Geospatial Multistate Archive and Preservation Partnership (GeoMAPP)

Best Practices for Geospatial Data Transfer for Digital Preservation

December 1, 2011
version: 1.0 (final)
# Table of Contents

1. Introduction ..........................................................................................................................................................1

2. Geospatial Data Transfer Planning and Preparations ......................................................................................3
   2.1. Getting Started with the Project – The Kick-off Meeting .................................................................3
   2.2. Assessments, Inventories, and Surveys .................................................................................................5
   2.3. Appraisal ..................................................................................................................................................6
   2.4. Geospatial Data Transfer Planning ......................................................................................................6
       2.4.1 Engage the Information Technology Organization ................................................................8
       2.4.2 Digital Object Considerations for Data Transfer and Preservation - Geospatial Metadata Requirements ..............................................................8
       2.4.3 Digital Object Considerations for Data Transfer and Preservation - Naming Conventions ........9
       2.4.4 Digital Object Considerations for Data Transfer and Preservation - Data Organization ..........9
       2.4.5 Dataset Verification .......................................................................................................................15
       2.4.6 Dataset Packaging Mechanism for Transfer and Post-transfer Validation ................................17
       2.4.7 File Transfer Size Requirements ..................................................................................................20
       2.4.8 File Transfer Mechanism .............................................................................................................20
       2.4.9 Identify Archival Submission Policies ..........................................................................................22
       2.4.10 Create a Written Document for Data Transfer to the Archival Organization ..........................23

3. Role-based Tasks for Geospatial Data Transfer ..............................................................................................24
   3.1 Geospatial Data Creators Tasks that can Facilitate Data Transfer and Preservation ..........................24
   3.2. GIS Clearinghouse Geospatial Data Transfer Tasks (or Submitter of geospatial data to theArchival Organization) .................................................................................................................................24
   3.3. Archival Organization Geospatial Data Receipt Tasks ........................................................................27

4. Conclusions .........................................................................................................................................................32

Appendix A – Geospatial Archiving Roadmap ................................................................................................. A-1
Appendix B: Process Checklist ........................................................................................................................... A-2
Appendix C - Resources ........................................................................................................................................ A-4
Appendix D - Archives 101: GIS Questions about Archiving ........................................................................ A-8
Appendix E - GIS 101: Archivists Questions about GIS ................................................................................ A-10
1. Introduction

The GeoMAPP project’s primary goal is to ensure the long term preservation of archived geospatial content.\(^1\) In order to preserve geospatial content, the GeoMAPP team explored processes, techniques and technologies to transfer geospatial data to an archival organization with state-wide preservation responsibility for geospatial data.

The following diagram illustrates a generalized version of the life cycle of a geospatial dataset from its creation to its final destination in an archival organization for long term preservation. Many states have established/are establishing a centralized GIS clearinghouse to provide centralized management and access to current geospatial datasets. Because of this emphasis on current data, when new versions of existing datasets become available, a clearinghouse may discard their superseded datasets. The state archives, or another institution with the mandate and mission to preserve state historical records, provides a logical location to house, preserve and provide access to these historical geospatial datasets.\(^2\) The archival organization may receive datasets from the clearinghouse, as illustrated below, or directly from state or local government data producers. In either case, the dataset submitter will be involved in similar planning, data transfer preparation, and data transfer tasks, as highlighted by the red box in the following diagram.

![Diagram of geospatial data life cycle](image)

Figure 1. Life cycle of geospatial data

The general organizational structure for the GeoMAPP partner states Kentucky, North Carolina, and Utah was a statewide GIS Clearinghouse contributing data to each state’s State Archives. Montana was structured differently, with the GIS Clearinghouse and the geospatial archival responsibilities co-located in the Montana State Library (MSL). Unlike Kentucky, North Carolina and Utah who will route their datasets through the clearinghouse, MSL plans to immediately archive datasets that they accept into their collection, and then produce the access derivatives, which will be deployed through the MSL clearinghouse. Another significant difference between the general architecture for the states, that is largely due to Montana’s centralized GIS and archival responsibilities, is

---

2 In some states, the State Library or State Historical Association may hold preservation responsibilities.
that MSL is planning to host a single GIS repository that will serve as both the GIS Clearinghouse (serving current datasets), as well as the archival preservation and access repository (serving superseded archived datasets).

