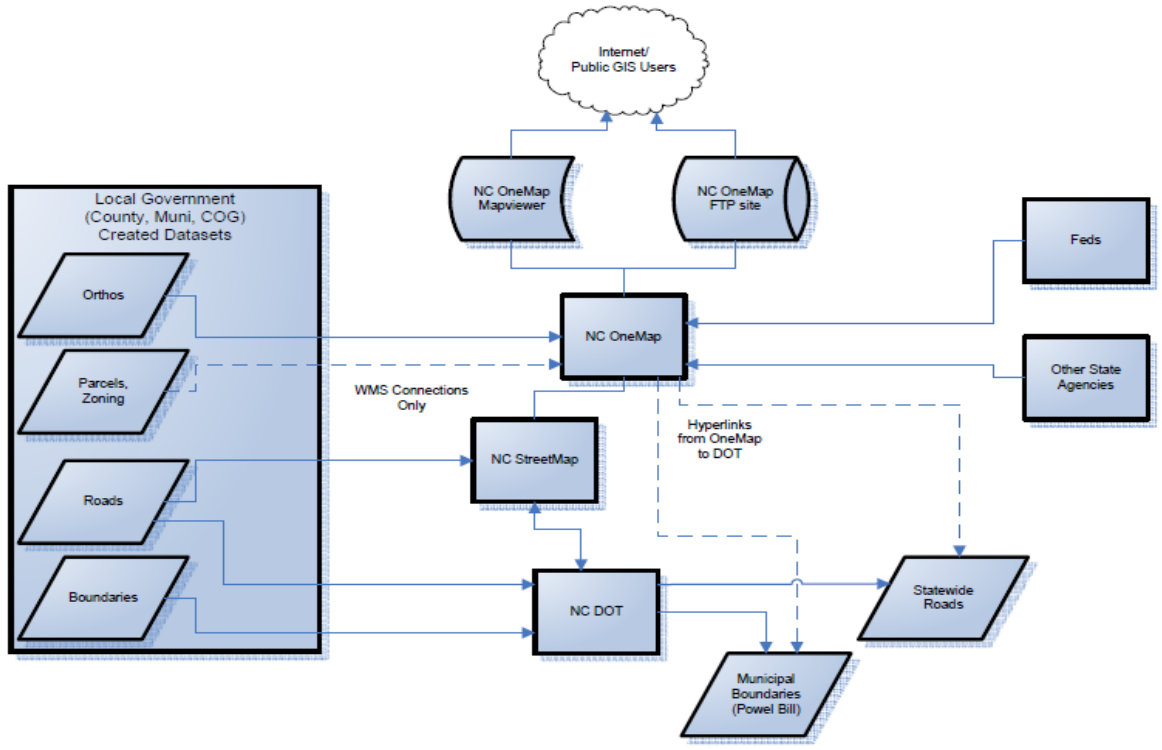


NC OneMAP allows users to locate and download geospatial content via a map viewer or FTP.

Despite having a robust centralized repository, most geospatial data in North Carolina are produced, maintained and hosted by data creators situated in a variety of state and local governments agencies, thus giving North Carolina a fairly decentralized approach to providing access to its geospatial content. NC OneMap uses Web Map Services (WMS) [11] to provide access to these remotely created and managed datasets via the Internet. In 2009, over 350 geographic data layers were accessible using the NC OneMap viewer[12] and NC OneMap had established relationships with over 100 partners who shared data either directly or via WMS, including a diverse mix of federal, state and local government agencies and academic institutions. More than 80% of these partners represent city or county government.

This decentralized approach to data hosting and management has made archiving this content and determining a location of capture a significant challenge for the North Carolina team.

[10] Geospatial metadata details on p.27
[11] NC OneMap WMS Data Catalog, http://www.nconemap.net/datacatalog/.
[12] NC OneMap Viewer, http://204.211.239.202/viewer/.



Mapping the complicated movement of geospatial content in North Carolina

Existing Geoarchives Activities

Before GeoMAPP's inception, there was no formal archiving of geospatial content by the North Carolina State Archives and the relationship between CGIA and the State Archives was in its infancy. However, CGIA and NCSU's involvement in the NCGDAP project and strong interest and support from the state's Geographic Information Coordinating Council (GICC)[13] had catalyzed a strong interest in preserving geospatial content in North Carolina prior to the GeoMAPP kickoff.

Prior to the Fall 2004 launch of the NCGDAP project, "archiving" of geospatial datasets in North Carolina was largely an localized effort driven by GIS data producers and often tied to regular back-up management practices or informal archives including personal collections of data stored on compact discs and tapes. There was little statewide awareness of the issue. Records retention schedules, which were adept at handling paper records, made little or no mention of digital geospatial records.

The GICC's early interest in preserving and providing access to superseded geospatial content was made evident in the May 2003 vision statement for the development of NC OneMap, in which it was noted as an essential characteristic of the emerging state mapping portal that "historic and temporal data will be maintained and available." Data archiving was formally addressed in 2007 in the findings of the GICC's Local/State/Regional/Federal Data Sharing ad hoc Committee. The Data Sharing Committee, comprised

[13] For more info about NC's GICC see: http://www.ncgicc.com/

of members from the local, state and federal government and university communities, focused on addressing the challenges that data producers face in providing access to their data. The Committee recommended that "data producers should evaluate and publish their long term access, retention, and archival strategies for historic data."[14]

Based on the findings of the Data Sharing Committee, the GICC created the Archival and Long Term Access Ad hoc Committee in November of 2007 to further investigate the issue of archiving geospatial data. In November of 2008 the group formally presented its findings to the GICC.[15] These findings included specific recommendations for data format, storage media, metadata, frequency of capture of the data, and next steps for the long-term preservation of geospatial content in the state.

Kentucky

The Kentucky GeoMAPP team is comprised of staff from the Department for Libraries and Archives (KDLA), the state's primary archival body, and the Department of Geographic Information (DGI) which manages the Kentucky Geography Network (KYGEONET), [16] Kentucky's geospatial data clearinghouse. The team also receives technical GIS training, consultation, and project assistance from Kentucky State University. Organizationally, DGI falls under Kentucky's Commonwealth Office of Technology (COT).

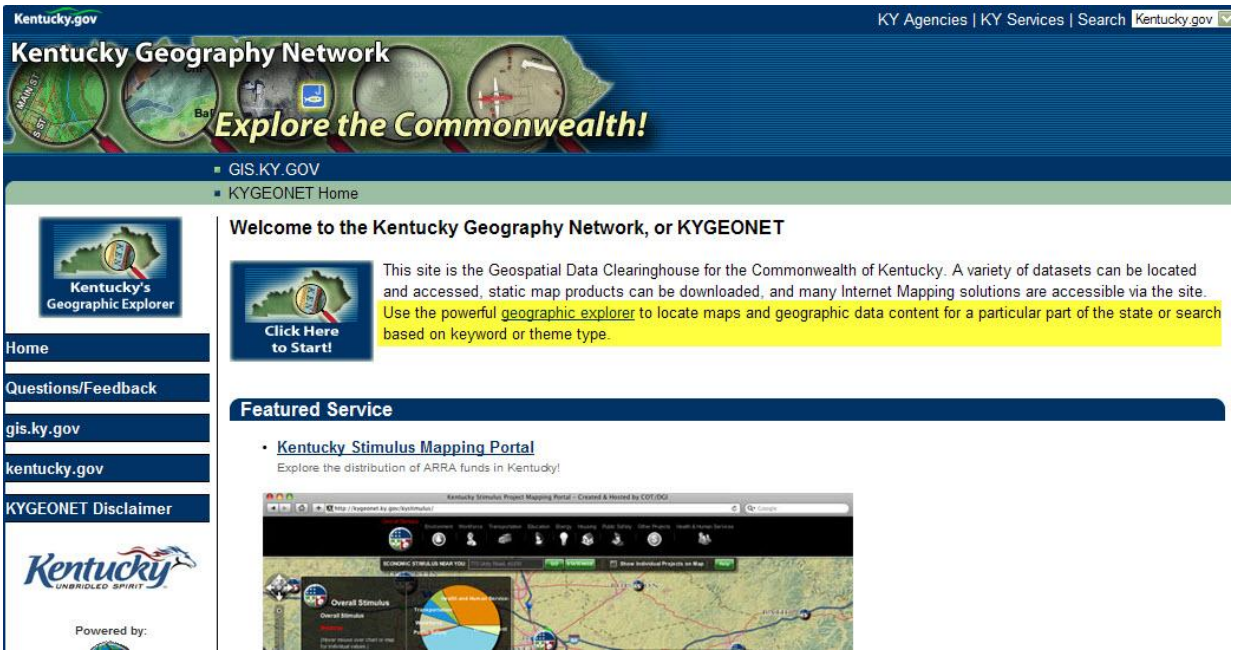
Electronic Records Program Background

At GeoMAPP's inception KDLA had 3 staff members accessioning geospatial data, e-mail, website snapshots, state publications, governor's records, and meeting minutes into their archive. GeoMAPP allowed Kentucky to continue expansion of its electronic records program through the financial support, sharing of ideas/techniques, and development of best practices, despite the loss of a team member during the project period. The team has developed a DSpace repository application that is housing GIS and other electronic records. The Kentucky DSpace repository stores shapefiles, small images and PDFs, and plans are in place to describe and reference file geodatabases and large image stores that are external to the DSpace instance. Throughout the project, Kentucky's electronic records holdings have continued to grow and the team is focusing on accessioning additional records.

Kentucky's Geospatial Architecture

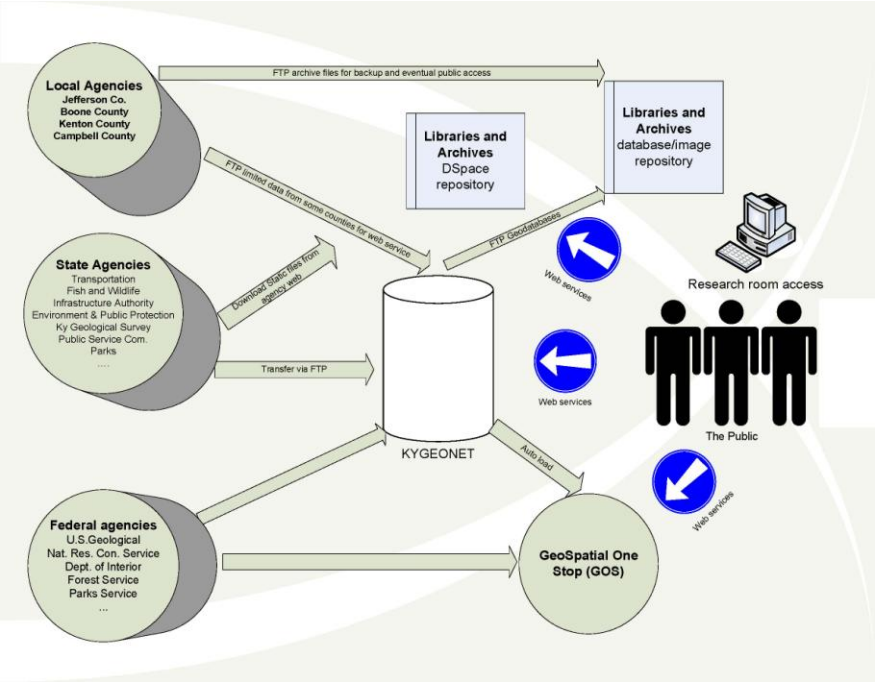
The Commonwealth of Kentucky takes a fairly centralized approach for their geospatial holdings and hosts data for local, regional, state and federal entities on the Kentucky Geography Network. All of the resources made available via the KYGEONET feed the Commonwealth's Enterprise GIS Databases, KyRaster and KyVector, which are managed by the Division of Geographic Information (DGI). These databases are accessed by hundreds of GIS users in State Government on a daily basis. There are no formal agreements in place nor do any mandates exist that require data producers to provide their geospatial data resources to the KYGEONET. Participation is voluntary; however, entities have chosen to contribute due to the exposure their data receives and the benefits that are realized from having the data accessible in a "self-serve" manner.

[14] The full data sharing report can be found here: http://www.ncgicc.com/LinkClick.aspx?link=156&tabid=306&mid=547
[15] The Archival and Long Term Access Committee's recommendations can be found here: http://www.ncgicc.com/Portals/3/documents/Archival_LongTermAccess_FINAL11_08_GICC.pdf
[16] KYGEONET, http://KYGEONET.ky.gov/



KYGEONET allows its users to locate maps and geographic content for a particular part of the state or by keyword or theme type.

In order for data to be ingested into the KYGEONET, geospatial data resources must include a minimum set of FGDC-compliant metadata. If the required metadata is not present, the data will not be ingested into the KYGEONET or the Enterprise Databases. In most all instances the data submitted for distribution is an ESRI shapefile or file and tile-based image datasets. Transfer of this data occurs via network shares, FTP, DVD/CD, and portable hard drives. One of the primary challenges the Kentucky team has faced in data acquisition has been with several regional agencies responsible for hosting local government data that charge for data access. This restricted access has limited the archiving efforts for this data, but participation in GeoMAPP has helped catalyze discussion between KDLA, DGI and the data providers.



How geospatial data moves within the Commonwealth of Kentucky

Existing Geoarchives Activities

Kentucky is unique among project partners as there was already an established working relationship in place between KDLA and DGI prior to the inception of the GeoMAPP project and data was actively being transferred to the archives. KDLA has been working with DGI and its predecessor agency since 2005, including producing records retention schedules and records transfers before GeoMAPP. This relationship has survived and flourished despite changes in leadership and absence of top level support due to the dedication of the DGI's staff to the principle of geospatial data value and the need for preservation measures. This relationship was initiated via the Kentucky Electronic Records Working Group and KDLA actively encouraged DGI and other holders of geospatial data resources to establish retention schedules for critical data. KDLA had 1 terabyte of storage in place prior to the project to help store their snapshots of DGI's vector data, and project funds allowed this capacity to expand to over 10 terabytes to handle extended vector holdings in addition to some raster imagery. The partnership with the GIS community which began before the grant has blossomed into an active data sharing and modeling relationship.

Utah

The Utah GeoMAPP team is comprised of staff from the Division of Archives and Records Service and the Automated Geographic Reference Center (AGRC). AGRC manages the State Geographic Information Database (SGID), [17] Utah's geospatial data clearinghouse. The Archives is a division within the Department of Administrative Services, while AGRC is part of the Department of Technology Services.

[17] Utah SGID, http://agrc.its.state.ut.us/

Electronic Records Program Background

Prior to kicking off the GeoMAPP effort, Utah was in the early stages of building an electronic records program. Selected records were submitted to the archives from a variety of sources, usually on compact discs placed in boxes with paper records. Utah Archives also received governors' records in electronic form and stored them on a hard drive. The files were typically desktop files, such as Word documents or spreadsheets. Additionally, the archives contracted with the Internet Archive to harvest state websites, but the archives have had only limited interactions with this data which is typically managed and harvested by the Utah State Library. Catalyzed by GeoMAPP project efforts the archives made a concerted effort to identify individual electronic datasets and record them in a catalog database.[18] The catalog functionality has expanded so it can be used for multiple formats including geospatial data. The archives staff has had ongoing discussions with its IT department with regard to preserving e-mail. The archives has also begun a pilot project with the state's Purchasing Division to classify agency e-mail messages and export them out of the existing proprietary e-mail system.

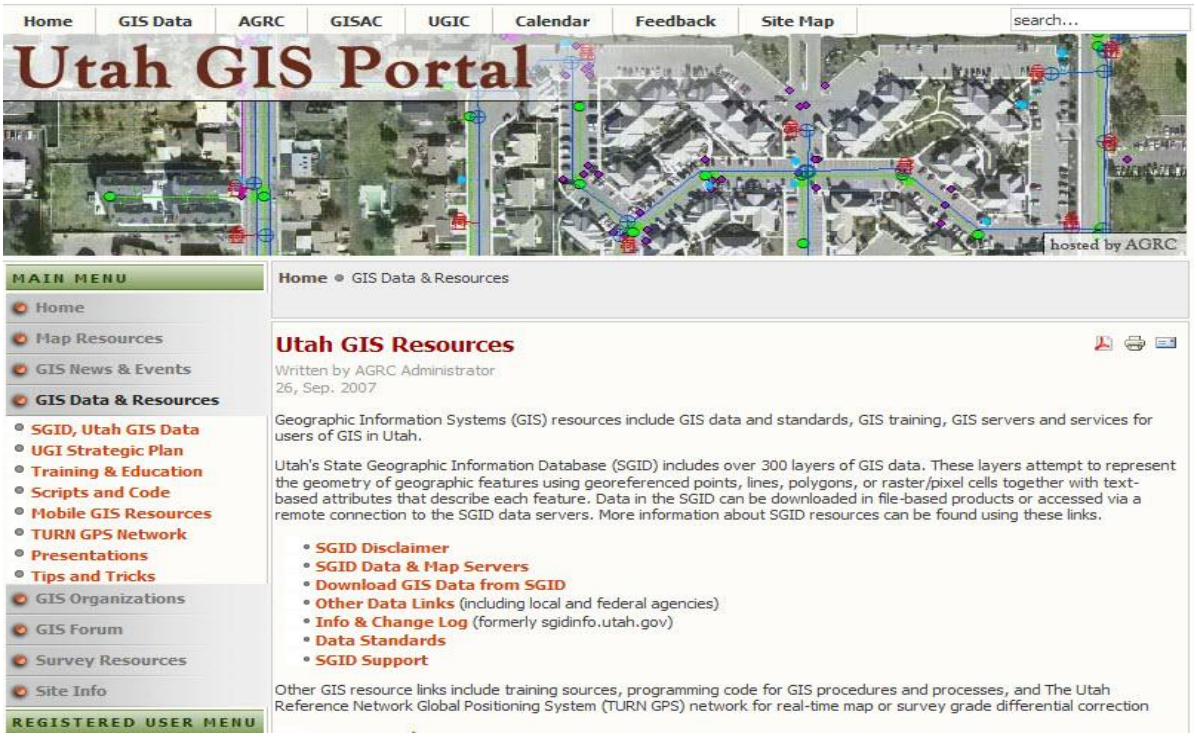
Utah's Geospatial Architecture

Utah began the project with a fairly federated approach to managing their state's geospatial holdings. Relationships between AGRC and state agencies and local governments were traditionally formed on a project-by-project basis. AGRC has managed large road and parcel data collection efforts, which has allowed for unprecedented opportunities to interact and build relationships with county governments. Many of the state agency relationships are built between people in each office. Because of these outreach efforts, the reputation and purpose of AGRC as a data clearinghouse has encouraged participation without prompting.

AGRC hosts any public or private data that data producers are willing to share, whether this data is from the local, federal or state level. The data focus has also shifted[19] for the SGID from being project driven to being more varied in type and focus.

AGRC receives and ingests raster and vector datasets ensuring that metadata is both complete and FGDC compliant. AGRC staff will enhance or refine existing metadata records transferred with datasets when they are missing critical information with input from data creator. If metadata is absent, AGRC will contact the owner or steward of the data so that the metadata is completed to meet FGDC standards. Additionally, the AGRC staff opens and checks the dataset to assess file validity, dataset projection and geographic extent. Once the dataset and metadata record have been validated, the data is made available for public access via FTP. The data listed can be downloaded for free and can be used by anyone without restriction.

