North Carolina Intrastate Data Transfer Design

Prepared by: NC CGIA and State Archives for the GeoMAPP Project

June 10, 2009

Revised: September 14, 2009

Introduction

This document was created to chronicle the full life cycle of intrastate data transfer, with the intent of following the standards recommended by the NC GICC (http://www.ncgicc.com/Portals/3/documents/Archival_LongTermAccess_FINAL11_08_GICC.pdf) for archiving geospatial public information. This design document is divided into two sections. The first section, the Data Transfer Pilot Design, was originally written and approved June 10, 2009 and it is the document that the North Carolina team used to build and implement the geoarchive demonstration. The following section, Data Transfer Evaluation, documents the actual execution and resulting issues found with original plan design, and provides next steps with data transfer and ingest.

Data Transfer Pilot Design

Introduction/ Background

The North Carolina demonstration Geoarchives system design process began in the summer of 2008 with discussions between the Center for Geographic Information and Analysis (CGIA), the North Carolina State Archives, and North Carolina State University Libraries about potential geospatial datasets that should be preserved, as well as tools and techniques that would be needed to capture, preserve and manage superseded content. The team also began an initial assessment of the infrastructure necessary to build the demonstration repository. In the first several months of 2009 the North Carolina team focused on dataset selection and sizing. In anticipation of the transfer of data and based on the size estimates, the NC 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 NC team based the initial database sizing in part on the size of the total holding of NC OneMap (~14 TB uncompressed). Available funding also influenced the capacity of the demonstration storage environment. 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.

Dataset Selection

The GeoMAPP Content Lifecycle and Data Transfer Working group, a group consisting of project members from each state, selected the types of datasets to capture as part of the demonstration portion of the project. The dataset categories include:

- Local Government datasets
- Orthoimagery
- Centralized datasets (Framework and Non-Framework data)
- Project Files
- Digitized Maps

The North Carolina team targeted framework local government datasets such as parcels, streets, and zoning from two counties: Wake and Dare.[1] The team selected these counties due to Wake's urbanized landscape and Dare's coastal positioning as well as the fact that that NC OneMap had multiple snapshots of orthoimagery for each of these counties. The selection of centralized GIS datasets was focused on two themes: framework datasets that were common among the partner states and non-framework datasets that were unique to North Carolina including a large number of coastal themed datasets. The coastal theme is continued with the selection of digitized aerial photos from Dare County from 1947 to meet the digitized maps obligation. With regards to project files, the team selected the CGIA produced "Sustainable Sandhills Land Use Modeling" project output products. It was selected because it contains a large and diverse collection of geospatial data, analytical scripts and documentation products that are well organized in a consolidated folder structure.

The total size of the demonstration data holdings is almost 1 terabyte including:

- 2.6 GB of Local Government vector datasets
- 33.4 GB of compressed Orthoimagery and 882 GB of uncompressed copies of the same imagery
- 15.5 GB of Framework and Non-Framework Centralized vector datasets
- 3.2 GB of Project Files
- 3.7 GB of Digitized Aerial Photos

Data Preparation Workflow

North Carolina's spatial data clearinghouse is known as NC OneMap (www.nconemap.com). The NC OneMap portal provides freely accessible data created by State, Local and Federal agencies. This data is available for download in ESRI shapefile format. Raster data is available

[1] A full list of the demonstration datasets can be found in Appendix A

in MrSID, JPEG, and IMG formats. There are currently approximately 225 data sets available for download. Data hosted through NC OneMap partners are accessible through a clearinghouse of web map services (WMS), hosted on the NC OneMap website.

CGIA ingests raster and vector datasets into NC OneMap using the following workflow:

- A virus check is performed to ensure there is no corruption. If a virus is detected in a file, it will be quarantined by the program. The storage medium may be returned to the submitting agency/donor, and the agency will then be asked to resubmit once the problem has been resolved. Conversely, the corrupted file(s) can also be deleted after receiving the agency's approval. Information concerning a rejected transfer or deleted file will be recorded in a text file or other appropriate place.
- Data is loaded from transfer media to a local server.
- The dataset is opened and checks are performed to assess file validity, dataset projection, and geographic extent.
- The dataset's associated metadata record (if it exists) is opened to verify the record's completeness and validity. The metadata record is then run through both the U.S. Geological Survey's CNS metadata pre-parser and the MP metadata parser to validate that the record is valid and FGDC compliant.
- If a metadata record is not transferred with the dataset, CGIA staff creates a new metadata record with input from the data creator. CGIA staff will also enhance or refine existing metadata records transferred with datasets when they are missing critical information with input from data creator.
    - If changes are made to a metadata record, the original record is replaced by the updated copy and the updated copy becomes the metadata entry of record for the dataset.
    - If significant changes are made to a metadata record or CGIA has to create a new metadata record for a dataset, CGIA places its contact information in the Metadata Contact fields of the Metadata Reference section of the metadata file.
- Once the dataset and metadata record have been validated, the data is made available for public access via FTP and WMS

Dataset Transfer Preparation

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

- The datasets to be transferred will be moved and consolidated at transfer staging server located at CGIA.

- All archived vector files will be in shapefile format. Any geodatabases will need to be converted to shapefile format in the staging environment due to the adoption of shapefiles by the Archives as the archival format for vector data.

The staging area will have a folder structure similar to that of the repository at State Archives for large transfers. *For more information on this folder structure, please see the Data Receipt and Ingest at the State Archives section below.*

- Ad hoc transfers of a few files or less will be stored in an Adhoc folder and will contain a readme.txt to inform the Archives staff where the file should be stored in their repository.
- Datasets, images and documents will be opened to validate that they are functional and viewable.
- There will be a check to see if a valid metadata record is included and has an entry for dataset sizing.
- A hashing generator (md5deep or MD5Summer) will be run against the data in the staging environment, which will generate hashes for each file and will detail the folder structure.
- The North Carolina team will investigate using the BagIt specification to build a transfer bag to serve as an additional manifest.

Data Transfer

The North Carolina team is planning to test two methods for moving data between CGIA and the State Archives. For smaller vector packages, the team will attempt 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 will use portable hard drives to transfer files. All vector, project, and map data will be transferred in an uncompressed format and imagery will be transferred as both TIFF and MrSID. Information to be included in the transfer package will be:

- the dataset itself
- metadata record for the dataset (for vector and raster data)
- the hashing results file
- the BagIt manifest

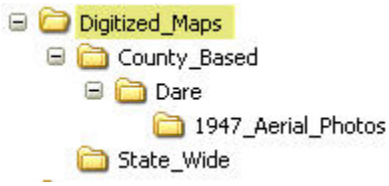
For hard drive/bulk deliveries, datasets will be stored and transferred within the folder structure hierarchy utilized by the State Archives repository. Ad hoc deliveries of data will only include the dataset, hashing file, and a readme text file to help aid filing the data within the repository and will not be contained in the ISO folder structure. Transfer rates for network transfers and loading/unloading rates for hard drive transfers will be captured during the data transfer process.

Data Receipt and Ingest at the State Archives

For hard drive transfers, once data has been received by the State Archives, the datasets will be transferred from the storage media to the Geoarchives SAN. Upon loading to the SAN, the data will be placed in a Staging folder for pre-ingest processing. CGIA will transfer data to the Staging folder directly via the network for Adhoc transfers.