Kentucky, Montana, North Carolina, and Utah, pursued different approaches to archiving their geospatial data, which in turn, influenced their specific data transfer processes, as well as methods they used to organize their data. Montana, North Carolina and Utah pursued a dataset-oriented archiving approach, while Kentucky pursued a geodatabase snapshot approach. These differing approaches yielded diverse best practices that should benefit other organizations considering geospatial archiving solutions. Details for each individual state’s test data transfer process can be found in each state’s data transfer white paper,3 and the GeoMAPP Interim Report: 2007-2009,4 and the GeoMAPP Final Report: 2010-20115.

This paper focuses on the data transfer aspect of the geospatial data life cycle, as indicated by the red box in the above diagram. It assembles several data transfer best practices for GIS professionals and archivists, as they partner and collaborate to transfer geospatial data to the archival organization for long-term preservation.

This paper offers:
- general practices that can apply to all participants in the geospatial data development life cycle as they collaboratively plan their geospatial data transfer tasks, tools, and infrastructure,
- recommendations for GIS professionals as they prepare, package and transfer their datasets to the archival organization, and
- techniques and tools for data preservationists and archivists as they receive and process the datasets.

Looking at a high level geoarchiving project roadmap, as illustrated in Figure 2, a geoarchiving project can be organized into several phases to manage the development of the cross-organizational geospatial data transfer and geospatial data archiving. Concurrent to the data transfer planning will be informational technology infrastructure planning and execution activities to prepare the infrastructure of both the contributing organization (e.g., the GIS Clearinghouse), and the receiving organization (e.g., a state archives). These are important activities to consider as there may be time-consuming dependencies such as procuring new storage, or establishing network infrastructures to enable the electronic transfer of data across organizational boundaries.

Part 2 of this paper focuses on the planning activities highlighted by the red boxes in Figure 2, and provides a collection of best practices realized by the GeoMAPP partner teams as they conducted their data transfer activities including:

3 State specific data transfer reports for the GeoMAPP partners can be found at: [http://www.geomapp.net/publications_categories.html#xfr](http://www.geomapp.net/publications_categories.html#xfr)

v1.0 12/1/2011
• Scheduling a cross-organizational kick-off meeting that promotes and initiates collaboration and joint planning between your GIS and archival professionals. Cross-organizational collaboration is a key geospatial data transfer prerequisite, and success factor, as datasets will likely cross organizational boundaries.

• Conducting assessments and inventories that can inform the participants about the types, formats, and sizes of the geospatial data that are being considered for transfer to the archival organization for long-term preservation.

• Evaluating and identifying the datasets that will be transferred to the archival organization.

• Planning and specifying the details of your data transfer, including data formats, dataset packaging mechanism(s) for transfer, transfer medium, dataset validation techniques and tools, and methods for functionally verifying datasets received by the archival organization.  

Following the planning activities, the organizations will proceed with the execution of their data transfer activities:

• The GIS clearinghouse (or state/local government submitter) prepares and transfers datasets, and

• The archival organization receives, validates, and functionally verifies the datasets.

Part 3 of this paper transitions from planning to execution, and offers role-based task summaries of the data transfer process for the GIS dataset developer, the geospatial dataset submitter (often a GIS clearinghouse), and the archival processor receiving the datasets. Part 3 may be useful in establishing the specific roles, responsibilities and tasks within each organization to support the tactical execution of the geospatial data transfer.

Following the receipt of the datasets, the archival organization proceeds with its archival processing activities to ingest, describe, and arrange the geospatial datasets into the archival repository. For more information on the archival processing activities, refer to the GeoMAPP whitepaper, Best Practices for Archival Processing for Geospatial Datasets. 

Note: There is an annotated version of Figure 2 found in Appendix A, that lists the subtasks associated with each of the geoarchiving project phases, as well as includes the alignment of software procurement and/or development activities that may be necessary to realize the full data transfer solution.

The Appendices offer several tools, including data transfer process checklists and resource lists that may be helpful in the planning and/or execution of your data transfer activities.