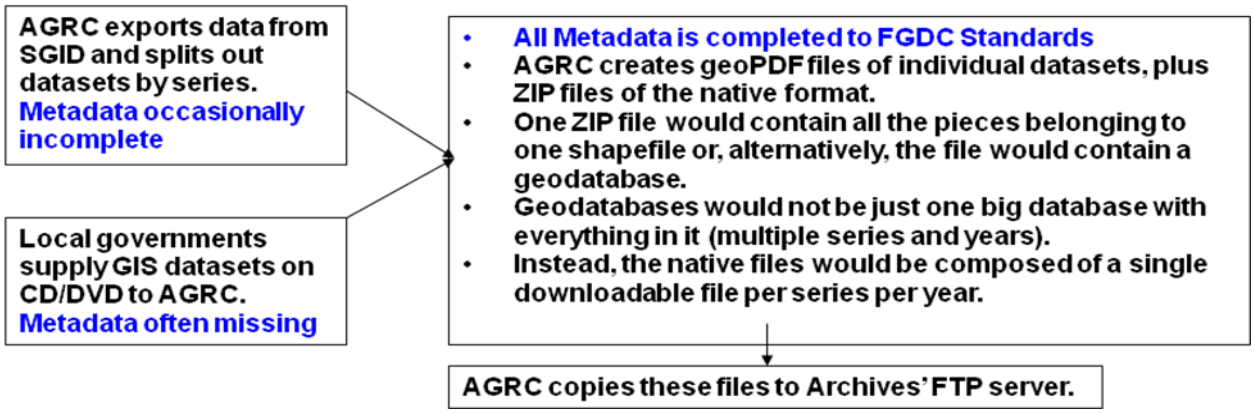
[18] Utah Archives e-records catalog: http://images.archives.utah.gov/cdm4/search.php
[19] The SGID Legislative mandate can be found here: http://www.le.utah.gov/UtahCode/getCodeSection?code=63F-1-507



Utah's GIS Portal enables its users to locate geospatial content by dataset type (vector/raster) and then by ISO Category (vector) or imagery product (raster).

The SGID is required to provide an accurate representation of all civil subdivision boundaries of the state. Each state agency that acquires, purchases, or produces digital geographic information data is required to inform AGRC about the existence of the data layers and their geographic extent and allow AGRC access to all data classified public. Additionally, the State Tax Commission annually delivers data relating to the creation or modification of the boundaries of political subdivisions. AGRC has also created a data sharing Memorandum of Understanding (MOU) with the federal government that has accepted by 13 federal agencies.

GeoMAPP has enabled the Archives and AGRC to extend their relationship with local data creators by supporting travel to localities and regional agencies statewide. During these visits, data were inventoried and added to the GIS Inventory, and targeted data were copied and transferred to the SGID and the Archives.



Utah's process for capturing data for the archives

Existing Geoarchives Activities

Prior the kickoff of GeoMAPP, there was no formal relationship between AGRC and the State Archives. There had been a few high-level interactions between the agencies including the appointment of a records officer within AGRC, but little had been done to address records management within the agency.

AGRC had also begun work on the implementation of a continuity of operations plan (COOP) when it was introduced to GeoMAPP. It was thought that the needs of AGRC and the needs of archives could both be met by aligning the disaster recovery mission with data preservation. As the project progressed, both agencies realized that the long-term plans for geoarchiving would benefit from the establishment of a formal agreement between the agencies. A Memorandum of Understanding (MOU)[20] was signed between the Utah Division of Archives and Records Service and AGRC outlining each agency's responsibilities to preserve the long-term availability of geospatial data.

Surveying the Geoarchiving Landscape

One of the first substantial efforts undertaken by the GeoMAPP team was to conduct surveys targeting GIS data producers, as well as archives and GIS leadership with the dual objectives of identifying the current state of geospatial archiving and becoming familiar with the preservation landscape both nationally and within state government. These GeoMAPP surveys were based on two previous survey efforts targeting local government GIS creators in North Carolina as part of the NCGDAP project in 2006 and 2008.

In the summer and fall of 2008, GeoMAPP released two national surveys targeting state government GIS leaders affiliated with the National States Geographic Information Council (NSGIC) and archives professionals with active membership in the Council of State Archivists (CoSA) and the National Association of Government Archives and Records Administrators (NAGARA). These national organizations were targeted due to the fact that their memberships include State Geographic Information Officers and State Archivists, the chief geospatial and archives decision makers in state government. The

[20] The Utah MOU can be found here: http://www.geomapp.net/docs/2009_AGRC_Archives_MOU.pdf

team also surveyed state agency GIS data producers in North Carolina and local government GIS practitioners in Kentucky. The invaluable data collected from these surveys not only informed the project team about the existing state of geoarchiving, but also served as a vessel for outreach to survey recipients by highlighting geoarchiving challenges and practices, and mutually informing GIS and archives professionals that there was a group keenly interested in and dedicated to investigating the preservation of geospatial data.

National Survey Highlights

After the formation of GeoMAPP when partners started building a common language and determining the capabilities within each state, the partners decided it would be useful to find out the status of geoarchives programs in other states in part to spread the word about the project, but also to see if there were things that were being implemented in other states that could benefit the work of the project. To this end, the team developed two similar surveys, one targeting state government archivists and records managers, the other distributed to state Geographic Information Officers (GIOs) and other state government GIS leaders. The surveys featured some questions unique to the target audience and others that were common across surveys for ease of comparison.

CoSA/NAGARA Survey[21]

The CoSA/ NAGARA survey was launched via e-mail to the CoSA and NAGARA listservs and was advertised heavily during a combined annual meeting of the two groups. The survey featured 12 questions with the aim of capturing data about the maturity of states' electronic records program, familiarity with geospatial data, interactions with GIS staff, and archival business drivers. Some key findings:

- 50% of respondents indicated that their state had an established electronic records program, 44% were in the planning or beginning phases of developing a program, while 6% did not have an e-records program nor had plans to implement one;
- Only 17% of respondents had **no** exposure to GIS data or staff. The remaining 83% had various levels of familiarity with GIS ranging from having a relationship with state GIS staff, to actually having an established program to archive geospatial data;
- Overall, 27% of the respondents specified that their archives were actively receiving and archiving geospatial data;
- Major business drivers for the preservation of geospatial data included change analysis, historic mapping, and cultural preservation, while records management policy placed forth out of a list of eight possible drivers.

A best practices/recommendations section at the end of the survey highlighted some of the challenges archival organizations were facing while trying to develop geoarchiving programs:

- "Management of GIS systems/data is largely within the individual agencies. The need for a centralized approach has been discussed, but progress has been limited;"

[21]CoSA/ NAGARA survey questions can be found here: http://www.geomapp.net/docs/geomapp_survey_nagara.pdf. Raw data results can be found here: http://www.geomapp.net/docs/geomapp_survey_nagara_rawdata.pdf.

- "Some agencies 'share' GIS systems and data and it is difficult to get a handle on how/when/what to preserve;"
- "The 'use value' of the data is of primary concern for justifying the cost of preservation. Outreach and collaboration are key components for a sustainable program."

NSGIC Survey[22]

In parallel with the CoSA/ NAGARA survey, GeoMAPP reached out to the state GIO community by launching a more geo-centric, 23 question survey distributed to the NSGIC listserv. The NSGIC survey had a variety of questions addressing geoarchiving status, familiarity with retention policies, several data and system centric questions as well as a request for business drivers and best practices. Key findings included:

- 55% of respondents indicated that their state has some form of a centralized archiving system. However, many of these formal geoarchiving programs were in their infancy as only 19% had been up and running for over 5 years and 57% were still in development;
- Only 29% of those archiving, have their data online and available for open public access;
- Orthoimagery, governmental (municipal, county, state, etc) boundaries and street centerlines were the three most commonly preserved datasets;
- The majority of respondents (62%) were not aware if geospatial data were included in agency records retention schedules, while one third said that geospatial data were addressed in **some** but not all agency schedules;
- The primary business drivers for archiving included: records retention policy, followed by utility of the data for change analysis and historic mapping.

A best practices/recommendations section at the end of the survey highlighted some of the challenges geospatial divisions were facing while trying to develop geoarchiving programs:

- Archiving geospatial data was determined to be a generally low priority issue with little funding support;
- One of the respondents noted that archiving is something that had been discussed within their state but had not been uniformly implemented across the board for state agencies or local organizations. They acknowledged the benefit of archiving some/all data layers, however noted that no standards or procedures existed to guide how frequently it should be done and could think of no enticements to do so.

State Specific Survey Highlights

Following the two successful local government surveys conducted by the NCGDAP project and launched in parallel with the two national surveys, Kentucky and North Carolina also conducted state-centric surveys targeting geospatial data creators within their borders.

[22] NSGIC survey questions can be found here: http://www.geomapp.net/docs/geomapp_survey_nsgic.pdf. Raw data results can be found here: http://www.geomapp.net/docs/geomapp_survey_nsgic_rawdata.pdf.

Kentucky Local Government Survey[23]

Borrowing heavily from the 2006 and 2008 North Carolina local government surveys, the Kentucky local government survey was conducted to broaden the Kentucky team's awareness of sources of geospatial records that were not being included in the state's geospatial repository, the KYGEONET. Although the survey was launched with assistance from two statewide local government organizations, the response rate was fairly low (18 total responses). Key responses came from two metropolitan area consortiums, some local area planning units and a handful of county governments. In Kentucky, local governments have had a history of not sharing their GIS records, largely to protect the cost recovery value of the records for resale, which may have negatively impacted the response rate.

While the low response rate made it difficult to draw conclusions about the extent of archiving by local government GIS data creators, the Kentucky team was able to glean some valuable information. While a majority of respondents (14) had archived files dating back over a year, few (3) had files older than five years old. While the majority of respondents did not indicate a frequency of capture, those that did generally captured files at least once per year with address point and utilities being the layers captured with the greatest frequency. The survey contributed to the development of MOUs between the archives and the two major metropolitan geospatial consortiums.

North Carolina State Agency Survey[24]

Armed with the findings from the two NCGDAP local government surveys, the North Carolina team wanted to branch out and find out more about data archiving practices in North Carolina state government. The result was an 18-question survey attempting to discover information about archiving status, familiarity with retention policies, data and system management questions as well as questions dealing with business drivers and best practices. The team received 58 responses from 6 state departments with multiple responses coming from agencies within the environmental, transportation, and commerce departments, who are also three of the larger GIS producing agencies. Some key findings include:

- 50% of agencies reported that they were archiving data, 26% were not, and 24% were not sure of their agency archiving practices;
- The most commonly archived data included: Biological/Environmental, Hydrologic, Boundary/ Ortho, Address, and Geodetic;
- 40% of respondents were either familiar with or were responsible for following their agency's records retention schedule. However only 19% of those archiving said that geospatial data was included in their agency records schedule;
- Primary business drivers included: historic mapping, records retention/ archival policy, change analysis, and legal or statutory purposes;
- The best practices section yielded several comments about issues related to data organization and tracking. Several of the respondents reported that not only was it difficult to locate the archived

[23] Kentucky agency questions and raw data results can be found here: http://www.geomapp.net/docs/KY_Local_Gov_Survey_9-29-2008.pdf and here: http://www.geomapp.net/docs/KY_Local_Government_survey_results.pdf.
[24] NC State agency questions: http://www.geomapp.net/docs/geomapp_survey_state.pdf Findings can be found here: http://www.geomapp.net/docs/StateAgency_GeoArchives_SurveyResults_NC.pdf.

data that they needed, but they were also not sure if the data that has been preserved was still valid or useful;

- One respondent said that it was difficult to find space to store archived data electronically, revise metadata and catalog amid all of the other demands of their daily duties. However they found that preserving superseded geospatial data is useful for some of the key responsibilities of their agency.

Conclusions

These surveys, while not meant to be scientific, did provide a useful overview on the state of the landscape regarding geoarchiving in both the archival and geographic divisions of state government, and gave the project a baseline to compare the results of their efforts going forward. The responses from the collective archives, GIS data creator and custodian communities highlighted an awareness and desire to preserve geospatial data, but also noted a lack of resources or understanding to formally preserve this content.

The archiving rates for each of surveys have promise. However it is suspected that while many of those who responded that they are "archiving" data may actually be saving copies of superseded data, it is likely that these are informal data stores with little management and validation of the data. This possibility seems especially strong in local and state government agencies where personnel and hardware resources are limited.

GeoMAPP Project Working Group Activities

Working Group Background

After the formative period of the project where the project participants were familiarizing themselves with their counterparts within their states and across the partnership, the project members developed an organized way to plan and track the work of the project and equitably distribute this workload. In the second half of 2008, the team members developed a consolidated workplan of all the items that the team wanted to explore during the project along with proposed deliverables and corresponding deadlines. The group formed six cross-functional working group teams with membership from each state to address critical areas of investigation for the project. The GeoMAPP working groups included:

- Business Case
- Inventory and Metadata
- Appraisal and Access
- Content Lifecycle and Data Transfer
- Communications
- Industry Outreach

The working groups became the backbone of the GeoMAPP effort and tasks included in the project workplan were divided among these groups. Each working group designated a team lead that was responsible for reporting group findings to the larger project, and interpreting and managing tasks assigned to the group. As anticipated, the working group approach to handling project tasks was fairly chaotic at its inception, but proved to be a good way to have individuals across the project collaborate on similar tasks and to collectively tackle the concerns of the project.

This section of the report details the background, focus areas, key findings, lessons learned and recommended practices from each of the GeoMAPP subject area working groups and discusses their path forward for GeoMAPP 2010:

Making the Case for Preservation – Business Case

<u>The Value of Preservation</u>

State governments have long understood the use of geospatial information in decision making processes and planning efforts. Agencies such as the North Carolina Department of Environment and Natural Resources, the Kentucky Transportation Cabinet, and the Utah Division of Forestry, Fire and State Lands have long embraced the use of GIS information to analyze real world problems like understanding the environmental impacts of major projects; the need to map roadways faster, more accurately and more cost effectively; and the desire to provide more efficient and effective services to their citizens. State governments are also recognizing the importance of having access to older data as it allows them to explore societal, environmental, and economic change geospatially over time. There are compelling business drivers to support the preservation of older data: tracking population, land, or vegetation changes over time; providing a cultural record of place over time; or avoiding the cost of re-creating datasets that were not preserved, are just a few of many drivers spurring users to seek out and use superseded geospatial content.

Suburban Development 1993/2002
Near Mecklenburg-Cabarrus County border

A comparison of these two orthoimagery tiles illustrates changes in the landscape in North Carolina.

State GIS organizations, archives and other agencies are making efforts to respond to this information need; however, they are facing serious obstacles. As revealed in the NSGIC Survey results[25], over 70% of respondents indicated that geospatial data was being archived within their state, however only 43% had a formal geoarchives program, with a majority of those programs being less than 5 years old. Limited resources, diminishing budgets, and in some cases a lack of understanding by key decision makers as to how preserving geospatial data will improve the "bottom line" can stifle the movement from commitment to implementation. As one respondent stated, "archiving is generally a low priority issue with little funding to support such efforts."

As such, it is necessary to make the case to legislators, financial supporters and users that geoarchiving has value, and that it can effectively leverage existing practices and processes to preserve and provide access to the data in a cost-effective manner. But how does an entity go about persuading these stakeholders that preserved geospatial data is not just "nice to have" but a "need to have"? GeoMAPP formed the Business Case working group to focus on creating tools and templates to help states and organizations develop formal business planning documents to justify and quantify the value of preserving superseded and temporal geospatial content. The end goal was to support the development of a compelling business case to encourage sustainable policy and funding support for preservation activities.

The working group recognized that leveraging existing practices and processes was a critical piece of justifying preservation programs, highlighting commonalities between existing activities and preservation efforts. For example, state governments already have continuity of operations (COOP) activities that provide for processes to secure data in the event of a disaster. The goals of COOP overlap those of data preservation for cultural heritage purposes, so it makes sense from a business perspective to align those activities in a meaningful way. The identification of shared goals became a starting point for complementary business case development between geospatial and archival groups. It became apparent that each state needed to have internal cooperation and collaboration to explore these existing value tracks and common goals that could show the largest return on investment focusing on benefiting the largest number of programs with the dollars allocated or spent.

[25] NSGIC survey raw data: http://www.geomapp.net/docs/geomapp_survey_nsgic_rawdata.pdf.

Pre-GeoMAPP Business Case Efforts

To facilitate the development of business planning tools, it was essential to first assess the various successes and obstacles each state partner faced in their attempts to gain support for preserving geospatial content. By 2008, the Utah Archives recognized that digitally-born records were being created at an ever-increasing rate and that they needed to implement an Electronic Records Management (ERM) strategy to proactively manage, preserve, and make accessible electronic records. Prior to this, electronic records had not been included in previous records management plans. In 2008, the archives produced a business case for Electronic Records Management[26] which was successfully adopted. While this document acknowledges geospatial data as an important electronic record, it did not address its management in detail. Driven by the involvement in the GeoMAPP project and using the ERM business case as a model, in December 2008, the Utah AGRC and Archives produced an initial draft of a Business Plan for Archival Preservation of Geospatial Data Resources. [27] This plan outlined a process for archiving digital geospatial data resources in Utah across all state and local agency partners, and includes goals, benefits, requirements, costs and an implementation plan. Due to state budget shortfalls, the business plan has not been fully addressed and the preservation of geospatial content is not adequately funded.

The Kentucky Archives took a different approach to business planning based on the maturity of their program before the grant started. In 2001 the Archives hired a consultant to produce a document used to gain initial funding for their Electronic Records Archives (ERA). While this document helped to create a process for archiving digital data, it did not provide a business case for the preservation of superseded geospatial data. Prior to the grant the archives did establish a relationship with the Division for Geographic Information, created a records schedule for their geospatial holdings and began archiving some records. While Kentucky is currently not planning to develop geo-centric business case documentation, it will use the success of the grant program to demonstrate the value in the archiving procedure.

While North Carolina Archives had been actively collecting electronic records, it had not specifically addressed the preservation of geospatial records until its involvement in GeoMAPP. During the course of the first phase of GeoMAPP, the archives staff had great success in securing support for its records management programs. In North Carolina's 2009 Legislative session, the General Assembly added an additional $5 fee on all deeds to be collected and sent to the Department of Cultural Resources to support the Archives and Records Management program; however, this legislation did not specifically call for the preservation of superseded geospatial data. The North Carolina team is hoping to test some of the existing and future business planning tools to develop documentation to support its long-term geoarchiving ambitions.

Drafting the Business Case Document

A goal of the Business Case working group was to develop templates to guide users in the creation of compelling business case documentation that would not only adequately capture both the quantifiable and unquantifiable aspects of a proposed preservation program, but would also include the consequences of

[26] To read the document in its entirety, see: http://www.geomapp.net/docs/ut_ERMBusinessCase.pdf
[27] To read the document in its entirely, see: http://www.geomapp.net/docs/Utah_Business_Plan_Geospatial_%20Archive_2008.pdf

doing nothing . In late 2008, the working group finished the first draft of a business case document for the preservation of digital geospatial data using the Utah Business Plan as a starting point. The initial draft incorporated information and ideas from a variety of resources, including the Utah Geospatial Infrastructure (UGI) Strategic Plan, [28] Utah Division of State Archives *Electronic Records Management Business Case,* and a set of strategic and business plan templates created by the National States Geographic Information Council (NSGIC). [29] The draft merged business case and business planning concepts into a single document. This effort was an iterative process with each discussion of the business plan resulting in new ideas and confirmation of direction and focus. The group also worked to make the original Utah-centric business drivers and supporting material more generic for broader comprehension and adaptability by different state entities.