The folder layout on the SAN will follow this model:

I. Root Directory
    a. Digitized Maps folder
        i. "County_Based" and "State_Wide" folders will be added to delineate between statewide and local datasets. These folders will contain subfolders for any digitized map products that will be archived as part of the demonstration, the actual pdf (or other format) product and any associated metadata record.



    b. Orthoimagery will be captured in a separate folder system. The master folder will be called Orthoimagery **(Figure 1)**.
        i. Under Orthoimagery there will be a subfolder for county, and then additional subfolders for each county or municipality that we have imagery for:
            a. Within the county folder will be folders for each year that the data exists.
                i. Within the year folder will be subfolders for compressed and uncompressed imagery.

Figure 1. Orthoimagery folder structure

c. Project Files Folder
    i. This folder will contain subfolders for any unique projects that will be archived.
    ii. The folders will have a descriptive name and date in their title (e.g. Sustainable_Sandhills_LandUse_2008).

d. Staging Folder
    i. This master folder will be the primary location for all incoming data from CGIA.
e. Vector Data
    i. Folders for each GIS ISO 19115 Category (Biota, Boundaries, etc., which encompass all 19 ISO categories) **(Figure 2)**.
        1. Each ISO Category folder will have subfolders for specific Ramona Data Layers.
            a. Within the data layers folder additional subfolders will be added to delineate between local and statewide datasets.
                i. In the case of local datasets there will be three subsets: Locality (e.g. Wake) → Dataset Name →Year (for each year that there is an available snapshot of that data). Datasets will be stored within the "Year" folder.

ii. In the case of statewide datasets, there will be two subsets: Dataset Name → Year (for each year there is an available snapshot of that data). Datasets will be stored within the "Year" folder.

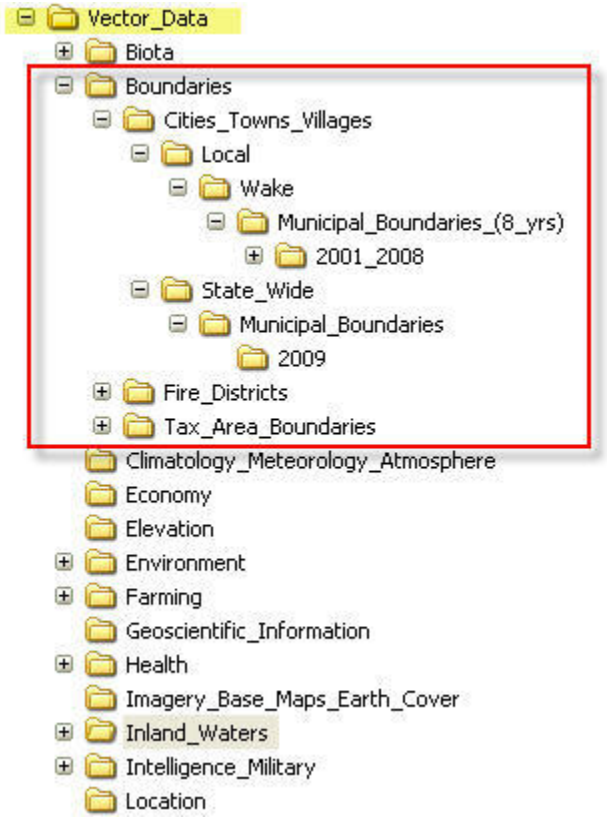


Figure 2. Vector folder structure

Once the data is loaded into the staging environment, the Archives staff will need to execute the following process:

- Perform a virus check to ensure there is no corruption in the transferred files.
- Run the BagIt validation tool to ensure that the files transferred match the manifest.
- Run the hash generator (md5deep or MD5summer) to measure current hashes and ensure that they match the pre transfer hashes.
- Rename files to match the naming convention standard: *Location_Datasetname_Create Year* (e.g. Wake_Parcel_2008).
- Open a random sampling of files to ensure that they are functioning and viewable.
- Run a metadata script to extract essential Federal Geographic Data Committee (FGDC) metadata elements from each dataset.

- Crosswalk extracted metadata elements into the North Carolina State Archives Manuscript and Archives Reference System (MARS) database.

An EAD (Encoded Archival Description) finding aid will be created at the collection level for the GIS datasets, projects and digitized maps and will include information about the collection such as acquisition and processing, provenance, organization and arrangement, and an inventory of the items in the collection. The finding aid facilitates the discovery of information on the World Wide Web, with the goal of guiding the user to the MARS database. It is within the MARS catalog that users can search the GIS collection using controlled vocabulary to find the specific datasets they seek. For the near future, the user will need to contact the State Archives to make arrangements to have access to the actual datasets.

Once the Archives staff successfully processes and validates the data in the staging environment, the files will then be migrated to their appropriate permanent location within the repository. For long term management of the data's integrity, the team will test the University of Maryland's ACE (Audit Control Environment) product. The purpose of ACE is to ensure the authenticity and integrity of digital objects in long term archives.

Demonstration Validation

In order to validate the success of the transfers, several validation steps are included in the data preparation transfer and ingest processes, including hashing/checksums, metadata validation and dataset opening and validation. Performance metrics will also be captured during these processes including data validation times, time to run hashing, and network transfer or data loading/unloading times.

The final validation check will attempt to measure the ease of data discovery. We will solicit several participants unfamiliar with the system, to try to discover, to access, and to open a specific geospatial dataset with little or no coaching. The participants will be asked to measure the time it takes to complete the process. We will then ask the participant to provide qualitative feedback by responding to a prescribed set of questions asking about the data retrieval process and challenges faced in data discovery and access.

The North Carolina project team will plan on moving a small subset of data to initially test these processes and will document the steps and results. Upon successful completion of this activity and after any adjustments to the workflow or process, the team will then attempt a bulk copy of the demonstration datasets and a few ad hoc transfers of data as well. This will also be documented, including the steps, results, experiences, and challenges. Future activities beyond the Intrastate demonstration transfer may include the copying of all superseded imagery from NC

OneMAP and based on outreach efforts, any additional vector datasets not included in the demonstration transfer.

Data Transfer Evaluation

Challenges Encountered with Data Preparation, Transfer and Ingest

The North Carolina team encountered multiple challenges when preparing, transferring and ingesting the demonstration datasets. Several of these were expected, such as the challenges associated with metadata, however many were unexpected. These challenges required the team to revise their way of thinking about the workflow which is outlined in the next section. Solutions to these challenges are addressed in both the Revised Workflow and Next Step sections.

1. *Virus Checking*. Virus checking was identified as an important step in the workflow, but this has not been fully implemented. Historically, CGIA has not performed virus checking and still needs to incorporate this step into their regular workflow. On the State Archives side, the virus scan software has not been installed on the GeoMAPP SAN. The virus scan was installed on the GeoMAPP Application Server, so datasets on an external drive can be scanned for viruses, however due to software limitations; the virus scanning software on the application server is unable to scan files on the GeoMAPP SAN.

2. *Data Migration*. It was decided that in lieu of downloading the datasets from the NC OneMap site CGIA would migrate the datasets from their original storage media to a local "staging" server for data preparation. As a temporary measure, a CGIA employee's work computer was appropriated as the CGIA local server. In preparation for this, 110 GB of space had to be cleared from the desktop for data preparation and transfer. This is not a feasible long-term solution for data preparation.