2. Geospatial Data Transfer Planning and Preparations

A geoarchiving project is a complex multi-phase initiative that requires the joint participation of GIS professionals, archival professionals, and their supporting information technology partners. Thoughtful, collaborative planning that includes all participants will promote the effectiveness, efficiency, and success of the project implementation. This section offers general guidance on the cross-organizational planning and preparation activities for a geoarchiving project, as well as examples taken from the experience of the GeoMAPP partner states. Appendix B provides a checklist summarizing these planning steps, and Appendix C provides links to resources supporting many of the planning activities.

2.1. Getting Started with the Project – The Kick-off Meeting

Schedule an inter-organizational kick-off meeting to meet your counterparts in the geospatial and archives organizations.

As the archiving of geospatial datasets is an inter-organizational collaborative effort between GIS and archives professionals, and the dataset transfers often cross the boundary between GIS and archival organizations, it is important to establish a relationship between these two teams early in the project. Developing successful and functional relationships is essential for the long-term success of the project.

6 While the GeoMAPP team performed limited data transfers of maps and projects, their work focused predominately on vector datasets with some minor explorations in transferring and verifying orthophotos (a raster format). Therefore, the guidance provided in this document focuses mostly on the handling of vector datasets, and to a lesser extent on orthophotos.

effective dataset transfer mechanisms requires cooperation and coordination among these partners from the earliest planning phases through to the actual execution of the dataset transfers. A kick-off project meeting can be very useful in establishing an interpersonal rapport and a baseline understanding of the extent and nature of the geospatial assets to be transferred to the archives.

This meeting will likely be organized by the sponsors or champions in both the GIS and the archival organizations who will lead the strategic initiative to preserve the state’s key geospatial assets. Use this meeting to get to know your counterparts in the archival and GIS organizations.

The GeoMAPP team highly recommends presenting cross-training for both types of participants to gain a better understanding about the details of their counterpart’s world. At a minimum, each partner needs to learn the basic terminology and concepts of the other’s environment. GIS personnel will benefit from understanding some of the basics of archiving and digital preservation (see Appendix D: Archives 101 questions), to streamline their work with archivists while preparing datasets for long-term preservation. Archivists require an introduction to the geospatial data formats and the tools used to create, render, and access geospatial datasets (see Appendix E: GIS 101 questions) so that they can:

- Understand how the geospatial information works,
- Determine the best mechanism for inspecting, preserving, organizing, and describing the datasets,
- Identify which tools and technologies archivists will need to implement within their systems for ongoing management of the datasets, and
- Identify and develop tools to enable public access to the datasets.

A key element of the orientation for the archivists will be an introduction to geospatial metadata. The GIS community is actively engaged in creating and promoting metadata standards and the use of metadata to describe geospatial data. Archivists know that metadata is a critical component in successful archiving initiatives, and will be interested in the breadth and richness of the geospatial metadata. The GIS personnel should provide an orientation for the archivists on the metadata standard used by the submitter (often the statewide clearinghouse, and hereafter referred to as the clearinghouse⁸), presumably based on the Content Standard for Digital Geospatial Metadata (CSDGM)⁹, established by the Federal Geographic Data Committee (FGDC) and endorsed by many GIS practitioners. Collectively, the GIS professionals and the archivists will need to understand, or even define, the guidelines and/or policies within the state regarding the FGDC metadata requirements for the geospatial datasets that will be preserved by an archival organization.

GIS professionals will benefit from understanding the state’s public records law as it pertains to GIS data, and should be introduced to any pertinent records retention schedules that apply to GIS organizations or GIS data. The archival staff should provide them with an introduction to appraisal to that they may aid in the dataset evaluation process, and an overview of archival processing and how the geospatial metadata will be applied to the archival metadata record, thus reinforcing the importance of providing a full geospatial metadata record.

Technical and computing infrastructure is a critical aspect of the overall project. You should, therefore, invite your information technology partners to participate in your kick-off meeting. This gives them an opportunity to meet the both the GIS and archival professionals they will be supporting.

The kick-off meeting should review the high level project activities and develop a preliminary project plan. At the conclusion of this meeting, liaisons for both the clearinghouse and the archival organization should be identified, as well as key leadership roles for the project.

If archiving geospatial data seems daunting, you may choose to begin your efforts with a smaller scale pilot or demonstration project that consists of a few representative datasets to help determine the resources needed and refine the processes required to transfer and archive geospatial data.