In early 2009 the working group engaged members of the geospatial and archival communities, including representatives of NSGIC, the Federal Geographic Data Committee, and the Society of American Archivists to solicit input on the early draft and gauge impressions of the direction of work in both the geospatial and archives/library communities. The plan received positive feedback and the partnership felt that it was on the right path.

What We Learned

The key lesson the team has learned at this point is that there is a shortage of existing information available to help archivists and GIS staff develop and create business plans to build and support sustainable archives. The community faces the same issues of how to secure continued support and seek new support to implement new programs when all programs face meticulous scrutiny based on budget shortfalls. Business planning documentation and justifications are critical for defending existing programs and the development of new ones. There is strong interest from both the archives and GIS communities for having sharable tools to help justify their archives programs. Each professional group GeoMAPP contacted in regards to the business plan effort were supportive of it and felt it had value. These groups also suggested that the working group needs to create a broad variety of generic tools to assist both archivists and GIS professionals get started with the business planning process. Examples of possible future tools include additional or refined business planning templates, tools to help capture use cases, cost benefit analysis, and cost estimation of programs. GeoMAPP plans to engage the support of an outside contractor to organize and enhance the technical and financial sections of the plan. The contractor would also create a cost benefit analysis with a view into the long term costs of the plan.

The partner states have common goals, but each has unique challenges. Each state's geospatial preservation processes and budget constraints within the partnership must be investigated individually to account for the unique intricacies within each state.

Next Steps

GeoMAPP 2010 efforts will concentrate on the continued development of a generic business plan toolkit that can be shared with other states. The business planning "toolbox" will include a model plan, a business planning template, a timeline tool and a series of templates to assist states in identifying the

[28] To read the document in its entirety, see: http://gis.utah.gov/docs/gisac/UGIStratPlanDraft0608.pdf
[29] http://www.nsgic.org/hottopics/fifty_states.cfm

return on investment of preserving geospatial data. The working group will continue its outreach to external partners for feedback on the progress of the toolkit, and continue to refine the iterative development of the business case documentation. New requirements will need to be included as they become known and further work on the associated suite of tools will need to be completed. This suite will include a business planning process map that can be used by those interested in developing a business plan to map out the steps that will need to be taken to develop personalized business planning documentation.

The partnership also proposes that each state identify a legislative champion in each individual partner's legislative body and ask them to comment on the quality and content of the business case. It will be imperative that the partners garner this support and receive suggestions on how to modify their business plan to help enable long term funding support. In GeoMAPP 2010, the partnership will be developing more cost benefit analysis and use cases to show why the preservation of superseded geospatial data is valuable now and in the future.

Knowing What You Have – Inventory and Metadata

Why it is Important to Inventory Your Data

After a state commits to a formal program for preserving geospatial content, it is tempting to forge ahead and begin transferring superseded data from the geospatial data clearinghouse to the archives. As appealing as this approach seems, before any geospatial records can be appraised, transferred or ingested into an archival repository, determining what data currently exists is critical. Understanding current holdings provides an accurate assessment of how much data exists (not just the number of datasets, but extent as well), its current format, and important details such as who is responsible for the data, when it was created, and where the data came from. All of these elements are essential for appraising which content is "at risk" and needs to be considered for long-term preservation and access.

The Inventory working group was given the responsibility of creating a master inventory of all three state's holdings. Creating, examining and analyzing individual holdings and then merging them not only served to identify the important elements that could be included in a shareable geospatial inventory tool, but was used to drive the appraisal and selection of datasets for later data transfer activities. Creating a master inventory also gave the group the opportunity to investigate similarities and differences in data classification, naming schemes, metadata, and metadata schemas. The findings of this analysis helped the Content Lifecycle and Data Transfer working group identify the most critical datasets for preservation while providing a framework to organize the data holdings and capture critical information about each dataset that would be included in the preservation process.

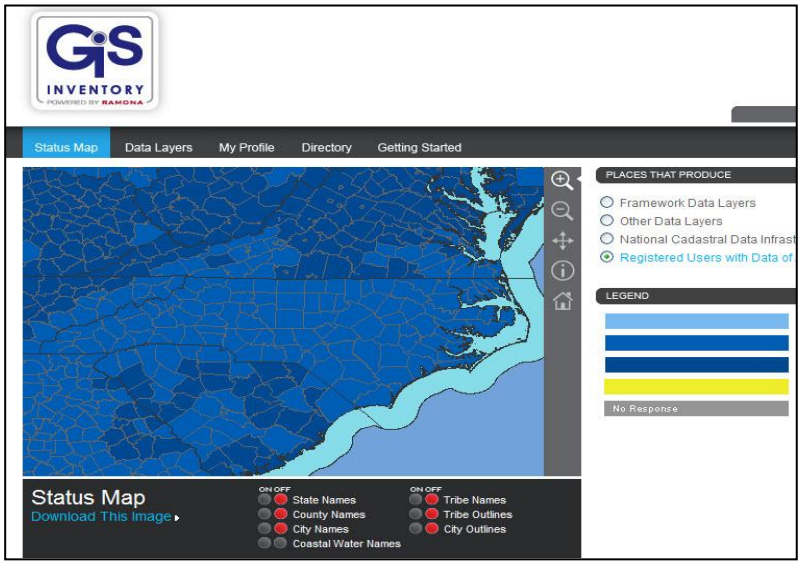
Partners Inventory

It came as no surprise that each state partner used a different means for tracking and inventorying their statewide geospatial data holdings. North Carolina's primary centralized inventory tool is the NC OneMap Inventory[30] powered by the national RAMONA database.[31] This database allows any local or state agency to enter important information about their geospatial data into a central web-based interface

[30] For additional information about the NC GIS Inventory tool, see: http://www.nc.gisinventory.net/.
[31] For more information about the RAMONA GIS Inventory Tool, see: http://www.gisinventory.net/.

that is national in scope and freely accessible to all. The GIS Inventory/RAMONA database is divided into 18 data categories, with over 200 specific data layer types available for users to select from to classify their data. In addition, the data types are delineated between the Framework and Non-Framework categories.[32] The 23 Framework data categories include commonly used datasets such as orthoimagery, boundary information and hydrography. From the information a user provides about a specific dataset, a starter Federal Geographic Data Committee (FGDC) metadata record is produced. The inventory tool also allows users the option of publishing the information about their data to the Geospatial One Stop (GOS).[33]

As an element of the NC OneMap program, the NC OneMap Inventory tool has been used by CGIA to record information about geospatial data across the state. CGIA also has additional methods to post data holdings to the GOS data discovery tool. Both of these inventory processes were in place prior to GeoMAPP.



As of December 2008, the NC GIS inventory included participation from 86 counties, 46 municipalities and 69 state agency representatives and was tracking over 2,200 geospatial datasets.

Kentucky had an established inventory and archival process in place for centralized geospatial content housed in the state's KYGEONET clearinghouse prior to joining the partnership. The Kentucky clearinghouse is modeled off USGS' Geospatial One Stop (GOS) Portal and currently has 19 publishers who provide data created by local, university, state and federal agencies. Information about these datasets is also posted to the GOS portal as another method of data discovery and access. Kentucky is currently not actively participating in the GIS Inventory (RAMONA), but did investigate the tool as part of the project.

In Utah, the State Geographic Information Database (SGID) had been established as a data repository to distribute all geospatial data created for Utah, but did not have a formal means to track this content. After joining GeoMAPP, Utah began a vigorous outreach program to engage county, state, and local agencies that were producing geospatial data. This outreach program afforded AGRC the opportunity to become more knowledgeable about what data were available (over 2000 datasets not in the SGID were collected), and realized that it would be important to select and utilize a tool to inventory these datasets to help with data management and the archiving process. Utah loaded each of the datasets discovered during their outreach efforts into the GIS Inventory and continues to use this system to inventory and track datasets

[32] List of RAMONA Data Categories and Layers: http://gisinventory.net/RAMONA_Data_Categories_and_Layers_2008.pdf
[33] See: http://gos2.geodata.gov/wps/portal/gos.

around the state.

While spending time using the RAMONA database to search and report on data holdings, project participants found there were some key elements missing from the GIS Inventory application that could benefit the archives process. The team proposed several enhancements to the RAMONA inventory including: functionality to send periodic email reminders to inventory participants to update their inventory holdings; inclusion of new fields in the data entry form such as "Layer Title," "Data Format", and a new standalone archiving section for each dataset; as well as enhanced reporting functionality to make better sense of data that was stored in the inventory. Building on efforts that began with the NCGDAP project, the team submitted a list of these recommended enhancements to the stewards of the GIS Inventory tool, and while there have been improvements to the application not all of the recommendations have been implemented.

Building the Project Inventory

Once each state established an internal process to inventory their data holdings, the next step was to develop a method for comparing datasets between states. Creating and analyzing this collective project inventory could not only be used as a basis for discussing what elements were important to capture in a generic inventory tool, but by becoming more familiar with each other's data, the team could keep an eye towards the later processes of appraisal and data transfer. Participation in the RAMONA inventory database afforded both North Carolina and Utah the opportunity to easily export information about their state data holdings into Excel spreadsheets; with each of the datasets organized into pre-defined categories. The format of the RAMONA data extract influenced the selection of many of the elements that the team chose to include in the project-wide inventory, in part to enable cutting and pasting from the state specific RAMONA extracts to the project's "common" inventory. For the project-wide inventory, the team utilized both the RAMONA data categorization and the International Organization on Standards (ISO)[34] 19115:2003 Geographic Information--Metadata standard to help classify each dataset while attempting to address the challenge of reconciling the occasionally conflicting resources. The project inventory also captured information about each dataset's name and description, the date the dataset was created, how often it was updated and if the dataset resided in the state's central clearinghouse.

State	Ramona Framework	Ramona Category	Ramona sub type	ISO Category	Keyword	Layer Title	Layer Description	Production Date	Update Frequency	Centrally Managed
Centralized Records										
KY	No	Biota	wildlife	Biota	wildlife	Wildlife_Management_Areas	Wildlife_Management_Areas		As Needed	Yes
NC	No	Biota	Rare, Threatened & Endangered Species	Biota	wildlife	Natural Heritage Element Occurrences	identify locations of rare and endangered species populations and occurrences of exemplary or unique natural ecosystems (terrestrial and palustrine) and special	2008	biannualy	Yes
UT	No	Biota	Plant Distribution and habitat	Biota	vegetation	Dominant Vegetation				Yes

KY | NC | Utah | Biota | Boundaries | Elevation | Environment | Ima

A snapshot of "Biota" data from the GeoMAPP Project Inventory

[34]Additional info on the ISO Geospatial Metadata standard: http://www.iso.org/iso/catalogue_detail.htm?csnumber=26020

Both Utah and North Carolina extracted the inventoried data information from their RAMONA database instances and were able to do a fair amount of cutting and pasting to populate their state-specific sections. Kentucky was able to extract some of the same types of information from their KYGEONET database into the spreadsheet, but much of the data was entered by hand.

Crosswalking

After hundreds of datasets from each partner were entered into the project inventory tool, the next step was to examine and compare each state's data holdings in order to assist the Content working group in the process of selecting the datasets to be used in both the Intrastate and Interstate transfer processes. In comparing the three inventories it was clear that each state categorized their individual datasets differently. There was a need for a crosswalk that could tie all the variously named datasets and unique data types into a single set of categories to compare the three partner's disparate datasets side by side. A crosswalk uses a set of categories that each state agrees describes their data. If all partners agree on the categories in the crosswalk, then data classifies as "boundary data" from one state should be similar to "boundary data" from another state.

Each dataset in the project inventory was assigned a RAMONA category and subtype as well as an ISO category and keyword. The team integrated tabs for the 19 unique ISO 19115:2003 categories[35] into the project-wide inventory with the existing state specific tabs and then imported data from each of the states, sorted into the respective ISO categories.

Metadata – Have to have it!

After the data about the different states' geospatial holdings had been inventoried and cross-walked using an internationally accepted standard, the Inventory group turned to investigating the role of metadata in managing and preserving geospatial data. Metadata is a critical element in understanding and managing geospatial data and was realized to be an essential component in the archiving process. Without complete metadata, it would be challenging to discover the "who, what, where, when or how" about any geospatial dataset, information that is necessary to have documented especially for data that is going to be preserved for many years.

The Inventory team compared and analyzed the FGDC Content Standard for Digital Geospatial Metadata (FGDC-STD-001-1998)[36] and ISO 15836:2003 Information and documentation -- The Dublin Core metadata element set [37] standard as potential "wrappers" [38] for the data. The team concluded while FGDC metadata is more robust for capturing the in-depth information about datasets for research purposes, Dublin Core works well for data discovery. The team created a metadata comparison document which proposed a simplified model merging optimal metadata for both the FGDC and Dublin Core standards was proposed.[39] After the study, Utah agreed to use completed FGDC metadata for all of its spatial data, while integrating Dublin Core metadata as a package descriptor explaining multi-faceted projects. North

[35] ISO Category info: https://www.ngdc.noaa.gov/wiki/index.php?title=ISO_19115_Topic_Categories
[36] For more information or questions about the FGDC Standard Metadata go to http://www.fgdc.gov/standards/projects/FGDC-standards-projects/metadata/base-metadata/index_html.
[37] See http://www.dublincore.org/documents/dces/ .
[38] In general, a metadata wrapper would contain all additional bits of metadata elements including descriptive, administrative, technical, and structural metadata.
[39] To read the complete study, go to http://www.geomapp.net/docs/MetadataComparison_200903.pdf.

Carolina and Kentucky would continue to use existing data processing workflows, which already required metadata for each dataset that met the FGDC standards.

Legwork and Lessons Learned

This team's detailed exercises in data inventory and metadata research were necessary to discover and formulate guidelines that could be followed by other states and their localities. There were several lessons learned:

- There is a critical need for an ongoing, dynamic and regularly updated inventory in each state, as knowing one's holdings aids in day to day "production" data management as well as for archival appraisal and ingest. The information captured in an inventory should include:
    - Name of the dataset
    - Description of the data
    - The name of the data creator or steward
    - How frequently the dataset is updated
    - Geographic scale of the data
    - ISO category that the dataset belongs to
    - If the data has been or is being continually archived
    - Frequency of archival capture
    - How to access the current and/or the archived data

- An inventory should allow its users easy access to the information about the individual datasets, and should be scalable enough to employ suggestions or accommodate changes that would improve the tool.
- A standard crosswalk that links the inventoried data to a national or international standard of categories can enable the sharing of multiple organizations' inventoried data for easy reference.
- Metadata (whether it is FGDC, ISO, or Dublin Core) must accompany each dataset being archived. Data that lacks significant supporting metadata risks being seen as inauthentic or inaccurate, and future users may not be able to understand the data or the context in which it was created.

Next Steps for Inventory Group:

Since the initial effort to inventory and compare the GeoMAPP partners data was completed during this phase of the project, the working group will not carry forward to the next phase of the project. Several themes that the group began to research will continue as the GeoMAPP efforts progress. These themes include:

- Continuing to update their own states' inventory database;
- Encouraging agencies that participate in an inventory to keep their data current;
- Providing guidance to new partners as they begin the process of inventorying their own geospatial data;
- Investigating critical geospatial elements that are essential to the preservation process;
- Examining how administrative metadata can be used in archival data management approaches.

Figuring out What to Preserve and Making it Available– Appraisal and Access

The Value of Appraisal

In an archival context, appraisal is the process of determining the value of records. During the appraisal process, archivists and records managers assign administrative, legal and research/historical value to records in order to decide retention periods for how long records need to be maintained, where they might be maintained and when they may transfer to the State Archives, if those records are archival in nature. Retention periods for records can range from storing a record for six months to keeping it permanently. Permanent records with historical or research value are often considered to be " archival" [40] and fall under the purview of the archives for long-term preservation, while permanent records addressing other administrative or safety and health matters are typically managed by the creating agency.

Since geospatial records can have high research value for analysis in areas such as land use, ecology, and a host of other vital topics, archivists and those creating geospatial records generally agree that these records are worthy of long-term retention. Because of the complexity of these data and the high cost in maintaining the records, a strong business case that assesses risk is important to both prove the value of the record and suggest the best strategies of preserving the record in the most cost-effective way.

Digital geospatial data inherits the preservation challenges that apply to electronic records in general, including the sheer volume of records and the rapid rate of technological change, but it also presents a number of domain-specific challenges to the preservation process. While key geospatial feature datasets such as land records, street centerlines, jurisdictional boundaries, and zoning are constantly changing, current data management practices by these data's creators commonly involves the overwriting of older versions of data which are then no longer available for historical or trends analysis. Other unique preservation risks for geospatial data include: unique and often proprietary geospatial data formats; spatial database complexity; the variety and complexity of digital cartographic representation methods; issues related to time-versioned content; and the absence or inconsistency of metadata. These factors not only increase the risk that data will be lost during the life of the record and increase the cost of reconstructing the data if they are lost or damaged, but also make the appraisal of these records exceptionally challenging.

Appraisal Models and Decisions on Preserving Permanent Records

The Appraisal and Access working group focused on evaluating "archival" geospatial datasets within the framework of established records management and archival procedures in each state. Traditionally governmental records managers have appraised from the perspective of selecting records for long-term retention that best document the activities of the government and society at large. While the group recognized the value of this more conventional approach to appraisal, the hope was to move beyond this approach to establish new best practices for geospatial records. To the extent that geospatial datasets may contain valuable information useful in a variety of research, appraisal goes beyond identifying records that document the transaction of public business. In this context geospatial data as electronic records can

[40] A Glossary of Archival and Records Terminology; Society of American Archivists; http://www.archivists.org/glossary/term_details.asp?DefinitionKey=3

serve numerous purposes beyond their original intent.[41] Geospatial systems, unlike the static maps that are outputs of these systems, are also more difficult to schedule because they are constantly changing and they are not arranged in traditional record series.

The working group used the geospatial datasets identified by the Inventory working group as the basis for its appraisal processing. The group also conferred with the Content Lifecycle/Data Transfer working group on data sharing once archival datasets were identified. In light of the NDIIPP mandate to identify "at risk" materials for preservation, the group evaluated current retention practice in other states[42] and began developing strategies for permanent preservation that minimized the risk from loss of the valuable records, which both document state activities and provide valuable resources for conducting research over time.

Each state team in the working group was tasked with appraising their geospatial records. During the appraisal process, each state transcended traditional appraisal by considering additional steps such as frequency of capture, scheduling of duplicate copies, developing creative disposition statements, and other modifications to the records retention schedule. During the course of appraisal, each team also reviewed their records retention scheduling processes. Records retention scheduling is an important part of state government records management and archival workflows and GeoMAPP has focused on exploring techniques to effectively integrate the scheduling of geospatial data under existing records retention regimes. Additionally, each of the state partners began development of records retention schedules specifically targeting geospatial information that can be shared with the wider archival community.

Kentucky

The Kentucky team appraised all of the records in the centralized KYGEONET as permanent and archival since they represent the most important geospatial records as assessed by a consortium of the record-producing agencies. Both KDLA and DGI decided to take snapshots of its centralized vector databases on a quarterly basis and maintain these permanently in the archives. This short frequency was thought to allow maximum practical capture of the complete set of vector datasets, some of which change with great frequency while others do not. During the grant period nearly two years of quarterly snapshots were archived. Since the database files in the KYGEONET are the point of collection for the archives, all other geospatial records that are duplicates of these records are evaluated as "Delete when no longer useful." [43] Raster image files that are currently regenerated every two years are also to be kept permanently either by the Division of Geographic Information or by the archives.