   Data migration included exporting shapefiles from the OneMap Production SDE database, uploading digitized maps from DVD's, and downloading orthoimagery from either an image server, copying them from CD's or providing an external drive to the data creator for them to copy the datasets from their storage media directly onto the external drive. This not only added more time and complexity to the data preparation workflow, but also made maintaining physical control of the three external drives more difficult. At any time the drives could be at CGIA, the State Archives or in the possession of a data creator.

3. *Dataset Validation*. It was initially decided that as part of the data preparation and ingest, dataset validation referred to the process of opening each dataset in ArcGIS to make sure

that it was functional and viewable within a GIS environment. As the data preparation and ingest of a variety of file formats and large datasets began, both CGIA and the State Archives quickly realized that they collectively needed to define specific steps required for data validation. Due to the large number of datasets being transferred it isn't always feasible to open every single dataset. Specifically orthoimagery is a challenge. A county's ortho flight can contain thousands of tile datasets.

4. *Metadata Challenges*. There were an assortment of complexities related to metadata, including:

    a. Incomplete and/or missing metadata on some of the datasets, primarily the local datasets and digitized maps.

    b. The minimum "mandatory" FGDC metadata elements necessary for data transfer from CGIA to the State Archives and for cross walking to the MARS catalog for data discovery.

    c. What "preservation" metadata needed be added to the datasets and which agency was responsible for adding this metadata. Once this was decided, we also needed to determine where these metadata elements were to be added in both the ArcCatalog metadata record and the MARS catalog.

    d. The "authoritative" subject/index terms critical for data discovery and which agency was responsible for adding the terms. Currently the subject/index terms in the datasets are user-created and therefore are not necessarily considered authoritative as they are not standard across data creators. These terms should be standardized in MARS for consistency in the data discovery process.

    e. Occasionally the metadata from provider was in .txt or .html format which is not recognized by ArcGIS 9.3.

    f. In ArcCatalog, the FGDC elements in "classic" view were not identical to the same FGDC elements in XML view. This caused confusion as to which version of the metadata element should be crosswalked to MARS **(Figure 3)**.

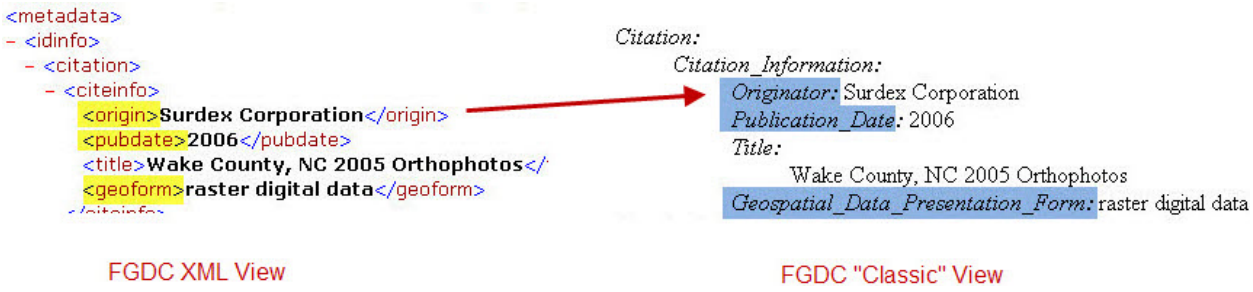


Figure 3. XML vs. "Classic" view

5. *Naming conventions*. According to the original data transfer design dataset renaming was supposed completed by the State Archives staff. It was determined that since CGIA has to export shapefiles from SDE feature classes it would be easier (and logically more feasible) for them to rename the files **(Figure 4)**. CGIA discovered that the datasets in the NC OneMap SDE environment were named with a universal system prefix (database.owner) and unique short abbreviation for each Feature Class (e.g. "sga" for "Shellfish Growing Areas"). When selecting multiple feature classes and exporting to multiple shapefiles, the output file could not be renamed. The resulting name contained a string of unnecessary prefix information such as onemap_prod.SDEADMIN.NAME.shp.

   To meet the naming convention requirements for data transfer, CGIA exported each feature class individually to a shapefile (batch exports don't easily allow for file renames) and during the process the file was renamed from its database identifier to a more intuitive name such as Shellfish_Growing_Areas_2004. While this worked effectively for vector datasets, there were additional questions about whether or not to rename both project files and orthoimagery since these types of files likely have intentional naming conventions (see #9 below for additional information on naming issues for orthoimagery).

   Issues with naming convention also emerged with "ad hoc" data transfers. All ad hoc folders were named "*YYYYMMDD_Adhoc*" with no reference to the type of datasets being transferred.

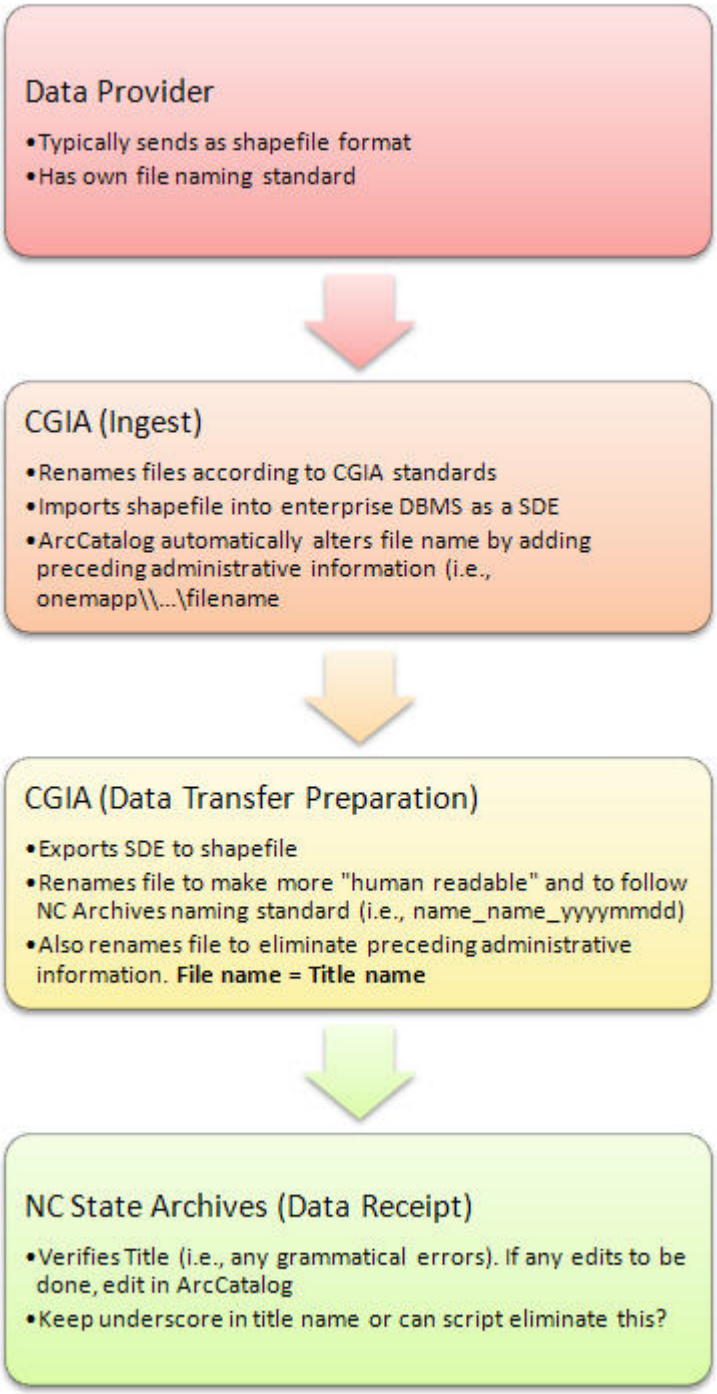


Figure 4. Data Flow.