---

⁸ It is understood that there will be geospatial submitters other than the GIS clearinghouse, but the most common model the GeoMAPP partners encountered was transferring datasets to the archival organization from the statewide GIS Clearinghouse, so “Clearinghouse” will be used hereafter to generally refer to any geospatial submitter, including individual state/local organizations that may directly submit datasets. The GIS clearinghouse has the benefit of being a centralized GIS management organization within the state, and thus will serve as a key strategic partner for the archival organization as they conduct their data transfer planning and design.


v1.0 12/1/2011
2.2. Assessments, Inventories, and Surveys

Conduct an initial archival capabilities assessment, geospatial inventory and technical infrastructure survey to identify the organizations (state, local), geospatial assets, systems, software, and personnel involved in geospatial data creation, and in archiving of electronic records.

As there are a variety of geospatial software packages, and geospatial data can be stored in a variety of formats, both inventories and surveys will inform the team what geospatial formats may be candidates for transfer to the archives. They will, therefore, be useful tools to accompany your data transfer planning. In addition, they can inform the archives on what software to purchase to support the functional verification of transferred datasets, as well as the ongoing preservation of the geospatial data. Refer to Appendix C: 2.2 for sample resources and templates for preparing and conducting assessments, inventories, and surveys.

Archival Capabilities Assessment

Archival organizations have varying degrees of capabilities, policies and procedures, and technology and infrastructure, and geospatial datasets present a complex digital format to preserve. Therefore, the GeoMAPP team advises that archival organizations conduct an archival self-assessment to evaluate their capabilities and readiness to archive digital materials, in general, and geospatial datasets in particular.

Baseline Inventory of Geospatial Datasets

As an important first step, you should conduct a baseline inventory of the geospatial assets that are in scope for your data transfer and preservation initiative. This inventory can provide insight to the current coverage, variety, and sizes of geospatial assets as well as provide a baseline for future planning discussions. Important data to collect in the inventory include the data creator, creation date, abstract or brief description, the geospatial format (e.g., shapefiles, uncompressed orthophotos, compressed orthophotos, geodatabases, projects, etc.), dataset size, and update frequency. Technical information that will be useful to archival planning should also be gathered such as the software application used to create geospatial datasets and the number and sizes of the datasets – information critical for repository storage planning. As you start your inventory project, consider investigating prior GIS inventories that you can leverage, such as your state’s prior contributions to the National States Geographic Information Council (NSGIC) GIS Inventory (http://www.gisinventory.net/). Additionally, data stewards and creators may also have created and maintained locally defined inventories that you may be able to reuse. As you conduct your geospatial inventory, if not already a contributor, you may also want to consider participating in and registering datasets with the NSGIC GIS Inventory.

For a pilot data transfer project, you may want to select a smaller domain of datasets to inventory and assess for transfer to the archival organization, and continue with the larger scale inventory activities in parallel with developing your initial data transfer prototype. For Montana, developing an inventory of their geospatial assets was a key objective, and they used this “pilot project” strategy to initiate a small geoarchiving project as they continued to conduct their larger inventory efforts.

Technical Infrastructure Survey

In conjunction with inventorying geospatial assets, you should conduct a survey of the technical computing environment and software tools used by the GIS data producers, the GIS clearinghouse, and the archival organization. This is important for evaluating your current hardware, software, and network capabilities. It also allows you to evaluate your current storage and network capacities and assess the need to expand your storage resources to support your geoarchiving initiative, either immediately or over time. It may also inform organizations about additional software may need to be purchased. For example, the archival organization, at a minimum, will likely need to acquire a GIS software package to verify received geospatial datasets.

---

2.3. Appraisal

Appraisal is the process of identifying which materials have sufficient value to be accessioned into the archives.\(^{12}\) As each state will likely be faced with a large number of geospatial assets being produced, it will be important to determine the criteria for identifying which datasets will be transferred to and preserved by the archives. Typical appraisal considerations include the evidential or enduring historical value of the geospatial datasets. Refer to your state’s general records management information to gather statutes, policies, and existing record retention and disposition schedules that may apply to geospatial data. In addition, you may want to consider the extent of geographic coverage of the dataset. For example, **statewide** geospatial initiatives such as a statewide orthophotography program might be a good candidate for retention by the state’s archival organization because it can potentially benefit a wider user community. Framework data layers\(^{13}\) may also be considered, especially geospatial assets that capture significant statewide infrastructure components or key boundaries that will be important for historical inquiry such as county boundaries, land ownership (cadastral), and land use change.