Kentucky's general schedule series applies to all state agencies and identifies the KYGEONET as the primary point of capture for the archives. This eliminates the need for data creating agencies participating in the KYGEONET to keep their contributed geospatial records permanently. Agencies with substantial records not included the KYGEONET are scheduled separately with agency specific records series. In an effort to identify older geospatial records that could come to the archives, Kentucky examined agency

[41] For a discussion of a related approach to appraisal of scientific datasets in the federal government see: http://www.joss.ucar.edu/daarwg/june08/NOAA_Appraisal_Approval_Procedure_V6_03a.pdf.
[42] Both Maine and Michigan furnished schedules and appraisal techniques to the working group. These can be found at: http://www.geomapp.net/docs/me_gis_schedule.pdf and http://www.geomapp.net/docs/MI_Schedule_EnterpriseReport6330.pdf.
[43] For additional information, see: http://www.geomapp.net/docs/ky_gis_schedule.pdf .

websites and talked to authorities in various agencies that produce GIS records to find valuable records, including static maps and project files that could come to the archives apart from what was in the KYGEONET.

In the case of large collections of static documents, such as those created by the Kentucky Geological Survey, the archives elected to work with the agency to support the agency repository rather than bring all the documents to the archives. Local GIS agencies in large metropolitan areas that were organized as consortiums also negotiated memorandums of agreement with the archives to retain their ability to generate receipts from the records and ensure that they would remain the primary access point for the first year of a record's life. In conjunction with DGI the archives also identified valuable geospatial records (such as parcel records) that have never come to the centralized repository due to local agencies desire to recuperate costs though sale of datasets.

North Carolina

The North Carolina appraisal team consisted of archives staff from the State Agency Services, the Local Records Unit, the State and University Records Unit, the Information Resources Branch, the Electronic Records Unit, as well as staff from the Center for Geographic Information and Analysis (CGIA). The team held a series of meetings over the course of six months. Initially, the North Carolina team discussed what data layers are, how GIS information is produced, how data flows within the state, and what processes CGIA employs when it receives new data and takes older data down from NC OneMap. Once the team established an understanding of these concepts, they reviewed North Carolina's centralized data holdings which had been loaded into and categorized in the project inventory. Once organized, the team began discussions about how to address the appraisal of the data. The main outcome of these discussions was that the majority of NC OneMap's holdings were classified as "permanent" or "archival" records since many of these datasets model statewide or regional features or have general research value. The team also discussed and began appraisal of typical data created by local governments, identifying several potential approaches on how to identify critical datasets and how frequently to capture it. The approach to local government drew heavily from the recommendations made by the state's GIS coordination council.[44]

The team also produced draft versions of records retention and disposition schedules[45] as existing schedules made little or no mention of digital geospatial records. Since all but two counties and many municipalities in North Carolina produce GIS data, the North Carolina team felt it would be beneficial if local government data producers participated in NC OneMap. Currently, participation in OneMap is voluntary; however, if the data producers were to participate in NC OneMap, the North Carolina team could work with CGIA to transfer all of this data in a consolidated fashion. Otherwise, counties could choose to preserve the data themselves but would need to consult the *GICC Archival and Long Term Access Ad Hoc Committee Final Report* adopted by the North Carolina Geographic Information Coordinating Council and follow the provisions for archiving. Another option discussed was to transfer confidential or sensitive data directly from the locality or agency to the archives. Since a robust

[44] NC GICC's archives recommendations: http://www.ncgicc.com/Portals/3/documents/Archival_LongTermAccess_FINAL11_08_GICC.pdf
[45] North Carolina created two draft schedules: http://www.geomapp.net/docs/ DENR_CGIA_NCOneMap_2009May26.pdf and http://www.geomapp.net/docs/ Local_gis_retentionschedule_items_v1.pdf

geoarchive environment has not been fully implemented and there is no clear immediate funding to support its long-term maintenance and update, the official scheduling of local government and state agency geospatial data was put on hold.

Utah

The Utah team relied on ISO 19115:2003 geospatial categories to develop recordkeeping instructions for their retention schedules[46] that apply to agencies which produce geospatial records. These categories provided a natural classification system for the records, or a framework that described records broadly enough so that each time a new dataset was identified for appraisal, an existing retention schedule was available to address that data. The types of retention schedules developed for this purpose are known as general schedules. A general schedule is applicable to records from similar state, county, or municipal offices; or departments that have many satellite offices where the work product is the same. Descriptions of records contained within a general schedule tend to fit a wide variety of data, but are recognizable by records creators. These schedules then go through an approval process, and when approved, agencies can destroy or transfer the records to the archives as instructed.

Once the general schedules were in place, another form of retention schedule was used, known as a specific or unique record series retention schedule. A record series identifies a set of records created for a common purpose, filed together, that shares a retention length. This type of schedule identifies records by the specific office that produces them, and describes the records in more focused detail. One reason to create a specific schedule is to obtain a control number (or record series number) that could then be used to interact with archives' services. If the records do not come to the archives, then this step is unnecessary. These specific schedules often cite a general schedule if there is an appropriate one that fits and so need not go through any further approval process. If a specific schedule is unique enough to not be related to a general schedule, then it would go through its own approval process. Once a record series retention schedule is established, geospatial datasets are identified in the archives' database and listed as they are accessioned. Any geospatial data producer outside of county or municipal government or AGRC is able to schedule their records individually without use of a general retention schedule, by creating a specific record series retention schedule. This will be the case for a handful of state agencies. In those cases the records are unique enough to only be applicable to that department, and so creation of a general schedule would not make the work more efficient.

In the very beginning of the appraisal process, it was difficult to determine series boundaries, or what constituted "a set of records filed together." Initially, the ISO categories were used as individual series. An inventory of all geospatial records created in the state was created, determining whether those records resided in the centralized SGID database or were housed in local, county or city offices. As the inventory process progressed, it became clear that since each category included so many disparate sets of information, the general schedule approach would be more realistic. This also allowed individual datasets of like kind (generally differentiated only by production date) to be grouped together appropriately, described as a unit, and assigned meaningful access terms. For instance, demographic data (such as race or income) and cemetery locations are two separate record series, but are all described under the ISO category of Society Records.

[46] Utah general schedules: http://archives.utah.gov/recordsmanagement/grs/AGRC-General-Schedule.pdf

Retention schedules[47] for specific record series identify the disposition instructions (or what happens after the retention has been met) for those records, including whether or not the records will be sent to AGRC, maintained locally, or sent to the archives. The disposition of duplicate copies is also identified where applicable. Potentially, multiple record series could be created for duplicate copies, as those copies only mirror each other at the moment of duplication. As soon as one copy is sent to another institution, data will change either at the creating agency or another organization using the records, such as AGRC (although not at the archives). So the datasets tend to morph even as they retain the same descriptive metadata. As soon as records are no longer "filed together, created for a common purpose," they are identified as a separate series. In this case the catalog record at the archives will contain a Related Materials note to tie them together.

Appraisal Commonalities and Differences among Partners

Comparison of Schedules

Utah's approach to appraisal is to identify all geospatial records by ISO 19115:2003 category and apply this classification to the records scheduling structure via general schedules that can be appropriately used by agencies. Kentucky's approach to scheduling is more agency-centric, starting with the existing record series which applies to DGI's KYGEONET and adding general schedule series to the local, state and university general schedules in order to declare that all geospatial files not captured in the KYGEONET system were also to be saved permanently. North Carolina, with a decentralized system of geospatial records, is a mixture of the Utah and Kentucky approach. ISO 19115:2003 categories are applied as records series and the point of capture is the individual agencies, which are permitted to store their own archived records according to a flexible disposition statement.

Retention periods

Kentucky declared all records maintained in the central database as archival. The North Carolina team appraised each dataset from the NC OneMap clearinghouse and were selective in the layers they declared to be archival since it was not clear whether every single dataset held equal value. The team also declared that all framework layers needed to be preserved. When data creators were visited personally by the archives in Utah, they found that the creators of the records believed that all data needed to be kept to provide context for the other data, and when asked if they would still use information from 100 years ago if they had it in their hands, they said yes. With a clearly established administrative need for the records, as well as historical value for the whole, the decision was made that the retention schedules for each dataset would be permanent. The only variation from this rule was for project files. Most projects have a short-term value of only10 years or less. Occasionally a project will have significant public interest, in which case those project files will be kept permanently.

Frequency of capture

For regularly changing datasets there is a critical question: how often should snapshots be taken of the data for inclusion in the archives? The challenge is to determine what granularity of change will be lost if a snapshot is not captured and saved each time the data are revised or updated versus ingesting and

[47] Utah agency records schedules can be found here: http://archives.utah.gov/recordsmanagement/rclist-a.html

managing multiple copies of a changing dataset that could be updated multiple times per day. Kentucky captures vector snapshots on a quarterly basis in order to capture all layers that could potentially change, and determined that yearly capture of vector data was adequate in most instances due to the typically static nature of those data. The North Carolina team decided to work within existing record keeping principles and deem that dynamic data would transfer when it was superseded by new datasets or was obsolete. In Utah, the decision of when to take a snapshot of the dataset was generally set to be once per year, although since some datasets rarely change, and others change quite frequently, that decision could vary from one dataset to the next.

Changes in Policies

Since Kentucky had a schedule for their centralized system before the project began, the only changes in policy that were made during the project were to add general record series for geospatial records to other agency groupings such as universities, local and state government. In the case of North Carolina and Utah, the development of schedules for geospatial records occurred during the grant period and both teams benefited from discussions of the group. During the project, it was learned that traditional scheduling provides a basis for appraisal of geospatial records, but the unique nature of geospatial records require different dispositions and capture procedures. Each of the states took slight different approaches to scheduling, but they learned from sharing those approaches.

Data Access

One of the key components in the lifecycle of a record is the ability for users to locate and retrieve it. Data access in an archival setting is usually accomplished utilizing tools such as catalogs, indexes and finding aids. If records are not accessible and the costs to render them are excessive, then they essentially have no value.[48] Because of the importance of accessibility, the Appraisal and Access Working group was also tasked with assessing various methods of making archived geospatial records available to users and offering suggestions as to which methods may be the most effective.

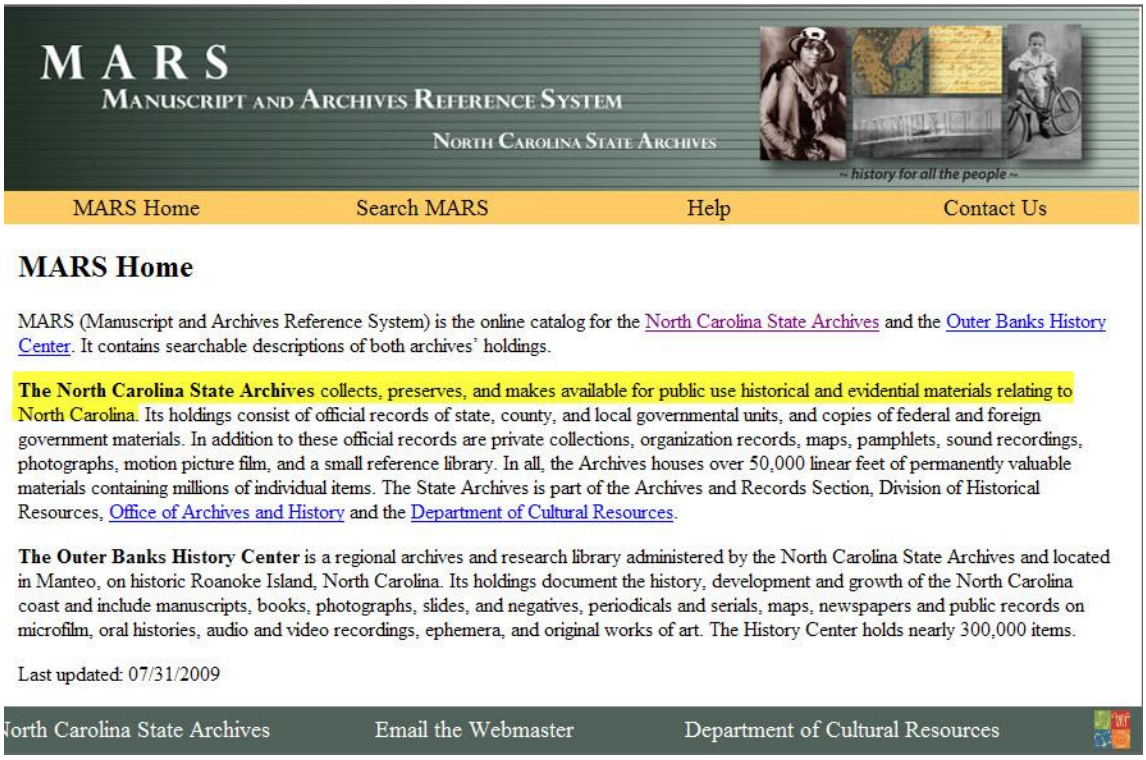
Kentucky uses the DSpace software as an open source data repository for all electronic records whether GIS, state publications, minutes or other electronic records Kentucky chose to bring into DSpace those geospatial records such as PDF maps, scanned map images, more dynamic geospatial files such as project files, and other shapefiles that are outside the realm of the KYGEONET.

KDLA created an e-archives in DSpace as a way of managing digital public records and making them available to their citizens.

[48] Kansas Electronic Records Management Guidelines: http://www.kshs.org/government/records/electronic/electronicrecordsguidelines.htm#7_2

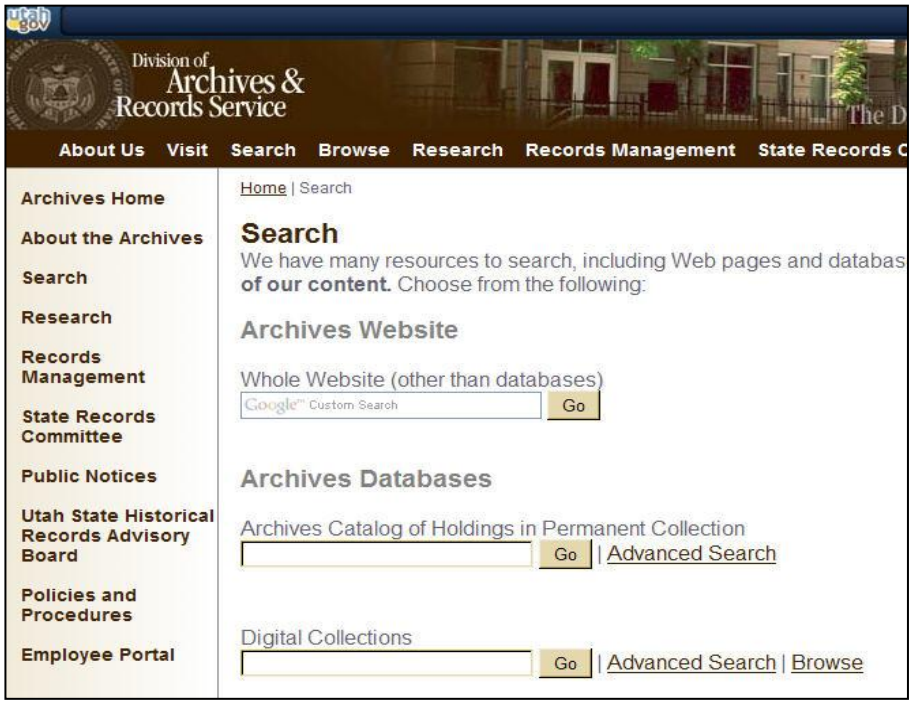
These records are grouped together where appropriate and searchable by agency, title, date, geographic name (county, city etc). Kentucky also plans to also use DSpace to reference database and image files in the archives that are only accessible using ESRI software. While these database and image files stored at the archives are currently only available to researchers through a research room workstation. Kentucky will investigate means of providing access directly to the database files through web mapping services during the next phase of the project.

Prior to the GeoMAPP project, the North Carolina State Archives did not have archived geospatial data in their holdings. As data began to be ingested as part of the project, the North Carolina team decided to use a combination of available access tools to expose the demonstration data transferred to the archives. The team cataloged the data into the Manuscript and Archives Reference System (MARS),[49] the online union catalog for the North Carolina State Archives which contains searchable descriptions of its widely varying archival holdings. In order to give the data maximum exposure for this part of GeoMAPP, the North Carolina Archives also created an Encoded Access Description (EAD) finding aid[50] which was indexed by the commercial search engine Google™. As a last step, usability test cases were created and conducted to test how well people could find the data using these tools. For additional information on this process, see the Content working group section starting on page 40.



Within the MARS catalog users have the capability of searching for the GIS data collection using a variety of basic and advanced methods to find the specific dataset(s) they seek.

[49] See: http://archives.ncdcr.gov/mars/.
[50] For info on EAD see: http://www.loc.gov/ead/



Utah's application allows its users to find information in a variety of ways.

Finding aids for Utah records are presented in two different ways. A Machine Readable Catalog (MARC) record[51] is created per record series, as well as an EAD finding aid. Both are generated by a custom application, APPX-based Archives Enterprise Manager (Axaem),[52] developed in part by staff from the Utah State Archives. Each GIS dataset within a series is entered into the database, where specific metadata are recorded (scale, projection, datum, type of GIS file, file size), as well as a URL to the FTP server where the dataset and its full metadata can be downloaded. These details are then made part of the finding aid. Cataloging processes ensure that these records are searchable via the standard means: creating agency, title, subject, scope and content, and other descriptive note fields.

The MARC record is uploaded (individually or as part of a batch) to the SirsiDynix Horizon commercial integrated library system (soon to be replaced by SirsiDynix Symphony). The EAD version is available on Utah's website both as a dynamically-generated XML file upon request by a browser when a database search is done, as well as static files posted to the web server and linked from other research guides. The dynamic version offers up-to-the-minute content as archivists add new accessions and other corrections to the finding aid. The static version undergoes more peer review before it is published, and is more easily harvested by search engines, although the GIS records may be directly harvested through an Open Archives Initiative Protocol Metadata Harvesting (OAI-PMH) connection[53].

Lessons Learned and Best Practices

A common insight from each state's appraisal process was the identification of a single centralized repository, if one exists, as the easiest point of capture for geospatial datasets to be archived. It was discovered that local data producers can be protective of access to their geospatial holdings, but typically arrangements can be made for the state archives to serve a preservation role for these restricted holdings, particularly for older files. The records retention schedule process serves not only as a legal basis for

[51] For more info on MARC see: http://www.loc.gov/marc/umb/
[52] See: http://archives.utah.gov/search.html.
[53] For more on OAI-PMH see: http://www.openarchives.org/

preservation, but can also be an organizing tool for developing a preservation strategy.

It is important for archivists to understand geospatial metadata standards such as the FGDC Content Standard and ISO 19115:2003 in order to interpret for patrons the detailed information imbedded in geospatial records as well as to help categorize and organize data within an archival repository. There is value in creating external metadata in the form of archival finding aids to facilitate data discovery. A number of standard descriptive formats such as Dublin Core, MARC and EAD can be used, depending on the capabilities of agency archival systems. Each of the three states instituted unique approaches to their geoarchive cataloging and data discovery (MARS, DSpace and APPX). While each of these systems have inherent strengths and weaknesses, each offers built-in description and tools that can be adapted for use with geospatial records.