6. *Folder structure.* We were able to follow the original folder structure, however during the transfer process we identified two additional delineations: (1) creating folders for orthoimagery to further delineate different scales (i.e., 100/400). See below in the Revised Workflow on when to add these folders, and (2) two root folders for the Preservation copy of the datasets and the Access copies of the datasets.

7. *Data transfer times/method of transfer.* Data transfer was perhaps the most challenging aspect of the workflow and multiple issues were identified:

    a. CGIA gets its information from a variety of storage media, which have to be loaded up to the CGIA local server. This can be time consuming particularly when loading up individual CD's.

    b. Vector datasets are transferred from CGIA to the GeoMAPP server "Staging" over a WAN. Since these datasets are small in both number and size it was assumed that the transfer would be relatively fast. However it took much longer to transfer due to a slower data transfer rate or throughput. For example, it took 10 minutes (or 600 seconds) to transfer 79.5 MB of data which is 17 times slower than expected.

    c. Orthoimagery presented its own challenges. Orthoimagery needs to be saved onto an external drive since it cannot be transferred via the WAN due to its size. The external drive needs to be directly connected to the GeoMAPP server (versus connecting it to the State Archives desktop) because it would take too long to transfer the data over the local network. As such, the external drive is connected directly into the GeoMAPP server. The datasets are then transferred from the external drive to the SAN "Staging" folder via the server. Several issues surfaced:

        i. Server network traffic is heavy during peak office hours, so if the State archives wants to transfer data during the day it is a VERY slow transfer.

        ii. We learned that there is actually a "host" machine connecting the GeoMAPP server to the SAN. The connection between the GeoMAPP server and the host machine is over the LAN (a regular network connection). Compounded with the server network traffic, it took an inordinate amount time to transfer large amounts of information (as an example it took approximately 15 hours to transfer 395 GB of data).This machine was running at 100 MB/s (which we weren't aware of). We recently experience connection issues and as a result of this issue being resolved, our bandwidth was increased to 2GB/s.

iii. The last major challenge was data validation using the BagIt tool. To validate the bag containing orthoimagery, BagIt creates a second manifest and compares it to the original manifest. This requires a large amount of RAM (random access memory). The amount of RAM is a critical factor because it will affect how much information can be cached. The more information that can be cached, the less a server has to rely on the comparatively slow process of opening and reading a file on disk. We discovered that the GeoMAPP server (since it is a "repurposed" server) only has 2 GB of RAM (newer machines have 4GB) which is not enough to validate the transferred orthoimagery.

8. *Orthoimagery*. Orthoimagery often has a complex organizational structure as there are a variety of ways to delineate the files. This includes Black/White vs. Color images, different scales (i.e., 100, 400, etc), and compressed (.tiff) vs. uncompressed (.sid). This structure is created by the data provider and it is not always clear how the data is organized. To complicate matters, orthoimagery also has a seemingly arcane file naming convention, consisting primarily of numbers ranging from 4 digits to 6 digits (e.g., *9860.tif* and *050514.tif*). While this numbering system may be both obvious and relevant to the data creator, it was oftentimes difficult for both CGIA and the State Archives to discern what this numbering scheme meant.

9. *Crosswalking*. Setting up the structure and creating metadata records in the MARS catalog was complex and time consuming.

    a. GIS records not only have a much more detailed and unique folder structure than most "paper" records **(Figure 5)**, but due to their digital nature are not easily translated to the more physical structure of paper records (i.e., series, box, folder, item). Determining how to mesh the two structures was complicated.

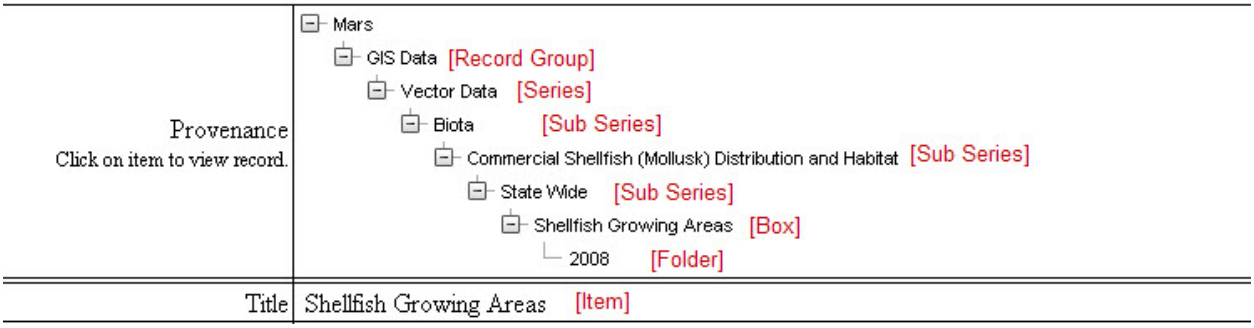


Figure 5. Folder structure

b. Specific FGDC metadata elements, including <geoform>, bounding coordinates (e.g., <westbc>, <eastbc>, etc.), <purpose>, and the preservation metadata (e.g., <cntorg>, <cntpos>, etc.) did not have a corresponding MARS metadata element.

c. There are two distinct metadata elements for title. <Title> is the more "human readable" title and <ftname> the file/table name. Which title element to crosswalk to MARS needed to be decided.

d. The date format in FGDC is not specified, resulting in a variety of creator-generated dates (including YYMMDD, YYYYMM, and YYYY with Month spelled out. The acceptable format in MARS is MM/YYYY.

Finally, no script to crosswalk FGDC metadata to the MARS catalog currently exists. It takes approximately 15 minutes to manually create one metadata record in MARS.

10. *Archival Finding Aid*. The collection level finding aid which will be available via the World Wide Web has not yet been written. Even when the finding aid is available, access issues still need to be addressed. As the datasets are not available online, there is a concern how the datasets will be delivered to the public if requested. Another outstanding question is how to serve out orthoimagery as it is so large both in size and the number of tiles. Copying the orthoimagery to CD's may not be a feasible solution.

11. *Data Integrity*. After a bag is validated, the datasets are moved to their appropriate folder structure resulting in the hashes no longer being associated with the relative path in the BagIt manifest. Essentially all the datasets will need to be re-hashed by another tool once they have been moved into the appropriate Preservation folder for long-term integrity.

12. *BagIt Bags*. Once a "bag" has been successfully transferred and the datasets moved to their appropriate location, there is still a question about how and where to store the remaining BagIt information (e.g., manifest). Currently, they are being organized chronologically under an "Unpacked Bags" folder, although there is still a question to whether or not the information is important to keep for the long term.