Involving both the GIS and archival staff when establishing a consistent set of questions and qualifying criteria will facilitate the appraisal process and ensure archival accessioning consistency as you evaluate new datasets.

Refer to **Appendix C: 2.3** for some additional resources on appraisal.

2.4. Geospatial Data Transfer Planning

There are several planning steps that will facilitate a smooth data transfer process. This next section outlines many of the specific planning activities to consider as you are preparing your geospatial archival project, in general, and your data transfer activities in particular.

* **A key first step is to determine what file formats your GIS organizations will transfer to the archival organization.**

Geospatial data are created in a variety of formats.\(^{14}\) Vector datasets include point data, such as locations of hospitals or airports, line data such as street centerlines, or polygon data such as county boundaries; raster data is comprised of an array (or grid) of cells that hold values for modeling continuous surfaces such as aerial or satellite imagery, or elevation models. The different geospatial vendors produce different file formats and most of the geospatial datasets comprise several files, which pose challenges for data transfer and preservation. Each file format has its strengths and weaknesses,\(^{15}\) and the geocaching team should understand the dataset formats used in your state, and work with the archivist to identify what dataset formats you will transfer and preserve. Your team should also consider any possible transformations required to prepare the datasets for preservation (e.g., exporting layers\(^{16}\) from a geodatabase as shapefiles, or converting datasets from one vendor’s format to a potentially more vendor-neutral format), as well as who will perform the transformations.

---


\(^{15}\) The shapefile format, originated by Esri™, has the benefits of a) being a relatively simple format, b) the format is documented, and c) and there are several geospatial software programs that can both produce and render shapefiles. There are also more complex geospatial formats, such as coverages and geodatabases, that can also maintain relationships within the geospatial data. In addition, there are derivative formats that can be produced, such as geospatial PDF, which may be useful for access purposes. Many of the geospatial software packages can produce the geospatial metadata in an xml file format, which provides a text-based tag file that is suitable for long-term preservation. For access purposes, some of the geospatial software packages can also transform the xml-based metadata file to a more user-friendly HTML version that is suitable for publishing to a website.

\(^{16}\) *layer:* the visual representation of a geographic dataset in any digital map environment. Conceptually, a layer is a slice or stratum of a 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.” from: Wade, Tash, and Shelly Sommer, eds. *A to Z GIS.* Redlands, California: ESRI Press. 2006.
Each of the GeoMAPP partner states established slightly different approaches for transferring vector geospatial datasets to their archives. Kentucky chose to transfer quarterly extracted snapshots of a complete Esri file geodatabase that contained a collection of layers, as well as some standalone shapefiles that were not included in the Esri geodatabase. Rather than a strict scheduled conglomerate snapshot approach, North Carolina selected and transferred batches of shapefiles (no geodatabases) of individual layers, when they were superseded in the North Carolina GIS clearinghouse, and if they were determined to be of archival value. Utah transferred both Esri shapefile and geodatabase formats for individual layers. Utah’s retention schedules direct datasets to be transferred annually. Montana chose an “archive on arrival” approach where data that they accept into their collection (in any format) is archived as it arrives at the Montana State Library, unless that data is formally archived by another agency (e.g., U.S. Census Bureau).

A determining factor in establishing these approaches, for the GeoMAPP states other than Montana, was the format structure of the original data, the difficulties involved in transforming the data into alternate formats, and the immediate usability of the data in its current format. These practical considerations need to be weighed against the long term preservation needs and costs. Details for each individual state’s data transfer process can be found in each individual state’s data transfer white paper, the GeoMAPP Interim Report:2007-2009,17 (Kentucky, North Carolina, Utah) and the GeoMAPP final report (Montana).

The GeoMAPP team conducted only a cursory exploration in transferring orthophotos. Due to the size of the imagery data captures, statewide or countywide orthoimage collections are often subdivided into small grid-based “tile”-ized files to increase portability. It is these collections of individual tile files that will need to be transferred and managed by the archival organization. There are two key aspects to consider when preparing to transfer orthophotos including:

1. What data formats of orthophotos you will transfer to address both preservation master file and archival access file needs, and
2. The file sizes associated with both uncompressed and derivative compressed orthophoto formats.