Recommended best practices for appraisal and access

- *Establish a relationship with data creators*. They can help with the appraisal process of determining which of their most valuable materials need to be archived and in establishing a long-tem preservation strategy;
- *Don't recreate the wheel when developing geo-centric records retention schedules.* Adapt existing records retention schedules for geospatial information (if they exist) but modify them as necessary with any special language or capture procedures, while also including any disposition statements that identify point of capture and frequency of capture;
- *Focus on superseded "at-risk" data.* Preserve dynamic (database or shapefiles) as opposed to static geospatial files where possible unless the static copy is the only file you have for older data or it is needed for access purposes;
- *Use metadata for access and understanding.* Describe geospatial records using archival descriptive tools and make them available via the internet where possible;
- *Capture data from reliable consolidation points.* Take advantage of existing GIS portals, clearinghouses and access tools where possible.

Next Steps

Since the first phase of GeoMAPP focused heavily on data inventory, records appraisal, and transfer of records to the archives, the following appraisal and access issues remain for the next phase of the project:

- Focusing on data access including use of geospatial web portals and other low cost web mapping tools for to provide access to archived data;
- Exploring how to configuring geospatial data to fit traditional schemes in terms of how archival data is arranged and accessed;
- Investigating enhanced methods of access in order to meet our user community expectations;
- Investigating the portability of archives management and cataloging tools to other states;
- Continuing to engage data producers within each state to raise awareness and increase buy-in for geoarchiving matters;
- Expanding the records scheduling process by identifying agencies not covered yet, particularly local records.

Preservation in Action – Content Lifecycle and Data Transfer

Storyboarding Data Movement

The first unofficial Content Lifecycle and Data Transfer working group was formed at the GeoMAPP kickoff meeting in March 2008. Presented with the challenge of storyboarding data movement within each state, the group attempted to model how data moved between: 1) different geospatial entities; 2) geospatial data creators or custodians and the archives; and, 3) between archival repositories. The group also discussed how those models could be extended for content transfer between states. Their efforts produced the *Multi-State Geo Archival Process Flow*[54] which charted how archives and data stewards could interact to move geospatial data.

The "official" Content working group evolved from this initial concentration on storyboarding to a focus on investigating and documenting the lifecycle of geospatial content and data transfer methodologies. Given the size and diversity of the scope of work that was needed to address the project's data transfer requirements, the Content group became the project's largest working group with each state committing at least one staff member from both the archives and the GIS organizations to the effort. The team drew heavily from the initial storyboarding efforts as well as the early findings of the Inventory and Appraisal groups in preparation for the eventual development of geoarchives systems within each state and the movement of geospatial content between states. Discussions catalyzed by the storyboarding effort helped build an understanding about the existing geospatial and archives processes within each state; a key prerequisite needed before geoarchives systems could be developed. The Inventory and Appraisal groups identified and organized information about each state's geospatial data holdings and made preliminary assessments about the kinds of data that needed to be preserved and began considering the "best" formats for preservation. Both teams noted the importance of FGDC Content Standard-compliant geospatial metadata in securing the long-term preservation and accessibility of the data in the archival packages ultimately submitted to the storage repository.

The efforts and feedback provided by the other working groups helped inform the development of "Intrastate" data transfer — a process intended to chronicle the full life cycle of acquiring, transferring, processing and archiving identified datasets of superseded geospatial data within each state. The Intrastate data transfer process included the design, description and documentation of results for each step involved in gathering data and transferring it to the archives, documenting management and preservation actions performed on the content, as well as capturing performance metrics for the network transfer, data loading/unloading and validation times for transferred data. Intrastate data transfer provided guidance on how geospatial data could potentially move from the data producers to a geospatial clearinghouse and from the clearinghouse to the archives.

The results from the Intrastate data transfer laid the groundwork for the second objective, which was the "Interstate" movement of archived geospatial data among the State Archives of the three GeoMAPP project partners. The purpose of this exercise was threefold:

1) To validate the data transfer methodologies of the other state partners;
2) To test the ingest of "foreign" data into the state's geoarchive to validate internal methodologies

[54] See: http://www.geomapp.net/docs/storyboard_dataflow.pdf.

and procedures, and

3) To investigate the viability of Interstate data transfer and provide lessons learned/ recommendations for sharing data for distributed archives or continuity of operations/disaster recovery purposes.

Preparatory Activities

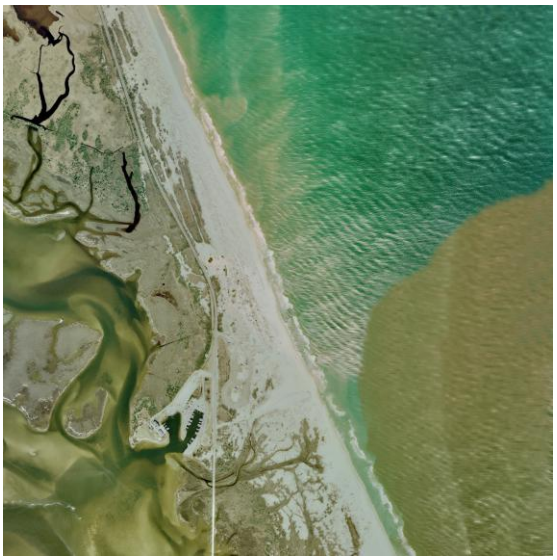
System Inventory

To prepare for data transfer, the Content group created a System Inventory spreadsheet template[55] to gather information about each state's existing geospatial and archival infrastructure. Information captured includes specifics about:

- Type of current and projected storage media;
- Amount of total space used on the storage media and the amount of free space allocated for future archiving;
- Types of servers and software used to manage and provide access to the data;
- Questions about network connectivity between the partner organizations (i.e., GIS and Archives) and to the Internet.

Data Sizing

The investigating the data storage element of the system inventory catalyzed a discussion about the sizing of geospatial datasets. The general consensus of the group was that raster digital aerial imagery products including county-based orthoimagery, and statewide imagery data such as National Agriculture Imagery Program (NAIP)[56] or Digital Ortho Quarter Quads (DOQQ's) posed a significant storage challenge due the size and complexity of the data. The size of imagery is proportional to the scale[57] or resolution of the image, meaning the more detailed the data, the larger the output file.

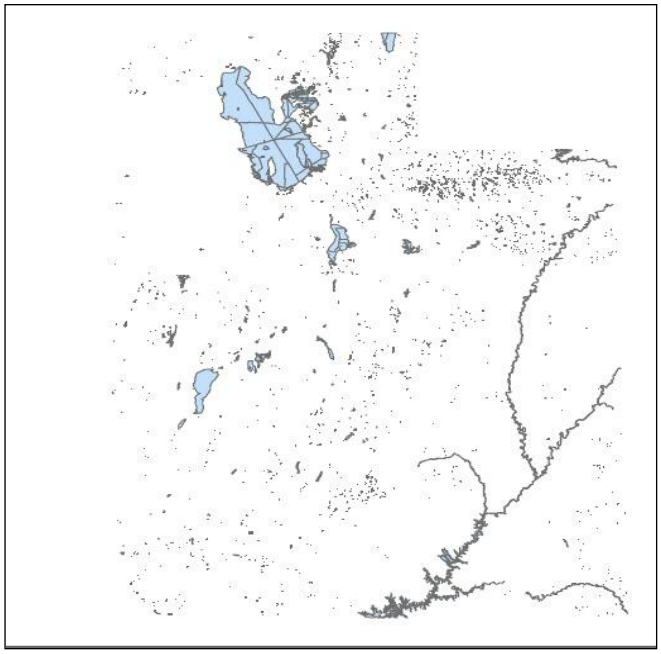


Uncompressed (.tiff) 2007 orthoimagery tile from Dare County (N.C.) captured at 400-scale. The size of this single tile is 300 MB. The size of the entire dataset, including the associated world (.tfw) files is 197 GB.

[55] To view inventory, see: http://www.geomapp.net/docs/GeoMAPP_System_Inventory_Template.pdf
[56] For more information about NAIP see: http://www.fsa.usda.gov/FSA/apfoapp?area=home&subject=prog&topic=nai
[57] Scale equates to the pixel size in the image to a measurement of what's being captured on the earth's surface. Local level imagery is typically flown with a scale ranging from 3 inches to 1 foot while statewide data is typically captured at a 1 meter ground resolution.

Newer, more detailed imagery in an uncompressed format can total several hundred gigabytes in size for one county's worth of imagery. To address the size challenge and to help ease in accessibility, state aerial imagery is typically "tiled" or broken down into smaller blocks; however, this merely adds to the data management complexity.

Vector data is comprised of points, lines, or polygons and underlying descriptive attributes, representing things varying from school locations, to river or road networks, to political boundaries. These data are typically much smaller. Simple point files demarking the x, y locations of things such as buildings with minimal descriptions or attributes about that location are typically very small, usually less than 1megabyte, while more complex data such as datasets capturing information about parcel locations descriptive information about each parcel are much larger, often having a footprint of several hundred megabytes to a few gigabytes. While vector data has other complexities that have to be accounted for, its size pales in comparison to that of imagery.



2002 vector dataset representing Utah water bodies. The size of this dataset is 1.5 MB.

Sweating the Small Stuff

Leading up to data transfer, each state partner framed the details for data storage, transfer methodology, and data validation.

While geospatial files had regularly been brought into the Kentucky State Archives before and during the early stages of the project, transfer of all of the files targeted by the grant for testing had to be delayed until after July 2009 when the State Archives purchased substantial additional data storage capacity using grant monies. To validate that transfer of the datasets, the staff of Kentucky installed hashing software including the BagIt[58] specification and MD5 Summer.[59]

For Interstate data transfer, Kentucky decided to use DVDs to transfer their vector data (stored in ESRI file Geodatabases), project files, and digitized maps. They also elected to provide these same files plus approximately 100 tiles of imagery for download via a file exchange website.

[58] Bag-It, developed by the Library of Congress, is a tool for creating and moving standardized digital containers, called "bags." A bag functions like a physical envelope that is used to send content through the mail but with bags, a user sends content from one computer to another. Bags have built-in inventory checking, to help ensure that content transferred intact. For more information, see: http://www.digitalpreservation.gov/videos/bagit0609.html.

[59] For additional information on MD5 Summer, see: http://www.md5summer.org/

In anticipation of the transfer of data, the North Carolina team spent the first several months of 2009 focused on dataset selection and sizing. Based on the size estimates, the North Carolina State Archives purchased and staged a storage environment consisting of 15 terabytes of Storage Area Network (SAN) storage and 3 portable drives totaling 7 terabytes. The team based the initial database sizing in part on the size of the total holdings (~14 TB uncompressed) of NC OneMap, North Carolina's spatial data clearinghouse. The Department of Cultural Resources Information Technology group (DCR-IT) also allocated a small application server to the project to help run scripts and manage the data.

The North Carolina team planned to test two methods for moving data between CGIA and the State Archives. For smaller vector packages, the team chose to transfer data across the network using the state Wide Area Network (WAN) to move the data between agencies. For full system transfers and for imagery, the team opted to use portable hard drives to transfer files. For Interstate data transfer, the team provided uncompressed orthoimagery via an external hard drive for Kentucky and Utah to transfer. All other types of data (vector, digitized maps and project files) were to be made available for download via a temporary FTP site.

To test the validation of both the Intrastate and Interstate data transfer, the North Carolina team installed three hashing generators (BagIt, MD5 Summer, and md5deep[60]) on the GeoMAPP server and on a local desktop at CGIA. After reviewing each of the tools, the team decided to use BagIt for both Intrastate and Interstate data transfer as it offered the most dynamic features for validating and transferring data. Using the tool allowed the team not only direct access to the BagIt development team if there were questions about using the tool, but also afforded the team the opportunity to provide relevant feedback to the development team for future releases of the BagIt specification. Additionally, ArcGIS version 9.3 was installed on several computers at the State Archives so that the geospatial data could be viewed and validated.

Utah's archiving process began to take form in June 2008 when AGRC entered into a partnership with the State Archives to purchase a new server to be located in the Richfield Utah Data Center and to share the AGRC's server in Salt Lake City Data Center. There was not a set storage capacity at that time. Capacity was to be added as needed, with a limited storage set for imagery. The Utah team configured the server to house all the geospatial vector data and eventually all imagery submitted to the archives for retention.

As the data submitted to the archives was to be placed in a directory on the AGRC's Salt Lake FTP site and "pushed" down to the archives' FTP site in Richfield, the open source software rsync[61] was installed on the Salt Lake FTP server. It was to be used to transfer the data to the server in Richfield for permanent retention. Rsync has a process that takes place over a Secure Shell (SSH) connection which encrypts the file on the sending end and de-encrypts it on the receiving end, thus checking the integrity of the file. The transfer also included the utilization of the checksum feature contained within rsync. Additionally, AGRC installed the BagIt application to be used for validation during the Interstate data transfer. For Interstate data transfer Utah opted to make all their data available via their FTP site.

[60] For additional information on md5deep, see: http://md5deep.sourceforge.net/
[61] For additional information on rsync, see: http://samba.anu.edu.au/rsync/

Ready for Primetime – Implementing Intrastate Data Transfer

The movement of geospatial content from each state's geospatial division to its associated division of archives tested the boundaries of wholesale and ad hoc transfer, management and ingest of identified "demonstration datasets" of geospatial content into a digital repository environment. For North Carolina and Utah, it was anticipated that findings from this exercise would foster the creation of a geoarchive repository, while for Kentucky it provided an opportunity for further review of their current data transfer processes. All three state partners expected that investigating and documenting Intrastate data transfer would drive the development of useful practices for archiving geospatial content.

Prior to the actual Intrastate data transfer, the Content working group developed a proposed outline for documenting Intrastate data transfer for the different types of datasets included in the demonstration portion of the project.[62] Based on the recommendations from the Inventory working group, each state partner selected the types of datasets to capture as part of the data transfer. The dataset categories included:

- Local Government datasets
- Orthoimagery
- Centralized datasets (Framework and Non-Framework data)[63]
- Project Files (source data, GIS and map outputs, and project documentation)
- Digitized Maps

The North Carolina team also discussed the viability of capturing data displayed in NC OneMap's web mapping viewer for the archives.[64] Much of this data is hosted remotely by local governments and state agencies and is made available via Web Map Services (WMS).[65] The team determined that data made available via WMS would have little archival value due to the fact that the standard does not easily allow for the capture of adequate spatial information about the data, its attributes, and other elements important for the archival process.

There were three different approaches to Intrastate data transfer as each state was at a different starting point. It was natural that each state partner follow their own data preparation and data flow models based on their state's existing processes and procedures instead of trying to create and push identical processes across the three partners as part of the project. The intrastate data transfer processes[66] for each state document the process design and execution and suggest some possible next steps.

Kentucky

Kentucky had an established data transfer process in place and decided to continue to leverage these existing workflows for the Intrastate data transfer demonstration.

[62] See the outline at: http://www.geomapp.net/docs/Intrastate_design_outline_20090515.pdf
[63] See the glossary at the end of the paper for more information about Framework v/s Non Framework data
[64] NC OneMap viewer: http://204.211.239.202/viewer/
[65] For more information about the OCC's WMS standard see: http://www.opengeospatial.org/standards/wms
[66] To see the full intrastate data transfer processes for each state, see http://www.geomapp.net/documents.htm under the header **Data Transfer Documentation**)

All data and services published to the KYGEONET are required to have FGDC Content Standard-compliant metadata. The KYGEONET metadata service ensures that key metadata fields are completed before the data is published. The metadata is validated before datasets can be accessed. After the metadata is vetted, the dataset is then loaded into an ESRI SDE "staging" database and data validation tasks are performed to check file validity, topology, completeness, dataset projection, and geographic extent. Once the dataset has been reviewed and accepted, they are loaded into the KyVector or KyRaster SDE database.

All Kentucky vector databases (eight quarterly snapshots) and selective Kentucky raster data are copied to the archives by DGI via direct network connection using Robocopy, a file replication tool that hashes the files and verifies a complete transfer. Each snapshot is accompanied by a report that lists layers in the database grouped by category. Additionally, archives staff downloads various shape files and map files in PDF format from agency's websites for import into DSpace. Once the State Archives receives the database snapshots, it is verified by rerunning the hashing software. As a last step, an access entry is created in DSpace along with the snapshot report which provides documentation in a format that all users can open.

Before the GeoMAPP project began, DGI had previously transferred two copies of older imagery to the archives via hard disk and CD. Due to their large volume, DGI had merged the bulk of individual tiles of orthoimagery to form a single statewide "mosaic" imagery file for access via KYGEONET. However due to the large size of this mosaiced imagery and the limited existing storage space at the archives, this data was not initially transferred for archival purposes. Once storage is available, the periodic image snapshots, which date back to the 1990's, will be transferred to the archives when they are superseded in the KYGEONET. For the purposes of the demonstration project, 100 tiled images from two separate years were copied by DGI and later used for the Interstate transfer.

North Carolina

For North Carolina, the project transfer of geospatial data was the first accessioning of digital geospatial data by the archives. As data transfer between CGIA and the State Archives had not previously existed, the NC team built and implemented the geoarchive demonstration from the ground up. The team selected 28 datasets to transfer, equivalent to 739 GB.

In preparation for the transfer of data from CGIA to the State Archives for ingest into the demonstration repository, CGIA staff executed the following steps for all vector, raster, project and digitized map files:

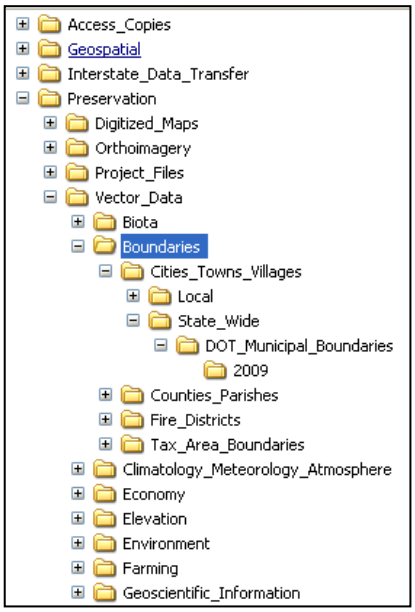
- Moved and consolidated identified demonstration datasets into a staging server located at CGIA;
- Converted all archived vector files to shapefile format. Any geodatabases were converted to shapefile format in the staging environment due to the adoption of shapefiles by the archives as the archival format for vector data;[67]
- The datasets were renamed to the State Archive's standard naming convention: ***Location*** (where appropriate) ***_Dataset name _Year_Month***;

[67] The NC State Archives followed the standards recommended by the North Caroline Geographic Information Coordinating Council (GICC). See http://www.ncgicc.com/Portals/3/documents/Archival_LongTermAccess_FINAL11_08_GICC.pdf for the standards for archiving geospatial public information.

- Utilized the BagIt application to bag the datasets and create a manifest. CGIA then transferred the vector data, project files and digitized maps via the Wide Area Network to a "Staging" folder on the GeoMAPP Storage Area Network (SAN). Orthoimagery datasets were transferred via an external hard drive.