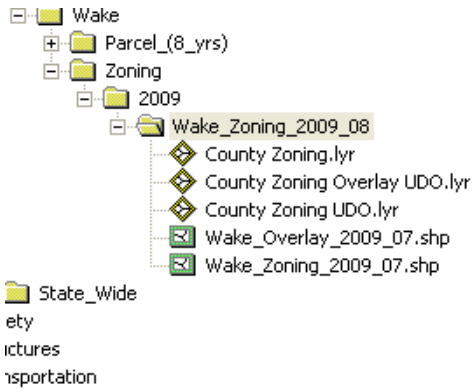
Revised Workflow

<u>CGIA Data Preparation and Transfer for Archiving</u>

1. Move datasets into staging environment on CGIA server (Vector datasets only)
    a. Prior to any transfers the "AdHoc" folder should be cleared out of any previous datasets before beginning a new transfer. To help aid State Archives in the storing and cataloging of geospatial datasets being archived, create folders for each of the

ISO categories that you will be transferring data for and place each dataset within its appropriate ISO Category folder. See: http://www.sco.wisc.edu/wisclinc/Layer_Assignments_060206.pdf to determine which ISO category the dataset belongs to.

b. For vector feature classes or projects received from other agencies or local governments, copy the dataset from the storage media to the staging server's "Data" folder as an initial staging location.
    i. Copy any datasets to be transferred to State Archives in the "AdHoc" transfer folder and place it into the appropriate ISO category folder. The Ad Hoc folder should follow the naming convention: *YYYYMMDD_Adhoc_Dataset Name or Type* (e.g., 20090826_Adhoc_2005_Wake_Orthos).
    ii. All vector datasets will need to be in shapefile format, so if the data received is in a geodatabase format or another convertible proprietary format, it will need to be converted to shapefile format.
        1. If additional files are associated with the shapefile (e.g., text file, layer file, etc.), they need to remain grouped with the shapefile(s). Text files may be data dictionaries or zoning codes and may be a plain .txt file or a Word document. Do not rename these files as it is unclear how/if this will affect the relationship between the files.



    iii. Rename the dataset to the archives standard naming convention: filename will follow this naming convention ***Location*** (where appropriate)***_Dataset name _Year_Month***. (e.g. Wake_Parcels_2005_09 or Shellfish_Growing_Areas_2008_08). Date should be taken from the metadata record (<pubdate>). **NOTE**: when a dataset is renamed using ArcCatalog, the feature class title will be changed in the metadata record to be the new shapefile's file name.

c. For datasets stored in the NC OneMAP SDE Environment:
    i. Export the feature class to a shapefile. As part of the export:
        1. Rename the shapefile to meet the archives naming convention listed above.
        2. Export the feature class to the "AdHoc" folder, moving it to the appropriate ISO category folder.
d. For very large datasets such as Orthoimagery (over 5 GB), copy the dataset to the portable hard drive that will be used for physical transfer to State Archives. **NOTE**: orthoimagery tiles should not be renamed.

2. Datasets, images, and documents will need to be opened to validate that they are functional and viewable before transferring.
    a. For GIS datasets:
        i. Open the dataset in ArcMap.
        ii. Validate that the dataset displays properly by opening the files in the Preview tab. Open the attribute table to make sure that it is populated.
        iii. Click on a feature using the Identify feature tool and make sure that attribute information displays.
        iv. Click on the properties for the dataset and make sure that a projection is defined for it.
    b. For Orthoimagery:
        i. Open 5-10 tiles in the Preview tab and validate that the dataset displays properly.
        ii. Validate that each of the tiles has an accompanying world file (.sdw or .tfw).
        iii. If there are both TIFF and MrSID copies of the same flight, compare world files between TIFF and MrSID to ensure that they are identical.
    c. Project files
        i. Open and review any available index or summary documentation.
        ii. Review folder/file structure/data organization to ensure that files are under the appropriate folders/subfolders.
        iii. Open and validate 5-10 thematic map exports (i.e., .pdf, .jpg)
        iv. Open 5-10 .mxd files using ArcGis to make sure that data is viewable
        v. Open 5-10 of any GIS datasets to make sure that data is viewable.
3. Metadata
    a. If metadata exists for a dataset to be archived, open the metadata record and check the following:
        i. Browse the metadata record for general completeness.
        ii. Critical fields for State Archives include: the citation information: (origin, pubdate, title, abstract, purpose) geoform, theme and place keywords,

Bounding coordinates, access and use constraints, and contact info for the dataset **(see Appendix B)**. It was concluded that the metadata that we're adding into MARS for data discovery would dictate the minimal metadata elements necessary for data transfer.

b. If metadata exists as a .txt or .html format:
    i. The .txt metadata needs to be imported into ArcCatalog and associated with the appropriate shapefile to make the association and create an XML version of the metadata.
    ii. .html metadata cannot be imported into ArcCatalog.
        1. One potential solution – use extern program to convert to xml (note: this was not tested as part of this demonstration)
c. If no metadata record exists for the dataset a minimal metadata record will need to be completed:
    i. Create a generic abstract and purpose that can be repurposed for a variety of datasets. For example:

    This starter metadata record was created by the NC Center for Geographic Information and Analysis to be used only for the purpose of ingesting this dataset into the archives. Please contact the data creator with any questions about the dataset and its attributes.

    ii. Add any available contact information for the data creator.
    iii. In the Citation details, add originator and online linkage if available.

4. BagIt
    a. For vector datasets, digitized maps, and project files, create a bag following the BagIt user manual.
    b. For orthoimagery create a bag by using the "baginplace" operation. This will bag the data files in the same location where the data files are stored, thus saving space.
5. Transferring the datasets
    a. For vector datasets, digitized maps, and project files, copy the bag from the CGIA server to the GeoMAPP SAN.
    b. For orthoimagery, physically transfer the external hard drive to a State Archives staff member.

<u>Data Receipt and Ingest at State Archives</u>

Vector Data, Digitized Maps and Project Files – transferred from CGIA over the WAN into the Staging folder on the GeoMAPP SAN