Once CGIA transferred the data the archives staff completed the following steps:

- Ran a virus scan on all transferred datasets;
- Executed BagIt validation to ensure that the files transferred matched the manifest;
- Examined a random sampling of datasets utilizing ESRI's ArcCatalog and ArcMap software, verifying the metadata record, the display (meaning the datasets were viewable), table attributes and geographic extent;
- Added subject terms (or "themekeys" and "placekeys" in the FGDC Content Standard terminology) that were standardized, so that the datasets could be more easily found in the State Archives online catalog. Additionally, staff included additional information in the metadata to reflect the transfer of the data to the State Archives;
- Moved datasets into a file structure arranged into 4 series: Digitized Maps, Orthoimagery, Project Files and Vector Data. Within each series are further delineations primarily distinguishing whether the data represented is county-based or statewide. The datasets in Vector Data have a different folder structure, beginning with the classification of items first by ISO 19115:2003 categories and then by RAMONA GIS Inventory Data Layers. The items are further distinguished as either county based or statewide, followed by the dataset name and then by the year the dataset was published. The team also created a mirrored file structure for access copies of the data to maintain a separate copy for access and viewing independent from the restricted preservation copy.



NC geoarchive folder structure

For data discovery, the archives staff created an EAD finding aid at the collection level for the GIS datasets, projects, and digitized maps. The finding aid included information about the collection such as acquisition and processing, provenance, organization, and arrangement.[68] Moreover, staff entered the datasets information into the MARS online catalog for the North Carolina State Archives containing searchable descriptions of its archival holdings.

To investigate data access, the Archives staff chose to conduct a usability study. The main objectives were: (1) to investigate the effectiveness and efficiency[69] of discovering and accessing GIS demonstration

[68] See the GIS Data Collection Finding Aid at: http://www.archives.ncdcr.gov/ead/eadxml/gis_data_coll.xml
[69] Effectiveness is the measure of the ability of a program, project or task to produce a specific desired effect or result that can be quantitatively measured. Efficiency, on the other hand, is the skillfulness in avoiding wasted time and effort.

datasets via an archival finding aid and the MARS online catalog; (2) to assess user satisfaction, specifically gauging the quality of the user's experience interacting with both the finding aid and the catalog; and, (3) to measure how users expect to access the data. Overall the data suggested that user satisfaction with both the finding aid and the MARS catalog was mixed. While participants expressed some frustration with using these tools to find geospatial datasets and identified specific functionality, display and content issues with the finding aid and the online catalog, the participants also provided positive comments and identified features that were easy-to-use and problem-free. Ultimately, the usability study provided recommendations[70] that could be used as a starting point for the next phase of the GeoMAPP project.

Utah

As data transfer between Utah's Automated Geographic Reference Center (AGRC) and the State Archives had not previously existed, the Utah team built and implemented the geoarchive demonstration from the ground up. The team selected 19 datasets to transfer, equivalent to approximately 3 GB.

Before any geospatial data was submitted to State Archives, the team followed specific archival tasks prior to data transfer. These included:

- Verifying that there was a general retention schedule with which the data could be associated. If there was, the data could be linked to ISO 19115:2003 category themes. If there was not a retention schedule, a new one had to be created and approved by the State Records Committee;
- Completing the validation process in the archives directory on the AGRC FTP site. AGRC staff submitted or transferred the data to the archives directory using the agency's preferred transfer media or method. After the transfer, the data was validated before it could formally be added to the archives database. This validation included:
    - Opening the data in ArcGIS to ensure the data file is spatially valid (opens and is readable within a GIS environment);
    - Checking for a defined spatial projection.
- Using the ArcCatalog metadata editor, the team checked that the data was FGDC Content Standard-compliant metadata, and included the following information:
    - All areas marked "Required" within the editor
    - Data Contact
    - Attribute Accuracy
    - Positional Accuracy
    - Data Source Information
    - Process Steps
    - Complete Definitions of all Attributes
    - Distributor
    - Metadata Contact
- If the metadata was omitted or incomplete, AGRC staff contacted the owner or steward of the data so that the metadata could be completed to meet FGDC standards.

[70] See the full report and an example of the usability study at: http://www.geomapp.net/docs/20091229_Usability_Test_Final_Report.pdf

- Once the data were spatially validated, archives staff used a crosswalk to associate the submitted data title with the ISO 19115:2003 nomenclature adopted by the archives for geospatial series.
- AGRC staff ran a Visual Basic script to create a GeoPDF, a shapefile, and a file-based geodatabase for each dataset and then moved all these files to their correct folders in the archives directory on Salt Lake FTP site.

Before the datasets were "pushed" down to Richfield, it was proposed that a second script be run to create a separate metadata .xml file containing the name of the file, a descriptive title, abstract , creation date, file size, whether it was a shapefile or geodatabase, scale or resolution, projection, datum, extent coordinates, and keywords for a search tool. After the folders were pushed down to the Richfield FTP site, it was anticipated that the script would extract the URL of each GeoPDF, the zipped shapefile and geodatabase moved and added to the second metadata .xml. This second metadata .xml would then be sent to archives to record and store all the information about every submitted dataset in a finding aid. Due to time and budget constraints, the Utah team was not able to create this script; however, the team is hoping to address this in the next phase of GeoMAPP.

After the geospatial vector data were transferred to the archives' FTP site, a sample set of the data was opened and checked, using the same validation checks applied to the data when it was first submitted to the archives. If this sample set of data passed these checks, the transfer was considered accurate and complete. When the imagery was submitted and transferred to the archives' Richfield FTP site, a sampling of the imagery was downloaded, opened and examined to check for metadata and imagery corruption. If the sample imagery passed this validation, the transfer was considered complete.

Across the Great Divide – Transferring Data between States

So why transfer data between states? From a national perspective, the replication and movement of content between states demonstrates the viability of a robust state-to-state content exchange network that can serve as a point of access to non-federal content, and help meet the nation's requirements for preservation of at-risk geospatial data. At the state level, the practice of moving content in an organized way across jurisdictional boundaries can further state interests in national spatial data infrastructure, which supports many business processes requiring access to geospatial content. And finally, from each GeoMAPP partner's view, interstate data transfer affords the opportunity to not only to test a distributed archives for potential data sharing between states, but allows each partner to test each other's processes and validate internal assumptions about their own processes.

The DCAPE's (Distributed Custodial Archival Preservation Environments) NHPRC sponsored grant project is developing archiving rules to use in a GRID storage environment that manages multiple copies of records.[71] The NDIIPP supported FACIT project (Federated Archive Cyber Infrastructure Testbed) demonstrated great improvements in transfer speeds using L-Store technology to transfer geospatial records.[72] Each of these research projects assumes a heavy investment in infrastructure in order to effectively manage high volume file transfer and preservation in a network environment. While the GeoMAPP project was unable to test these methodologies using demonstration datasets, it would be

[71] For more information on Dcape, please see: http://www.dcape.org/
[72] For more information on FACIT, please see: http://www.ngda.org/facit.html

necessary to make the investment if institutions wanted a sustainable means for managing geospatial records exchanges between states. The project elected to test less costly means of transfer that have traditionally been used by the geospatial community.

The results from the Intrastate data transfer laid the groundwork for the second objective-- the movement of archived geospatial data between the state archives of the three GeoMAPP project partners. Using the *GeoMAPP Interstate Data Transfer Design Overview*[73] as a guide, each step of this process was to be documented, and include metrics, observations, issues encountered, lessons learned and recommendations for moving forward.

The selection of specific demonstration datasets was left to the discretion of each state partner; however each had to ensure they included datasets for the following categories:

- Local Government data;
- Centralized data including framework datasets common between partners (e.g. statewide roads, municipal boundaries, etc) and non-framework datasets unique to the host state;
- Orthoimagery;
- A consolidated geospatial Project unique to the state;
- Digitized products including scanned/digitized maps or aerial photographs.

The partners did not limit the type of format(s) (e.g., shapefiles, file geodatabases, etc.) or size to be included in the datasets. State partners were encouraged to share a diverse collection of datasets via FTP and external hard drive. Prior to the data transfer, the state partners discussed and gained consensus on the methodology for transferring each other's data. While data transfer options such as 3rd party SAN "landing" areas or other "drop services" for data transfer were explored, it was determined that implementation of one or several of these alternate methodologies was not feasible due to their complexity. External hard drives were chosen for the transfer of orthoimagery due to its large size.

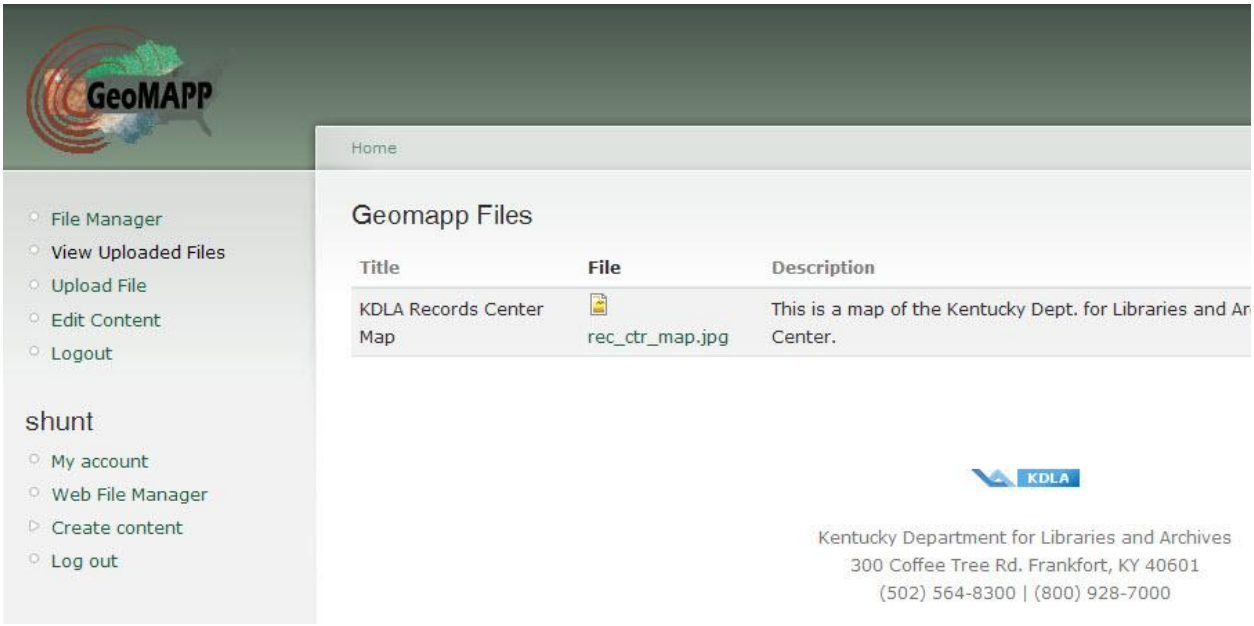
Kentucky

The Kentucky Department for Libraries and Archives (KDLA) transferred approximately 10 GB of files to North Carolina and Utah, including orthoimagery, 12 thematic centralized datasets, two project files and scanned/georeferenced digitized maps. All of the files were "bagged" (using the BagIt application) by Kentucky unless the volume of the files made the process too difficult, as was the case with mosaic orthoimagery files that had trouble being transferred even when they were not bagged. Each bag was verified before it was put on one of the transfer media.

During its utilization of BagIt, Kentucky encountered some difficulties but were able to correct them. While BagIt is a good means for institutions with various environments and software to use the software to exchange data, it does not perform a full format validation. Creators of bags need to be cognizant of file size. Kentucky never tested a bag that was larger than 6 GB. Bagging in general took two hours or less to transfer via FTP. Kentucky gave the North Carolina team vector data (file geodatabases), project files, and digitized maps on 3 DVDs during the partners meeting in September. They agreed that they would provide these same files plus approximately100 tiles of imagery for download via a file exchange

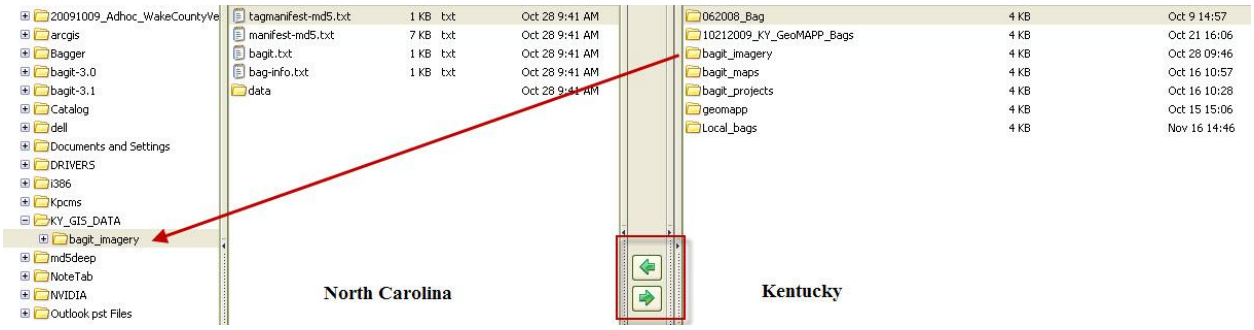
[73] GeoMAPP Interstate Design document: http://www.geomapp.net/docs/Interstate_Design_Guidelines_Final.pdf

website once they returned from the meeting. It was thought that the website could allow each state partner to both upload and download demonstration datasets.



The proposed Kentucky file exchange website will allow partners to create an account, upload/download demonstration datasets, and edit content.

Each of transfer methodologies had its own set of issues. The use of DVDs was effective for small bags but could not be successfully reconstructed into a verifiable bag if they spanned multiple DVDs. The hard disk transfer method worked satisfactorily when there was time to stage the transfer and the volume was not too big (1/2 a TB for example) to bog down the network in transferring between network and removable hard disk. The Kentucky file exchange service has to be used in conjunction with a secure ftp site. When datasets are "bagged," the bag can contain a large number of files, and the team file exchange web services were unable to accept the bag as a single unit. Resolution of this will be explored in the next phase of GeoMAPP. The alternate secure FTP was selected to make the Kentucky datasets available for download. This transfer methodology worked best when the transfer was done in incremental amounts and when using FTP transfer software.



Kentucky's SFTP server provided a simple way to transfer data. By selecting the dataset and then clicking the right arrow, the data could be easily transferred to where North Carolina stores its data.

At the very end of the project, KDLA and DGI began constructing a public access workstation that has ESRI software. The workstation will provide user access to open ERSI formatted file geodatabases and images stored both in the archives and the KYGEONET site. An instruction manual is planned for the workstation so that with some help from a knowledgeable archivist, the user can construct their own set of maps that can show change over time.

North Carolina

The North Carolina State Archives prepared over 15,000 files for data transfer, equivalent to 128.7 GB. This included compressed and uncompressed orthoimagery datasets, project files, digitized maps (both compressed and uncompressed digitized aerial photography), as well as five local and six centralized vector datasets.

Overall, it took approximately 15 hours to bag the data, with the uncompressed orthoimagery (5,831 files/117 GB) taking the most time (~14 hours). As this imagery was so large, the project team decided that North Carolina would make it available to both Kentucky and Utah via an external hard drive (versus an FTP server) for them to download to their respective hard drives during the face-to-face GeoMAPP partners meeting in September 2009. The process of copying the bagged orthoimagery to an external hard drive took nearly 4 hours as the data was being transferred over the network in the middle of the day (the highest peak of network traffic). Staff uploaded the remaining datasets to the NC FTP server, which took less than 15 minutes.

North Carolina provided the uncompressed orthoimagery via an external hard drive for Kentucky and Utah to transfer. All other types of data (vector, digitized maps and project files) were to be made available for download via a temporary FTP site. North Carolina DCR-IT policy does not allow for FTP sites, however for the purposes of this data transfer, the department Chief Information Officer did allow the team to establish a temporary FTP site which was disabled after the data had been successfully transferred. Initially, firewall and security issues prevented this site from functioning; however these issues were resolved and the data was made available for download.

The transfer of the uncompressed orthoimagery also presented challenges. The Utah team did not have enough storage on the laptop they had brought during the partners meeting to transfer the orthoimagery. As a result, they were only able to transfer the compressed orthoimagery which was much smaller in size.

Due to time constraints the Kentucky team was not able to copy the uncompressed orthoimagery to the external hard drive that they had brought for the partners meeting. As such, the Kentucky team took North Carolina's hard drive with them and then after copying the imagery mailed the hard drive back to North Carolina.

Utah

All the datasets were prepared for transfer by AGRC rather than the archives since the geoarchiving environment resides on AGRC servers. The transfer set included three local government datasets, three sets of NAIP imagery, 14 centralized vector datasets, project files, and USGS quadrangle maps (digitized maps). AGRC bagged the data using BagIt and copied it onto a hard drive, which traveled with them to North Carolina for the GeoMAPP partners meeting in September. Only some of the Utah data were contained on the hard drive. Utah agreed to make all their data available for download via an FTP site.

The Utah Archives tried to install BagIt on a desktop for the validation process, but could not get it to install properly. After several attempts at trying to configure it correctly, the decision was made to just move on and document their difficulties with BagIt.

Since BagIt was not able to be used for validation, the Utah team decided to try another free tool: Karen's Directory Printer.[74] Use of this tool provided a spreadsheet of all the files that had been downloaded, their location, file type, and totals (number of files and bytes) per folder, plus an MD5 hash string. When spot-checking, it appeared that the hash string generated in the BagIt manifest that came with the records did not match the hash string generated by Karen's Directory Printer. However, while thinking about how the archives could use the information obtained, the realization came that having this data in spreadsheet form would make it much easier to auto ingest into its database. If the agency sending the files uses a spreadsheet to outline the metadata needed per dataset (or the native XML metadata is parsed into delimited form using a tool such as Altova's MapForce or XSLT, and that is merged that with Karen's output), the archives could have the data in a format that could be imported into a series record with a single click.

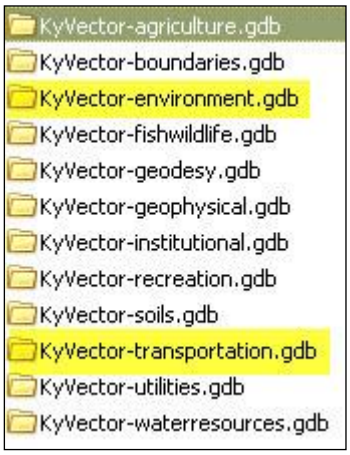
Looking Under the Hood – Technical Details and Challenges

Each state partner analyzed each other materials, paying particular attention to similarities/differences in file structure (or how the data is organized), file formats, naming conventions and metadata.

File Structure

For all three partners, file structure was the most similar when organizing vector datasets. The use of ISO categories to classify these datasets was implemented by each state. North Carolina appeared to be the most "traditional", using categories directly from the ISO standard, while both Utah and Kentucky used a mix of both standard ISO categories and state-generated categories. Regardless, it was relatively easy to assess how these datasets were organized.

[74] For additional information, see: http://www.karenware.com



Kentucky

North Carolina

Utah

While some ISO categories varied, several of the categories were identical.

The review, handling and archiving of the complex and varied formats contained within the GIS project files proved be a challenge to the group. In general, folder names were not always recognizable, and file structure was inconsistent and at times many layers deep. What made these file structures particularly challenging was the lack of accompanying documentation providing clues to the way the information was organized. Since quite a bit of "born digital" geospatial data is created as a result of projects, capturing the data, analyses and decisions that went into creating these datasets and the supporting project documentation will need to be an area of investigation for the project moving forward.

File Formats

All three state partners included a wide variety of file formats in their demonstration datasets. While a majority of them were shapefiles, file geodatabases or geospatial PDFs there were at least two dozen other "non-geospatial" such as PDFs and Microsoft Word documents, plus several unidentifiable file formats.

The presence of PDF files and other file types were not discoverable in ArcCatalog – only through Windows Explorer, which meant both tools had to be used to see what was in the root folders (not "one stop" in ArcCatalog). The geospatial PDFs provided by Utah added an extra layer of complexity as these types of files required the latest version of Adobe Acrobat Reader with a TerraGo plug-in. North

Carolina, did not have Adobe Acrobat Reader installed, therefore the team could not view the attributes associated with the geospatial PDFs, just the image itself.

The North Carolina team also encountered problems with file geodatabases. A "buffer overrun" error in ArcGIS would not allow the team to open several of the file geodatabases. The team discovered that they did not have the most recent version of ArcGIS (v. 9.3.1) which was needed to view those file geodatabases created in ArcGIS 9.3.1.

Naming Conventions

Each state approached naming conventions differently, as such some datasets were more challenging to discern than others. Deciphering orthoimagery file names presented the greatest challenge as they are not typically "human readable" names, but instead may be based on geographic position (such as Northing and Easting), geographic coordinates (such as longitude and latitude), state plane coordinate systems, scale size or any combination of these elements. Examples included "fsa_n07e079" (Kentucky), "0607_004" (North Carolina), and "q1217_drg24" (Utah).

Kentucky has begun to address the "mystery" of orthoimagery naming conventions by providing a grid map with all their orthoimagery. While not necessarily providing the user a "legend" for translating the file name, the grid map does assist users in determining where in Kentucky the tile is located.

The Kentucky grid map helps users find the right tile by providing the tile name in each cell.

Metadata

Encouraging data creators to provide a thorough metadata record is a challenge. This was apparent when evaluating each state's metadata records. Vector data seemed to have the most "complete" metadata, although completeness was in the eye of the beholder – meaning, what was considered required for one state may not have been required by another. In general, metadata was sparse for orthoimagery (with the exception of Kentucky), digitized maps and project files. Project files from all three states were the most

perplexing as nothing in the in the metadata records told why the project was undertaken, how the different fields related to one another, or whether the file hierarchy and naming scheme was required to maintain project integrity. As a test case, these records came to the archives without the paperwork (including a records retention schedule) that would have normally been required with transferred electronic records. Without this paperwork to give clues to the context of the data, at times each partner struggled to determine what exactly they were looking at.

What We Learned

Overall the both Intrastate and Interstate data transfer exposed similar challenges for all three state partners. Key lessons learned include:

- ***Metadata presents a myriad of complexitie***s that need to be addressed before any type of data transfer between the data consolidator and the state archives. For the state partners, the data consolidator was responsible for providing thorough and accurate metadata. However, as both entities have a vested interest in metadata, the challenges and subsequent solutions regarding metadata need to be discussed jointly. Challenges include:
  - How to handle missing and/or incomplete metadata, particularly from local data producers and digitized maps.
  - Determining which minimum "mandatory" FGDC Content Standard metadata elements are necessary for data transfer so that the archives can understand and interpret the data it received.
  - Deciding whether additional custodial and preservation metadata needs to be added and which agency is responsible for this task.
- ***Promotion of naming convention standards.*** While the archives can only influence naming conventions, meeting with data creators/ consolidators to explain how naming standards can affect preservation and access is critical as it can make the data more preservation ready at the source.
- ***Orthoimagery has a complex organizational structure*** as there are a variety of ways to delineate the files. These include color, scales, compression, and organizational structure. As these elements are driven by data creators, making an effort to communicate with data providers is a critical step in understanding orthoimagery.
- ***Data preparation and transfer times are time consuming***. The time necessary to prepare data for ingest into an archives increased as the complexity and size of the dataset increased. Bagging the data, while a very important step in the data preparation process, was admittedly a time consuming step, particularly for large datasets like orthoimagery. Data transfer itself was another time consuming process as it relied heavily on existing state networking infrastructure and transfer capabilities. As such, there was very little control over increasing the speed of data transfer. It is critical to meet with information technology staff to discuss the most effective methodology for data transfer.
- ***Data transfer requires detailed planning***, including:
  - Preparing for the unexpected. This can include technical issues, staffing issues – these can occur during the process, so build in extra time for the process.

  - Deciding lines of communication. It is important to determine at the outset which team members are going to be involved, at what level, and how best to communicate. At points in the process, there were too many individuals involved in the process thus muddying the lines of communication.
  - Determining who is doing what. While it helped that an individual on the North Carolina team was the "lead" for Interstate data transfer, when technical or content issues arose for each state, it wasn't always clear who needed to be contacted for the issue to be resolved.
- ***Each state's data is distinctly different*** which could present additional challenges for Interstate data transfer and ingest. This includes:
  - Naming conventions –particularly challenging were the often complex naming conventions for orthoimagery.
  - Organizational structure –while each state used some variation of the ISO categories for the root folder often the subcategories were completely different.
  - Metadata – when metadata was present, each state tended to include what it thought relevant. Regardless, each state partner recognized the importance of having "good" metadata.
- ***Current geospatial formats have both benefits and drawbacks.*** It seems likely that migration will be necessary for each of the formats and the difficulty of reformatting will depend on unpredictable future factors. The type of geospatial format depends on the goals established for long-term preservation; however, emphasis should be placed on format "openness" (whether proprietary or not), data portability, and the ease of data migration. In particular:

  - *Shapefiles* —ESRI's proprietary format has a single data theme or layer comprised of 5 required ancillary files needed to properly render the data. Shapefiles are an open, commonly used format with a metadata record that is easily identifiable/accessible .xml file contained with the other ancillary files. However, they are a complex object as multiple files are needed for dataset to function properly. Since it is a standalone dataset it is not possible to maintain relationships with other data. It is an older format that may be replaced by newer Geodatabase technologies.
  - *Personal Geodatabases* — an alternative ESRI format that uses Microsoft Jet Engine technology (MS Access) to store multiple geospatial datasets in a consolidated relational database. It uses a single .mdb file to store and manage datasets, and its database model allows for storage of multiple datasets in one place and to create/ maintain relationships between those data. Despite these positive features, personal geodatabases are in a proprietary format with size limitations of the MS Jet engine. Metadata is stored internally in tabular format and cannot be accessed without proprietary software. Personal Geodatabases have been replaced by File Geodatabases in ESRI's development priorities.
  - *File Geodatabases* — are the latest method by ESRI to store and manage data in a portable fashion. Uses proprietary system to create a folder and multiple ancillary files to store and manage geospatial data in a relational database format. It uses a relational database model to allow multiple datasets to be stored and managed in one place and it is possible to create/ maintain relationships between those data. File Geodatabases do not have the storage limitations that Personal Geodatabases have, however as with Personal Geodatabases, File Geodatabases are proprietary and are complex objects with multiple files that cannot be renamed as the system generated names are required for the database to function properly. Its

metadata is stored internally in tabular format and cannot be accessed without proprietary software.

- ***Additional "non-geospatial" formats created obstacles***. With a few exceptions, each state could open and view each other's geospatial formats. However, non-geospatial formats such as Excel spreadsheets and PDF documents and several unidentifiable formats (e.g., .gas, .aes) could not be opened in ArcGIS thus requiring the use of different applications to view the files. We should be cognizant of the different file types included with geospatial data and ensure that they are not only "openable" but also give some indication of how they are all related.
- ***Each of the data transfer methodologies had its own set of issues***. The hard disk transfer method worked when there was time to stage the transfer and the volume was not too big (1/2 a TB for example). It worked best during off hours when it did not bog down the network in transferring between network and removable hard disk. Secure FTP transfer seemed to work best when the transfer was done in incremental amounts and when an FTP client was installed.
- ***Moving data between states does not remove our obligation to provide full-context finding aids***. The finding aid can provide a direct link and context to the data regardless of where and how it is stored or what it is named, which is why obtaining the necessary information upfront to produce the finding aid is critical.
- ***Training on geospatial software is essential.*** It takes a certain skill set to understand what to do with these unique electronic records, and most archivists may not interact with them frequently enough to remember any training they do receive. Additionally, it is critical to ensure that the archives staff has the appropriate software to view various geospatial formats and its associated metadata.

Where We Go from Here

The process of data transfer not only exposed issues and challenges associated with moving and ingesting geospatial data, but also informed each of the state partners about gaps in their existing data transfer models and procedures and beneficial refinements to the processes that will need to be addressed for GeoMAPP 2010. It will be critical to continue transferring data internally and between states to further develop and improve best practices for content transfer efficiencies. The project will be adding new state partners who will be tasked with testing and validating current state partners' data transfer findings by moving content amongst themselves and also sharing their content with the existing partners and possibly the Library of Congress. Additionally, by further analyzing the similarities and differences of naming conventions, file structure, and metadata, the team can continue to explore various technologies and applications that provide the best access, discovery, and display of aggregated superseded content.

Getting the Word Out – Industry Outreach and Communication

Beyond exploring technical workflows, one of the most important elements of the GeoMAPP project has been conducting outreach to share the project's findings and experiences. The goal of outreach was in part to simply get the word out to the GIS, archives, and IT communities to raise awareness about the need to preserve digital geospatial data and the risks of not saving it. After these issues were identified to these communities, the team was then able to share specific recommendations about how to begin the archiving process based on project outcomes. The audience for the project's outreach efforts was intentionally broad and featured engagement with the GIS and archives communities nationally, statewide, within state

government and locally as well as informing GIS and archives vendors, the IT community and the public at large.

Engaging Industry

GeoMAPP engaged with the geospatial information technology industry to encourage awareness of the business value of preserved geospatial content and to help solution providers to understand the maintenance and use of preserved data as an important customer problem. The project prepared a draft industry outreach white paper that encourages the private sector to address data preservation-related use cases in their products, with the long-term hope that solution providers will find value in added capabilities that enhance the ability of users to interact with historic and superseded data and enable its preservation. GeoMAPP held talks with ESRI, Progressive Technology Federal Systems, Inc., APPX, and others to communicate issues that could potentially be addressed by industry engagement.

To support in-depth discussions with ESRI, the project developed a document outlining general requirements for consideration, with the goal of initiating a discussion surrounding the topic of enabling geospatial content preservation, retention, discovery, and long-term access. The document outlined a set of general requirements that addressed preservation and access needs with regard to superseded data, documented analysis processes, software, and output products. The expansion of the focus of discussion from just data to associated processes, tools, and products reflected project partner concerns about the repeatability of processes and the retention of outputs from those processes. These concerns reflected an understanding that there are legal requirements for some agencies or projects to emulate historic processes and results exactly as depicted at the time they were used for a decision. Other points of discussion outlined in the document included requirements for retrospective support of data and projects, the need for recommendations on procedures to be used in defined archival scenarios, and requirements for a content packaging solution supporting both data and its ancillary components.

Future project work in the area of industry outreach will include continued efforts to educate solution providers about the value of preserved data and customer needs with regard to maintenance and use of preserved data. There will be opportunities to engage in further discussions around specific technical concerns that will need to be addressed in the ongoing development of commercial products. One targeted area of interest includes support for long-term and open access to spatial databases. The project will also have an opportunity to continue the previous work of the North Carolina Geospatial Data Archiving Project in engaging the Open Geospatial Consortium (OGC) industry standards body on preservation issues related to existing and evolving OGC specifications.

Communication Methods

To support project outreach activities, GeoMAPP created the Communications team as one of the project's first working groups. The team was primarily responsible for external communications regarding the work of the GeoMAPP project. In order to facilitate this, the group established the GeoMAPP website (www.geomapp.net) and provided regular updates to the site with news, project documents, reports, and presentations. Early in the project, Kentucky established a wiki for the group to communicate and share information, however due to technical barriers; the tool was not utilized to its full potential.

The Communications team also took the lead on producing a project logo for GeoMAPP and a project brochure. The colorful brochure provides introductory material regarding the work that GeoMAPP is doing and the challenges it faces. In the fall of 2008, 3,000 copies were printed and the brochure has been widely distributed at events across the country and within each state. Brochures have also been shared with decision makers including state and national legislators.

The responsibilities of this team will transfer to a new working group named Outreach and Mentoring for GeoMAPP 2010. Responsibilities will include creating a new logo with new state partners, continued website updates, publication of an updated brochure and a possible venture into Web 2.0 technologies and a project video.

Project Outreach

A significant focus of the GeoMAPP effort has been to conduct outreach to highlight the issues surrounding the preservation of geospatial data. This outreach includes sharing project findings and lessons learned, simply to make people aware that it is important to preserve geospatial data and not just delete or overwrite older datasets when updates are made. Project participants have been conducting both formal and informal outreach since the project's inception and keeping these lines of communication open will be a critical component of future efforts.

Outreach within the Partner States

One of the most important relationships developed during the project was between the project team and the geospatial data producers within each state. In Kentucky, the initial outreach performed during the project entailed notifying all KYGEONET publishers that their data would now be archived. Many of the publishers have found great value in this and some see it as one of their main reasons for publishing their resources and creating all associated metadata records. The Kentucky team followed this effort with the launch of its Local Government Geoarchiving survey and followed up with meetings with various agencies using GIS to develop records retention schedule items. Additionally, two long-standing local government GIS creators were approached regarding the project. An overview of GeoMAPP was given to the entities and then each was asked if they would like to participate in data sharing. Both agreed to work with DGI and KDLA on archiving some select datasets. A Memorandum of Understanding was developed and executed with each entity before the data sharing process began. Superseded data has now been acquired from each entity.

The North Carolina team leveraged its state Geospatial Information Coordination Council to reach out to the state's GIS community. The North Carolina GeoMAPP team provides regular project updates to the full council and has worked directly with both the Local Government Committee and the State Government Users Committee to share project findings, explore records retention scheduling options and to seek partners for participation with project tasks. The project team also leveraged the local and state agency surveys to help spread the work about the project aims. Additionally the team presented at the 2009 North Carolina GIS Conference while also performing outreach at several 2009 GIS Day events at local universities and for Wake County. The North Carolina team has also worked on internal outreach and marketing within CGIA's technical staff and among State Archives and State Library records analysts and reference staff.

Utah has conducted extensive outreach with local governments and regional organizations by travelling and talking to data creators across the state. This outreach has yielded positive outcomes as the project team was able to take inventory of data being created by localities and capture critical data for the archives. Talking to local governments helped make the localities aware of GeoMAPP and the internal Utah archiving efforts, and by extension, awareness of the State Archives' other services. Rural counties have been very receptive to Utah's archiving program and open to the inventorying of data, but have required the State Archives or AGRC's help to perform inventory entries. The urban cities and counties have a more complex situation and have their own archives so they were likely to follow their own procedures rather than that of the state. Some of these more urban counties did not inventory and archive geospatial data, as they did not feel the need to do so.

Outreach to local and state government GIS creators is a critical element of the project, as these groups produce the majority of essential data within each state. Each state will continue their Intrastate outreach during the next phases of the project, with the intent of furthering relationships and potentially opening the door to gain access to data that was not previously being archived.

National Outreach

In addition to reaching out to state data producers, the project also engaged other states and GIS and archives decision makers and thought leaders nationally through the project surveys and participation in national conferences. During the course of the project, the team has directly engaged with the following national organizations/events through either direct discussions or presentations at events:

- The National States Geographic Information Council
- The Society of American Archivists
- The Council of State Archivists
- The National Association of Government Archives and Records Administrators
- The Open Geospatial Consortium
- The National Association of State Chief Information Officers
- ESRI
- Society for Imaging Science and Technology
- American Society of Photogrammetry and Remote Sensing
- American Congress in Surveying and Mapping
- Best Practices Exchange for Government Digital Information
- National Digital Information Infrastructure and Preservation Program

This national engagement has provided great feedback to the GeoMAPP effort and has been a good opportunity to share project findings with a wide audience. The project intends to continue its outreach moving forward with a greater focus on engaging the GIS community and sharing best practices and recommendations to make data more "preservation ready" at the source. In addition to planned presentations to the ESRI Federal Users Conference and the American Society of Photogrammetry and Remote Sensing in early 2010, an article providing a recap of the project's accomplishments during its first two years was published in the *Journal of Map and Geography Libraries* in January 2010.[75]

[75] Article abstract and ordering info can be found here: http://www.informaworld.com/smpp/title~db=all~content=g918168567~tab=toc

GeoMAPP Lessons Learned/ Key Findings/ Best Practices

A key component of the GeoMAPP effort has been to test and implement preservation workflows, document key findings and compile useful recommendations and observations that can be shared with other state and local governments to assist them in addressing their geospatial preservation concerns. Highlights of these efforts are listed here:

Collaboration is Key

- **Learning from sharing --** While each state partner took a unique approach to investigating and approaching the preservation and long-term accessibility of superseded geospatial data, each also realized the value of sharing information and lessons learned. Frequent communication, in the form of monthly team meetings, gave state partners the opportunity to communicate experiences and learn about things that did or did not work in the other states. This not only made implementing policies and systems much easier than starting from scratch, but also provided ample opportunities for the partners to come to agreement on project-wide generalized recommendations, best practices and standards.
- **Get to know your partners --** GeoMAPP enabled each of the state partners to build or enhance the relationship between their state archives and GIS agencies. Providing formal cross-training between the two agencies on archival processes and GIS tools and technologies, integrating both archives and GIS staff when conducting outreach, and holding regular meetings as well as less formal interactions, allowed the groups to build familiarity with each discipline's standard terms, jargon, workflows and responsibilities. These interactions highlighted similarities and began the process of diminishing differences between the two groups in each state. By understanding each other's language, responsibilities and goals the state teams were better prepared to tackle the challenge of preserving geospatial content together.
- **Spread the word using existing networks --** Leveraging existing geospatial/archival relationships is critical for developing a unified approach to preserving geospatial content. Each state partner reevaluated the relationships between the statewide GIS coordination bodies and the internal state and local government archives staff and their counterparts (i.e., chief records officers, clerks, etc.) in state and local government agencies that produce GIS data. Tapping into these relationships can catalyze a strong interest in preserving geospatial content.
- **Get IT buy in --** Engaging agency information technology staff at the beginning of the project and continuing discussions throughout the project aided with the selection, implementation and support of various technologies that enable archiving of geospatial content.

Find Out What You Know and What You Don't Know

- **Say hello with a survey --** While time-consuming to develop, manage and analyze, conducting surveys targeting GIS data producers as well as GIS and archival leadership and codifying the results can help identify the current state of geospatial archiving and the preservation landscape within state and local government. Surveys also perform critical outreach as they inform groups that the preservation of geospatial data is important and is worthy of being investigated.

- **Keeping track of things with an inventory --** Having a method to track which geospatial datasets exist in your organization is an essential first step in deciphering the preservation puzzle, whether you use a simple spreadsheet or a complex inventory database such as the national Ramona GIS Inventory. An ideal inventory should contain basic information such as the title of the dataset, the creation date, who created it/owns it, where the dataset physically resides, how often the dataset is updated and ideally provide a method to categorize/organize the data based on a widely recognized standard and/or significant keywords.

Understand What Needs to Be Preserved, and Why

- **Think "archival record" --** Recognizing that geospatial content needs to be ingested, preserved, and made accessible for long-term future use, should compel both GIS data consolidators and their state archives counterparts to think carefully about the utility of the information as a record.
- **Connect with internal resources to get the ball rolling --** Engaging internal state and local government archives staff to work with data producers to investigate an organization's geospatial data holdings and to begin discussion of how to integrate these data into a record retention schedule is an important early step in building a geoarchive. These initial discussions not only open lines of communication and inform the process of what type of content is out there, but also encourages staff to regard geospatial data as something that should be preserved. These early discussions can help shape the approach to data ingest and management. Important things to capture during these conversations include: which data are important to archive; how often these datasets need to be archived (frequency of capture); and the mechanism by which the data can be transferred to the archives.
- **Make it official - put it in writing --** Developing geo-centric records retention schedules is an effective way to ensure that geospatial data worthy of long-term preservation is retained and transferred to the archives. Retention schedules give archives staff the opportunity to talk to the data creators about retaining this data, and helps the data creators identify, organize and dispose of geospatial records once they've reached the end of their active lifecycle.