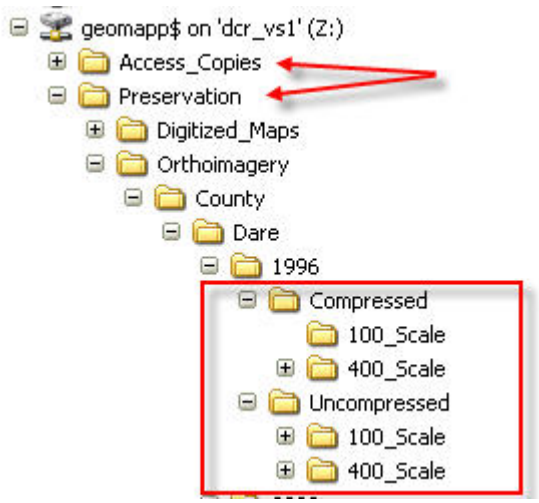
1. Virus scan should be done on all transferred datasets, but currently the virus scanning software has not been installed on SAN so this has not been done.
2. Run BagIt validation to ensure that the files transferred match the manifest.
    a. If the result of the validation is "False", identify the issue and contact CGIA to determine if the datasets need to be re-bagged and validated on their end. If yes, repeat Step 4 above. If no, include a "readme.txt" file with the other bag information specifying the identified problem(s) and the ending result.
3. Depending on the number of datasets included in the transfer, open all datasets or a random sampling of datasets in either ArcCatalog or ArcMap. Verify the following:
    a. For vector datasets, validate that the dataset displays properly by opening the shapefile (.shp) in the Preview tab.
    b. Open the attribute table to make sure that it's populated.
    c. Click on a feature using the Identify tool and make sure that the appropriate attribute information populates.
    d. Right-click on the dataset and select Properties. In the Shapefile Properties window, ensure that a projection is defined for it in the Details pane.
    e. For project files:
        i. Open and review any available index or summary documentation.
        ii. Review folder/file structure/data organization to ensure that files are under the appropriate folders/subfolders.
        iii. Open and validate 5-10 thematic map exports (i.e., .pdf, .jpg)
        iv. Open 5-10 .mxd files using ArcGis to make sure that data is viewable
        v. Open 5-10 of any GIS datasets to make sure that data is viewable.
4. Open the metadata record for each dataset transferred, and complete the following:
    a. Verify that CGIA has completed "mandatory" metadata elements that will be crosswalked to MARS **(Appendix B**).
    b. It was determined that the State Archives would be responsible for adding subject terms (or "themekeys" and "placekeys" in FGDC) that are standardized. Add bounding coordinates (if applicable), preservation metadata and supplementary note to each metadata record **(Appendix C)**. Additionally, include appropriate <themekt>, <themekey>, <placekt>, and <placekey> metadata elements (or search index terms) to each metadata record. Themekey should be ISO Category and Ramona Data Layer and the LOC Name Authority Headings (if necessary). Placekey should be "North Carolina" and the county name if applicable.
    c. If the shapefile is not projected and came from CGIA, the State Archives will need to contact CGIA about what projection to use.
5. Move the datasets to their designated locations under the *Preservation* and *Access* root folders.

6. In the Staging folder, delete the empty Data folder in transferred Bag, but retain the folder that holds the manifest and other BagIt text files. Move this to the "Unpacked Bags" folder.
7. In the MARS client, manually create a metadata record for each dataset.

Orthoimagery – transferred from external drive → GeoMAPP SAN via GeoMAPP server

1. Physically transfer the external drive to the Server room at the State Archives building and have technical staff connect it directly to the GeoMAPP server.
2. While logged into the GeoMAPP server, perform virus check on all orthoimagery on the external hard drive.
3. Copy entire bag from the external drive to the Staging folder on the GeoMAPP SAN using a batch process (Data_Transfer.bat). Using the batch process will allow for scheduling of the task at a time when the server network traffic is light (with the intent of shortening the time to copy). The batch script will also calculate the length of time it takes to copy all files by using timestamps.
4. Run BagIt validation to ensure that the files transferred match the manifest. Validation needs to be run from a server different from the GeoMAPP server. Remotely connect to the other server (DCR-sv3 which is mapped to the GeoMAPP SAN) to run the BagIt script.
    a. If the result of the validation is "False", identify the issue and contact CGIA to determine if the datasets need to be re-bagged and validated on their end. If yes, repeat Step 4 above. If no, include a "readme.txt" file with the other bag information specifying the identified problem(s) and the ending result.
5. Open either ArcCatalog or ArcMap and perform the following:
    a. Open 5-10 tiles for each orthoimagery dataset (i.e., Wake_2005) in the Preview tab to validate that they display properly. This includes 5-10 tiles for both.tiff and .sid files if applicable. If there are both black and white and color images, open 5-10 tiles for each of these as well. Additionally, if there is an index.shp file, preview this as well to make sure it is displayed properly.
6. Open metadata record for each dataset transferred, and complete the following:
    a. Verify that CGIA completed the "mandatory" metadata elements that will be cross-walked to the MARS catalog.
    b. In ArcCatalog, add appropriate preservation metadata and supplementary note to each metadata record. Additionally, include appropriate <themekt>, <themekey>, <placekt>, and <placekey> metadata elements to each metadata record.
7. **Prior to moving** the datasets from the *Staging* folder to their designated locations under the *Preservation* and *Access* root folders (i.e., moving compressed files into Access folder and uncompressed into the Preservation folder) perform the following:
    a. If there are scale differences (i.e., 100 vs. 400):

i. If the scales *are easily identifiable* (i.e., 100 scale images have a six-character naming convention and 400 scale have a four-character naming convention):
    1. If applicable, create distinct scale folders under both the Compressed and Uncompressed folders.



    2. Move the files to the appropriate scale folder.
    3. Include the correct metadata record in each scale folder. In general, there will be only one metadata record for each 100/400 scale for .tiff files only. These metadata records first need to be renamed so that which set of data they are describing is easily identifiable (i.e., *dare_1996_ortho_Metadata_100scale_SID.xml*). The metadata records then need to be copied, renamed, and updated for .sid files (if applicable).

ii. If the scales *are NOT easily identifiable,* do not attempt to separate out but instead keep in the original order. If two metadata records are included, include both records for compressed and uncompressed.

b. If there is an Index .shp file:
    i. Move the Index .shp and all its auxiliary files to both the root "Uncompressed" and "Compressed" folders.

c. If there are additional Index files (i.e., .dbf, .prj, etc.), move the entire Index folder to the appropriate "Uncompressed" and "Compressed" root folders.

8. In the Staging folder, delete the empty Data folder in transferred bag, but retain the folder that holds the manifest and other BagIt text files. Move this to the "Unpacked Bags" folder.
9. Delete all orthoimagery data from the external drive.
10. Retrieve the external drive from the Server room at the State Archives.

11. In the MARS client, manually create a metadata record for each dataset.

Data Transfer Metrics

Data Preparation and Transfer Time

For local government and centralized (framework/non-framework) datasets, project files and digitized maps, data preparation and transfer time includes exporting datasets from a variety of sources including the OneMap Production SDE database and/or DVD/CD's to a CGIA Staging folder, renaming the datasets, validating the integrity of both the data and metadata, adding or editing metadata if necessary, creating a BagIt bag, and transferring the bag to the Staging folder on the GeoMAPP SAN.

Data preparation and transfer time for orthoimagery is the same as above, except that includes the initial step of transferring the data from either an Image server, CD's or the data creator's storage media onto a portable hard drive.

Processing Time

For local government and centralized (framework/non-framework) datasets, project files and digitized maps, processing time includes validating the BagIt bag, reviewing and validating the integrity of the data, moving the data from the Staging folder to the Preservation folder on the GeoMAPP SAN, adding preservation metadata to the dataset, and creating a metadata record in the MARS catalog.

Processing time for orthoimagery is the same as above, except that includes the initial step of transferring the dataset from the external hard drive to the GeoMAPP SAN via the GeoMAPP server.