Spread the Word

- **Develop a web presence --** Employing a variety of methods for external communication about your project informs project participants and can "get the word out" to other interested parties. Wikis can be beneficial to an internal team, while a project website can provide regular updates with news, project documents, reports and presentations to a wider audience.
- **Hit the road --** Conducting extensive face-to-face outreach efforts with local governments and regional professional organizations can yield positive results by creating a greater understanding of the need to preserve superseded geospatial content. Outreach to local and state government GIS creators--the leading data creators within each state--is a crucial element in developing a successful geoarchive. Sharing best practices, whether in the form of presentations, training sessions, or one-on-one meetings, will help make the geospatial data more preservation ready at the source.
- **Influence change --** Actively communicating with the organizations and vendors that create the tools and technologies used in the archival and GIS communities about the challenges of archiving geospatial content could influence future product changes.

- **Attend conferences--** Local and national conferences provide an opportunity to share project findings with a wider audience and inform others about techniques that could be implemented in their preservation practices. It is an important feedback mechanism for learning about the successes and/or obstacles others may be facing when capturing and preserving geospatial content.

Put Data "In Motion"

Designing and testing workflows to move geospatial content between agencies can compel both GIS and archival agencies to address data-related challenges. The experiences of the GeoMAPP team members with the handling and ingest of geospatial data into an archival setting yielded the following observations and recommendations:

- **Metadata** - Just do it
  - Ideally, an FGDC Content Standard-compliant metadata record should be included with any dataset that is to be shared or archived to assure future access and use; however, prior to any movement of data, a decision must be made to determine the level of completion that is acceptable to both the GIS data consolidator and the archives.
  - As the level of metadata completion appears to vary with the complexity of geospatial format it is critical to be consistent with all data formats and provide the most complete metadata possible.
  - Moving data between states does not remove our obligation to provide full-context finding aids or the paperwork normally required with accessioning records. The finding aid can provide a direct link and context to the data regardless of where and how it is stored, or what it is named. Obtaining the necessary metadata upfront to produce the finding aid is critical.
- **What's in a name? -** File Naming Conventions
  - Assigning a logical file name to a geospatial dataset is important for identification and management of that data. Use a system that works best for your workflow, but information about geographic extent/location, data theme, and creation date are useful attributes that may be worth capturing in the file name. (e.g. SaltLake_Parcels_2006)
- **Format matters -** Versioning and Format Awareness
  - The type and version of the GIS software being used to view geospatial content does matter. Interoperability between different vendors is always a concern in the digital world. The GeoMAPP team, which relied on ESRI software for data review and analysis, also discovered that differences between product releases can impact data preservation and access.
  - Selection of a type of geospatial data format for preservation depends on the goals established for long-term preservation; priority emphasis should be placed on format "openness" (whether proprietary or not), community uptake, data portability, and the ease of data migration.
- **Putting it all together -** Data Packaging
  - Geospatial datasets are often comprised of a number of related files that must be present in order for the set to be complete and functioning. One fundamental challenge of archiving geospatial data is the lack of existing standards for packaging these disparate files in a consistent way. While each state had a unique approach for determining what information should be included in the archival package, it was universally agreed that the most complete

metadata available for the dataset must accompany the data. Trying out new tools not only allowed the state partners to successfully transfer geospatial content internally and externally, but also reaffirmed the importance of data validity and authenticity when moving and ingesting electronic records.

- **Size matters -**Data Storage
  - Knowing the overall size of your geospatial content, understanding your state's network infrastructure and capacity, determining how the data is to be transferred between entities (i.e., over the network versus external hard drives), and examining the pros and cons of different storage media and their costs are critical pieces of information when selecting data storage. It is essential to work with information technology staff from the beginning of the process to ensure the most effective outcome for data storage.

Overall Project Observations

- **While collaboration is fruitful, it can be a challenge to manage.** Building a successful geoarchive is a complex process that requires time, resources and both personal and organizational commitment to get things accomplished. In addition to the uniqueness of having disparate state agencies collaborate within each state and with other states as well, tight budgets, conflicting priorities, and limited staffing resources can create additional challenges.
- **There is no "one size fits all" approach to preserving geospatial content.** Currently there are no out-of-the-box solutions for archiving geospatial data; however, this is not necessarily negative. Using a customized approach allows an organization to take advantage of existing workflows and relationships and was more effective than deploying a unified "cookie cutter" approach.
- **Eat the elephant in small chunks.** It is best to take a modular approach to archiving geospatial data, starting with small steps and building the program over time. Trying to address all of the challenges of inventory, appraisal, outreach, system design, system implementation, data transfer, long-term management, and data access at one time is extremely challenging.
- **How you handle unanticipated events can affect the project outcome**. During the project, each state partner faced the loss of key project staff, including project leaders and champions, general staff and budget reductions, and organizational changes. To help cope with the "unexpected":
  - Build extra time in the project plan to handle technical, staffing, and budget issues. Know that these can and likely will occur during the process, so try to be prepared.
  - It is important to determine at the outset which team members are going to be involved, at what level, and how best to communicate when something unexpected arises. Particularly during stressful times, knowing who is responsible for what can solve problems more quickly.
  - Through your outreach efforts, develop champions within and outside of your organization who can help keep the effort alive if critical staff departs.

GeoMAPP Next Steps: A Growing Partnership

Creating a New Venue for Discussion and Outreach: GeoMAPP Informational Partnership

On October 21, 2009, GeoMAPP kicked off a new method of project engagement with the creation of an Informational Partners Program. The program aims to engage a new set of state archivists and geospatial data managers with the activities of GeoMAPP by sharing findings from the project and providing mentorship to participating states. The new partners will also expand perspectives on developing best practices for the preservation and long-term stewardship of geospatial data.

The engagement will take the form of regularly scheduled collaborative sessions and meetings that will allow GeoMAPP to share the concepts, business planning methods and strategies that have been explored over the course of the project, including ideas to assist the informational partner states in garnering support for their own geoarchiving efforts. The team held two sessions in late 2009: a program kickoff where existing partners shared information and fielded questions about GeoMAPP's initial activities and findings, and a second session where each informational partner described how their geospatial and archives groups were organized, the status of geoarchiving efforts in their state and their expectations for the project. Comments and questions from both sessions highlighted a general need for documentation to help states get started with building geoarchiving programs and to craft business planning documentation to help with program funding.

A significant outcome of GeoMAPP's early efforts has been the project's ability to build collaboration across institutional and state boundaries, and these efforts will continue with the informational partners. It is anticipated that expanding the project to include the voices of ten new states will benefit the findings and output of the project while also providing a resource for shared knowledge and forum for discussion for the informational partners as they explore implementing or expanding geospatial preservation efforts within their states.

The complete list of new informational partners:

- District of Columbia Office of Public Records and the Office of the Chief Technology Officer
- Georgia Archives and Information Technology Outreach Services Division
- Maine State Archives and the Maine GIS Office
- Maryland State Archives and the Maryland Department of Natural Resources
- Minnesota Historical Society and the Minnesota Dept. of Administration, Geospatial Information Office
- Montana State Library
- New York State Archives and the New York State Office of Cyber Security and Critical Infrastructure Coordination
- Texas State Library and Archives Commission and the Texas Natural Resource Information System
- Wisconsin Department of Administration and the University of Wisconsin-Madison
- University of Wyoming

Extending, Expanding and Refocusing the Partnership: GeoMAPP 2010

Based on the initial success of the GeoMAPP project, the Library of Congress awarded additional grant funding to the GeoMAPP team to extend its investigation. The initial partnership built a solid foundation though forging relationships between archivists and GIS practitioners. The team identified some initial challenges with inventorying, appraising, transferring and ingesting geospatial data and created unique approaches to begin to address these issues. GeoMAPP's research and outreach aims will continue in 2010 with at least two new full-time partners and ten informational partners joining North Carolina, Kentucky and Utah in the GeoMAPP 2010 effort.

GeoMAPP 2010 will continue to explore areas of investigation from the initial phase of the project as well as a more heightened focus on the following key areas:

- Development of portable tools for creation of business planning documentation in support of building sustainable geoarchives.
- Exploring tools and techniques to provide better access to superseded geospatial content.
- National, statewide and local outreach and engagement with the geospatial and archives communities.
- Mentoring of new full-time partners and informational partners.
- Validation of existing best practice recommendations and creation of new lessons learned and project best practices.
- Investigation of tools and techniques to better support content transfer, data packaging, records discovery and cataloging.
- Compilation of a final project report that details the findings across the entire project.

The highlights of next phase of the project, including testing and refining the initial best practices, refocusing efforts to investigate business planning documentation, and providing access to data that has been archived, are aimed to help better inform other states who may be interested in developing geoarchives while also expanding the abilities and offerings of the participating states.
The team will also continue to support the Library's "Geospatial Data for the National Collection Initiative." This effort kicked off in November of 2009 with a two day GeoSUMMIT in Washington, DC, which drew leaders from federal, state, and local government agencies; archives and libraries; creators, disseminators and users of geospatial data; and other stakeholders to develop a national strategy for preserving and supporting enhanced access to geospatial data.

Glossary of Archival and GIS Terms

Archival Record - materials created or received by a person, family, or organization, public or private, in the conduct of their affairs that are preserved because of the enduring value contained in the information they contain or as evidence of the functions and responsibilities of their creator. Archival records may be in any format, including text on paper or in electronic formats, photographs, motion pictures, videos, sound recordings. (Society of American Archivists [SAA] Glossary)

Archival Value - the ongoing usefulness or significance of records, based on the administrative, legal, fiscal, evidential, or historical information they contain, justifying their continued preservation. In general, records with archival value are estimated to make up only three to five percent of an organization's records. (SAA Glossary)

Attribute Data - generally defined as additional information about each spatial feature housed in tabular format.

Checksum - a checksum is calculated from the data using a known formula that returns a single-digit value and is stored with the data. At any point the checksum can be recalculated to see if the value has changed. (SAA Glossary)

EAD (Encoded Archival Description) –standard used to mark up (encode) finding aids that reflects the hierarchical nature of archival collections and that provides a structure for describing the whole of a collection, as well as its components. (SAA Glossary)

ESRI (Environmental Systems Research Institute) - providers of GIS Software (such as ArcGIS).

Feature - natural and man-made geographic features represented by points/symbols, lines, and areas on a map. Object in a geographic or spatial database with a distinct set of characteristics. For example, a road segment, manhole, building, or area designated having the same soil type. *(http://www.kansasmappers.org/kam/services/gisdictionary.cfm#P)*

FGDC (Federal Geographic Data Committee) – promotes sharing of the nation's geospatial resources. The FGDC metadata standard was adopted in 1994, and is a standard developed to determine the robustness, the method of accessing, and the successful transfer of geospatial data. *(http://www.fgdc.gov/metadata)*

Finding Aid - a description of records that gives the repository physical and intellectual control over the materials and that assists users to gain access to and understand the materials. (SAA Glossary)

FTP (File Transfer Protocol) - is used to transfer files between computers on a network, such as the Internet. You can use FTP to exchange files between computer accounts, to transfer files between an account and a desktop computer, or to access software archives on the Internet. Keep in mind, however, that many FTP sites are heavily used and require several attempts before connecting.

Geoarchive – refers to a digital records repository designed to ingest and manage archived geospatial content.

GIS (Geographical Information Systems) – incorporates graphical features with tabular data in order to assess real-world problems (e.g., prioritizing sensitive species habitat to determining optimal real estate locations for new businesses).

At the simplest level, GIS can be thought of as a high-tech equivalent of a map. The key word to this technology is Geography – this usually means that the data (or at least some proportion of the data) is spatial, in other words, data that is in some way referenced to locations on the earth. Coupled with this data is usually tabular data known as attribute data. Attribute data is generally defined as additional information about each of the features, which then can be tied to spatial data *(GIS Lounge: http://gislounge.com/what-is-gis/).*

Framework Data – geospatial datasets deemed to be the most critical or commonly used for a wide variety of mapping and analytical purposes. The RAMONA GIS Inventory database delineated the following 23 datasets as being "framework": **Boundaries**- American Indian Reservation, Cities/Towns/Villages, Civil Township, Counties/Parishes, and State; **Elevation**- Bathymetric Contours, Contours, Digital Elevation Model (DEM), Digital Surface/ Terrain Models (DSM/DTM), and Spot Elevations; **Imagery/Base Maps/Earth Cover-** Digital Orthophotography/Orthoimagery, and Land Cover; **Inland Waters-** Hydrography, and Watershed Boundaries; **Location-** Address Points, Geodetic Control Points, Geodetic Networks, and Geographic Place Names; **Planning/Cadastral-** Parcel/Cadastral/Land Ownership, and PLSS Townships & Sections; **Transportation-** Airports & Airfields, Railroad Lines, and Roads/Streets/Street Centerlines.

Geospatial Metadata – provides a way to describe geospatial data and other electronic records. It contains such information as the coordinate system, when the data was created, when it was last updated, who created it and how to contact them and definitions for any of the code attribute data.

Hash function - a mathematical algorithm that takes an electronic document and creates a document fingerprint. The document fingerprint is much smaller than the original document, and does not allow the reconstitution of the original document from the fingerprint. A slightly different document, processed through the same hash function, would produce a very different document fingerprint. A hash function helps to secure data by providing a way to ensure that data are not compromised (*http://www.records.ncdcr.gov/guides/AH_Best_Practices_Digital_Preservation_Final_2008_04_01.pdf*)

ISO (International Organization for Standardization) 19115 - defines the schema required for describing geographic information and services. It provides information about the identification, the extent, the quality, the spatial and temporal schema, spatial reference, and distribution of digital geographic data. (*http://www.iso.org/iso/catalogue_detail.htm?csnumber=26020*)

Layer - the visual representation of a geographic dataset in any digital map environment. Conceptually, a layer is a slice or stratum of the geographic reality in a particular area, and is more or less equivalent to a legend item on a paper map. On a road map, for example, roads, national parks, political boundaries, and rivers might be considered different layers. (ESRI Glossary: *http://support.esri.com/*)

MARC (Machine Readable Cataloging) - data communications format that specifies a data structure for bibliographic description, authority, classification, community information, and holdings data. (SAA Glossary)

Metadata Schema - A metadata schema defines a framework for representing metadata. In general it includes definition of terms used in the schema, structural constraints and data structure definitions, and bindings to physical description syntax. For additional information, see: http://www.ieee-tcdl.org/Bulletin/v3n1/nagamori/nagamori.html.

Metadata Wrapper - In general, a metadata wrapper would contain all additional bits of metadata elements including descriptive, administrative, technical, and structural metadata.

Network - a number of computers connected together to share information and hardware. A Local Area Network (LAN) is small, usually confined to a single building or group of buildings. A Wide Area Network (WAN) is a system of LAN's. It is large, with many computers linked.

Non-Framework Data – see Framework Data. Any geospatial datasets that are not included in the list of 23 Framework datasets listed above.

OAI-PMH (Open Archives Initiative – Protocol for Metadata Harvesting) – published by OAI, the protocol defines an application-independent interoperability framework based on metadata harvesting. The framework is used by data providers, who expose metadata about information held in a repository, and by service providers, who use that metadata to build value-added services. See http://www.openarchives.org/.

Orthoimagery – digital imagery in which distortion from the camera angle and topography have been removed, thus equalizing the distances represented on the image. A rectified copy of a photograph (typically an aerial photograph), showing image features corrected for variations in scale and height displacements. (*From http://www.websters-online-dictionary.org/or/orthophotography.html)* Aerial photographs that more precisely show the features of the landscape, including those that might be important for agriculture such as slope or size of gullies, because they are corrected for distortion caused by tilt, curvature, and ground relief.

Permanent Record – see Archival Record.

Projection - a system to portray all or part of the earth, which is an irregular sphere, on a planar or flat surface.

Raster Data – cell-based data such as aerial imagery and digital elevation models. Raster data is characterized by pixel values. Basically, a raster file is a giant table, where each pixel is assigned a specific value from 0 to 255. The meaning behind these values is specified by the user- they can represent elevations, temperatures, hydrography, etc. Satellite imagery uses raster data to record different wavelengths of light. Raster data is advantageous to vector data in constructing 3D images, as the values for every pixel are calculated through a process called *interpolation (http://www.umich.edu/~ipcaa/GIS/General%20GIS%20Concepts.htm).*

Record - data or information in a fixed form that is created or received in the course of individual or institutional activity and set aside (preserved) as evidence of that activity for future reference. A record has fixed content, structure, and context. (SAA Glossary)

Retention and Disposition Schedule - a document that identifies and describes an organization's records, usually at the series level, provides instructions for the disposition of records throughout their life cycle. (SAA Glossary)

Spatial Data - also known as *geospatial data* or *geographic information* it is the data or information that identifies the geographic location of features and boundaries on Earth, such as natural or constructed features, oceans, and more. Spatial data is usually stored as coordinates and topology, and is data that can be mapped. Spatial data is often accessed, manipulated or analyzed through Geographic Information Systems (GIS). *http://www.webopedia.com/TERM/S/spatial_data.html.*

Spatial data = Spatial (Where) + Data (What)

Spatial Data Clearinghouse – repository structure, physical or virtual, that collects, stores, and disseminates information, metadata, and data. A clearinghouse provides widespread access to information and is generally thought of as reaching or existing outside organizational boundaries. (Wade, T. and Sommer, S. eds. *A to Z GIS)*

Storage Area Network (SAN) - is a network specifically dedicated to the task of transporting data for storage and retrieval. SAN architectures are alternatives to storing data on disks directly attached to servers or storing data on Network Attached Storage (NAS) devices which are connected through general purpose networks.

SDE (Spatial Database Engine) - refers to ESRI's spatial database engine. It is a relational database management system that provides a formal structure for storing and managing information in tables. For additional information, see: http://www.esri.com/software/arcgis/geodatabase/storage-in-an-rdbms.html.

Temporal – existing for a time only. As an example, in the Dublin Core elements, temporal can be defined by a date, date range or a named period.

Vector Data –spatial data represented as points, lines and polygons. This system of recording features is based on the interaction between arcs and nodes, represented by *points, lines, and polygons.* A point is a single node, a line is two nodes with an arc between them, and a polygon is a closed group of three or more arcs. With these three elements, it is possible to record most all necessary information (*http://www.umich.edu/~ipcaa/GIS/General%20GIS%20Concepts.htm*).

Web Map Service (WMS) – is an Open Geospatial Consortium (OCG) Web service standard for exchanging map information as map images. WMS allows a user to request map images over the web using open standards. WMS supports the use of datasets without the need to keep a local copy. *(http://www.lib.virginia.edu/scholarslab/resources/class/mlbs)*

XML (Extensible Markup Language) - a standard to promote sharing information over the Internet by specifying ways to describe the information's semantic structure and to validate that the structure is well formed. XML is described as extensible because it is a metalanguage that allows the creation of tags to be used for semantic markup. The greatest value of XML can be realized through common vocabularies of tags, so that applications can use those vocabularies to understand documents from different sources.