Data Type	**Total Number of Datasets/File Size Transferred (in MB)**	**Data Preparation Time**	**Transfer Time**	**Processing Time**
Local Government Datasets	5/79.45 MB	2 hours	10 minutes	1.5 hours

Orthoimagery	5/747,701 MB	48 hours	36 hours	38.5 hours
Centralized Datasets	15/1,728.2 MB	3 hours	3.5 hours	5 hours
Project Files	1/3,072 MB	35 minutes	2 hr 20 minutes	1 hour
Digitized Maps (TIFF/sid)	2/4,295 MB	1 hour	20 minutes	2 hours
Totals	28/756,929.7 MB	54 hours 35 min	42 hrs 20 min	48 hours

Next Steps with Data Transfer and Ingest

The following are items that the North Carolina team continues to improve upon or is investigating with regards to our transfer and ingest processes:

- CGIA is planning to incorporate virus checking in the data preparation workflow. The State Archives is planning to install virus checking software on GeoMAPP SAN for future dataset transfers.
- Investigate avenues for improving the data transfer rate between CGIA and the Staging folder on the SAN.
- Create Python script to crosswalk FGDC metadata to the MARS catalog.
- Create a collection level finding aid and devise a plan for making all types of datasets available to the public.
- Provide to the State Archives reference staff a basic primer on geospatial data to aid in access and discovery of data, and to better assist users of the archives. (Completed)
- Install specialized network cards (a pair of single port or a single dual port 4 GB/s fiber channel host bus adapter cards and the necessary cables) that will allow for direct access to the GeoMAPP SAN from the GeoMAPP application server.
- Continue investigating the ACE audit manager for long-term data integrity.
- Transfer additional ad hoc datasets.
    - The remaining NC OneMapp vector datasets not included in this demonstration.
    - The remaining superseded orthoimagery collection not included in this demonstration.
- The long-term vision of the NC geoarchive system is to archive critical local government and state agency datasets. The team has written draft records retention schedules for local government geospatial datasets as well as schedules for other state agency GIS data

creators. If these schedules are approved, CGIA would act as a conduit for these records to be transferred and permanently preserved by the State Archives. Interim steps for this implementation would be additional test transfers from local government and state agencies to validate these procedures.

- Address zipping the datasets prior to transfer to determine if this would make a difference in the transfer time.
- Complete a full system data transfer (i.e., one that is not "Ad hoc").
- Work with the data providers of orthoimagery to decipher naming conventions and potentially formulate a plan for more standardized naming conventions.

Appendix A: North Carolina Demonstration Datasets

North Carolina's Test Datasets:

The North Carolina team decided to focus on a coastal processes theme for many of the datasets that it will be including in the data migration demonstration portion of the project. This coastal theme was driven by the fact that North Carolina is currently the only ocean bordering state in the project and by the fact that they host a number of coastal datasets in their centralized GIS clearinghouse, NC OneMap.

Local Government Datasets:

North Carolina will initially be focusing on two counties data for inclusion for the demonstration:

- Wake, which is the NC's second most populous county and includes state capital Raleigh, provides an urban perspective and also has superseded data in their holdings
- Dare County, which includes the coastal Outer Banks, provides a combination of highly developed coastal property and large undeveloped areas reserved for preservation

Datasets targeted for collection from the counties and for inclusion in the demonstration archive include "at risk" regularly updated datasets such as parcels, zoning, boundaries, and street centerlines. Both counties have several snapshots of countywide Orthoimagery, which also made them appealing for integrating into the demonstration.

NC is currently investigating potentially reaching out to two other counties to get data with a focus on rural border counties (to have a rural perspective and to have data that may be worth sharing with a neighboring state) and counties from the western part of the state (to capture an Appalachian/ mountain county) in hopes of discovering datasets that are unique to these counties population's needs that would also be good to archive.

Centralized Datasets:

North Carolina's centralized datasets will be drawn from the collections of NC OneMap, which is NC's statewide GIS data clearinghouse. The NC team selected framework datasets that were potentially common to the three state partners; datasets that were both important to archive from an NC perspective and would also be good to use to compare to similar datasets created and maintained by the other state partners. The non framework datasets were selected due to their coastal, hydrologic, or disaster recovery themes and all of the datasets include data points and coverage for the coastal parts of the state.

Project Files:

North Carolina selected the **Sustainable Sandhills** project to be its primary project file for archiving. The Sustainable Sandhills project looked at land use in a several county region in south central North Carolina that could be potentially impacted by the expansion of Fort Bragg, a large US Army base in Fayetteville. Ft. Bragg will be expanding significantly to the BRAC base realignment effort and this project was an initial look at existing land cover types in the regions and made recommendations for future use of the land for various purposes (i.e. Agriculture, Industry, Residential, etc) based of GIS modeling and analysis. This project was selected due to its value for preservation, the multiple, inputs and processes that took

place for creating output and the fact that the project files had been packaged, organized and documented for delivery to project customers and planners.

Scanned/Georeferenced/Digitized Maps:

The North Carolina team decided to keep with its coastal theme for the scanned maps selection and will be archiving digital coastal maps, including the 1947 TIFF images from Dare County.

GeoMAPP North Carolina Demonstration Datasets				
Source	**Dataset Name**	**Creation Date**	**Compressed Size (MB)**	**Uncompressed Size (MB)**
Local Government				
Dare County	Tax Districts	2005		1
Dare County	Parcel	2008		67
Dare County	Flood Zones	2005		4
Dare County	Fire Districts	2005		5.5
Dare County	Streets	2008		2.6
Wake County	Parcel (8yrs)	(2001-2008)	944	2280
Wake County	Zoning	2009	0.7	1.2
Wake County	Streets (8yrs)	(2001-2008)	59	212
Wake County	Municipal Boundaries (8 yrs)	(2001-2008)	10.5	15
Orthoimagery				
Dare County	Orthoimagery (1996)	1996	192	3768
Dare County	Orthoimagery (2002)	2002	9216	N/A
Dare County	Orthoimagery (2007)	2007	7055	201728
Wake County	Orthoimagery (1999)	1999	3400	118784

Wake County	Orthoimagery (2005)	2005	14438	389120
Centralized Datasets				
Framework				
DOT[2]	Municipal Boundaries	2009	5.8	7.6
CGIA/ DENR- Water Qualty	Hydrology 1:24k	2006	276.5	1334
CGIA	Federal Land Ownership	2006	4.3	6.1
CGIA	County Boundaries Shoreline	2006	N/A	7.8
DOT[2]	Statewide Roads (Arc)	2009	N/A	224
CGIA/ State Property	State Owned Lands	2008	7.6	11.6
Non Framework				
CGIA/ DENR- Marine Fisheries	Shellfish Growing Areas	2008	36.2	51
CGIA/ DENR- Waste Mgmt	Haz. Substance Disposal Sites (Superfund)	1998	0.3	0.8
CGIA/ DENR- Water Qualty	Ntl Pollutant Discharge Elimination System Sites	2006	0.1	0.8
CGIA/ Ntl Hurricane Cntr	Hurricane Storm Surge Areas- Fast Moving Storms	1999	31.9	67.7
CGIA/ Ntl Hurricane Cntr	Hurricane Storm Surge Areas- Slow Moving Storms	1999	29.6	66.2
CGIA	Emergency Operations Centers	2007	0.1	0.2
CGIA	Potential Emergency Shelters	2003	0.2	0.7
CGIA	Hurricane Evacuation Routes	2007	0.9	1.4
CGIA/ DENR Parks & Rec	Paddle Trails- Coastal Plain	2001	0.9	3.4

[2] These datasets were acquired by NCDOT.

Project Files				
CGIA	Sustainable Sandhills		687.5	3072
Digitized Maps				
Dare County	1947 Aerial Photos	1947	199	3983
		Total Sizing	**36,596.1**	**724,826.6**

Appendix B: FGDC Metadata Crosswalk

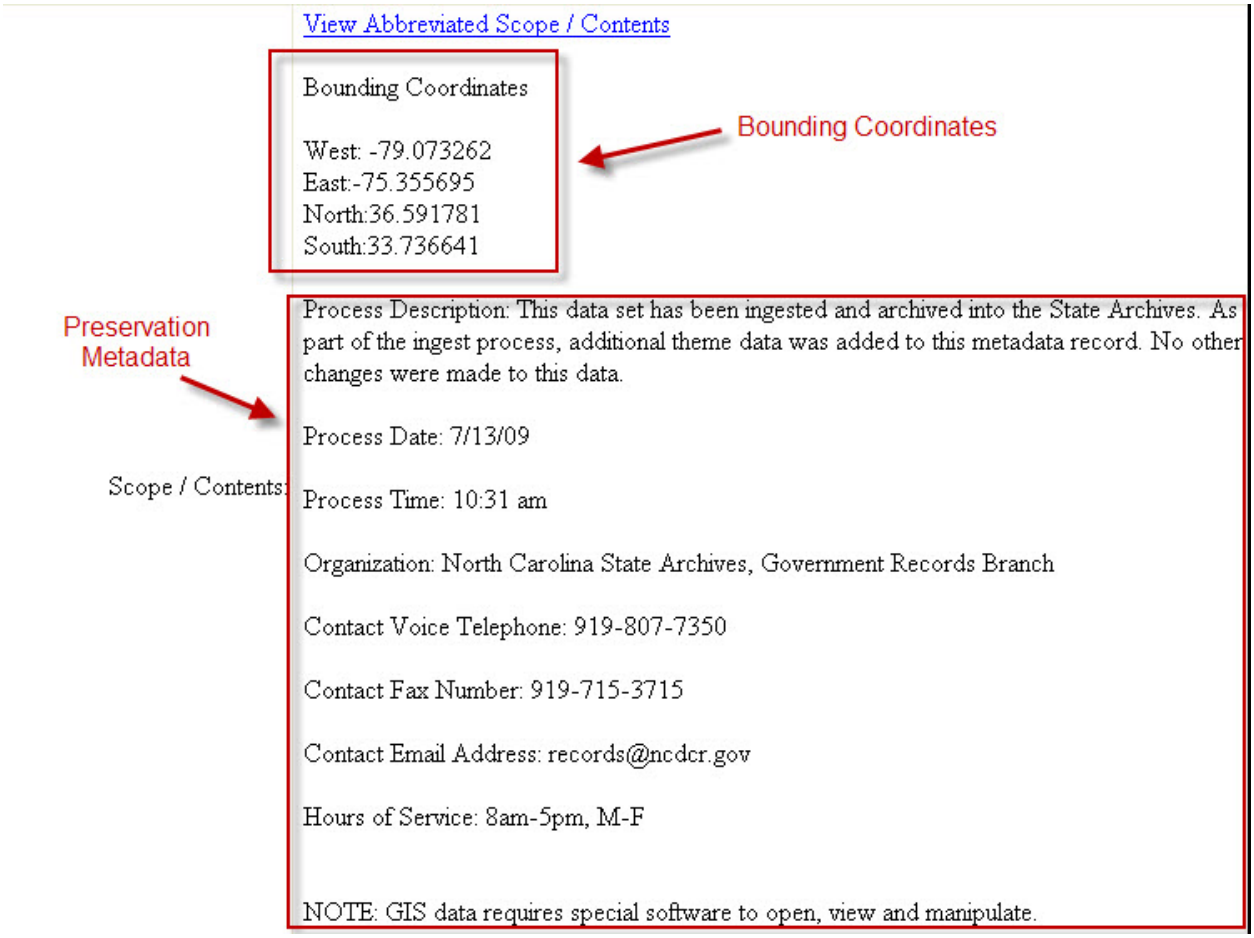
FGDC XML Elements	FGDC Element Definition	FGDC Current Format	Mars Fields	MARS Format
citation: citeinfo: origin	The name of an organization or individual that developed the data set.	No standard	Originator	For government agencies, appear to use LOC Authorities headings when applicable.
citation: citeinfo: pubdate	The date when the data set is published or otherwise made available for release.	Recommended that the YYYYMMDD format is used, however, have seen following format: YYYY Month (spelled out - i.e., March)	Years	YYYY
citation: citeinfo: title	Name of the dataset (without the extension)	Currently no naming standard	Title	Not sure if a standard naming convention is followed.
citation: citeinfo: geoform	Geospatial data representation form.	Drop down list in ArcCatalog with various forms to choose from.	Form of Record	
descript: abstract	A brief narrative summary of the data set. *NOTE: for orthoimagery we should indicate how many files/tiles associated with the dataset.*	Free text	Abstract	Free text
descript: purpose	A summary of the intentions with which the data set was developed.	Free text	Abstract	Free text

spdom: bounding: westbc	Western-most coordinate of the limit of coverage expressed in longitude.	Numeric with decimal point.	Scope	
spdom: bounding: eastbc	Eastern-most coordinate of the limit of coverage expressed in longitude.	Numeric with decimal point.	Scope	
spdom: bounding: northbc	Northern-most coordinate of the limit of coverage expressed in latitude.	Numeric with decimal point.	Scope	
spdom: bounding: southbc	Southern-most coordinate of the limit of coverage expressed in latitude.	Numeric with decimal point.	Scope	
keywords: theme: themekt and themekey	Themekt makes reference to a formally registered thesaurus or a similar authoritative source of theme keywords. Themekey is the common-use word or phrase to describe the subject of the data set.	Can either can be "authoritative" word (meaning a word from an authoritative source) or user-created word (can be various parts of speech including noun, verb, adjective, etc. - this is not dictated).	Subject (Subject)	From authoritative source? Have the choice between HICATS Subject Headings and State Archives Subject Headings.
keywords: place: placekt and placekey	Common-use word or phrase used to describe the regional reference of the data set.	Can either can be "authoritative" word (meaning a word from an authoritative source) or user-created.	Subject (Geographic)	
accconst	Restrictions and legal prerequisites for accessing the data set.	Free text	Access Restrictions	

useconst	Restrictions and legal prerequisites for using the data set after access is granted	Free text	Use Restrictions	
Preservation Metadata				
dataqual: lineage: procstep: proccont: cntinfo: cntorgp: cntorg	Data Quality information essentially tracks the lineage or history of the dataset. This may describe processes such as importing metadata, whether or not the dataset was copied and ingesting and archiving into the State Archives		Scope	
dataqual: lineage: procstep: proccont: cntinfo: cntpos			Scope	
dataqual: lineage: procstep: proccont: cntinfo: cntvoice			Scope	
dataqual: lineage: procstep: proccont: cntinfo: cntfax			Scope	
dataqual: lineage: procstep: proccont: cntinfo: cntemail			Scope	
dataqual: lineage: procstep: proccont: cntinfo: hours			Scope	
dataqual: lineage: procstep: procdesc			Scope	
dataqual: lineage: procstep: procdate			Scope	
dataqual: lineage: procstep: proctime			Scope	

Appendix C: Additional State Archives Metadata

Bounding Coordinate/Preservation Metadata



Supplementary Note Field

Note:	Currently GIS datasets are NOT available online. If you need to access any of the GIS datasets, please make a request via the Government Records Branch of the State Archives. Please note that the Government Records Branch is not responsible for the technical accuracy of the datasets. If upon receipt of the GIS datasets there are technical questions, please consult the metadata record for contact information on the data producer.