The sizes of your orthophoto collection can influence your choice of data transfer mechanism, and your data storage calculations for your data transmittal and archival staging and preservation repositories. The following offers some additional considerations related to the general orthophoto formats you will likely encounter:

- A collection of uncompressed orthophoto tiles (TIFF) can require terabytes of space. We recommend TIFF files for the preservation copies as they are the best representatives of the original data source. e.g., as an example, a 2010 statewide orthoimage flight of North Carolina, with six inch pixel resolution, totaled 17 terabytes of uncompressed data, and was subsequently subdivided into approximately 62,000 equal area TIFF-formatted tile files, each 286 megabytes in size.

- Compressed orthophoto tile files (commonly MrSID or JPEG), with their significantly smaller space requirements, are recommended as archival access copies. e.g., for North Carolina, 20:1 compressed MrSID-formatted tiles were produced which were 14 megabytes per tile.

- Other derivative formats, such as a tile mosaic which offers a seamless, stitched-together view of multiple tiles, may be considered as an additional or alternative access format. e.g., for each of North Carolina’s 100 counties, a 50:1 compressed MrSID-formatted tile mosaic was produced for each county, where each mosaic was approximately 10 gigabytes in size; Kentucky has a one terabyte collection of National Agriculture Imagery Program (NAIP) imagery stored in an Esri-formatted mosaic raster dataset within an Esri file geodatabase.

Note: You may want to consider contracting with your orthoimage vendor to provide both compressed and uncompressed images. If they are not available from the vendor, then the GeoMAPP team recommends that the GIS clearinghouse (or submitter) prepare the compressed images as part of the geospatial data preparation.

For a pilot or demonstration implementation, you may want to identify a small subset of geospatial datasets (e.g., 10-50) that are representative of the variety of geospatial formats targeted for archiving (e.g., shapefiles, orthophotos, geodatabases, projects, etc.). Your selections may also include representatives from the variety of dataset categories or themes (e.g., Biota, Boundaries, Transportation) to be archived.

Create a file transfer flow chart that illustrates the transfer of the datasets from the GIS data creators to the archival organization.

Visually mapping the workflow of the GIS data creation, use, and “archiving” is a critical tool to understanding how data is used, manipulated, and might be transferred to the archives. In the GeoMAPP project, the team determined that the most effective workflow for transferring data was for GIS creators to submit their data to the GIS clearinghouse. The GIS professional in the clearinghouse inspected the datasets, and performed clean-up and other preparations, such as ensuring that datasets had projections assigned. In Montana, the Montana State Library may also perform analysis on some data that will be helpful to subsequent users of the dataset. For the GeoMAPP project, GIS datasets identified as “permanent” were then transferred to the archival organization directly from the Clearinghouse.

The data flow diagram can start out very simply, and then be expanded as more of the details are defined through the course of the project. Refer to Appendix C: 2.4 for examples of simple flow charts produced by Kentucky and North Carolina.

2.4.1 Engage the Information Technology Organization
Connect with your Information Technology organization to discuss technical computing and infrastructure issues.

How you transfer your geospatial assets has a significant dependence on your technical computing and network infrastructure, so you’ll want to engage your Information Technology organization early. Including your information technology partners in your planning process can improve the selection and deployment of your overall archival solution. Solicit your information technology partners’ input on:

- Infrastructure planning, storage purchase, and software purchase, installation and configuration.
- Recommending and selecting tools for data transfer.
- Obtaining network access, firewall clearances, file system or network security access rights.

2.4.2. Digital Object Considerations for Data Transfer and Preservation - Geospatial Metadata Requirements
To promote long-term access and preservation, establish geospatial metadata requirements, and populate geospatial datasets following FGDC guidelines.

The best source for geospatial metadata is the dataset creator. Therefore, the ideal scenario is to provide metadata guidelines to geospatial creators, and rely on them to populate the metadata. GIS software, such as Esri ArcCatalog, can automatically populate a substantial subset of the metadata elements, especially the technically-oriented metadata, thus relieving the GIS professional from some of the data input burden.

The GIS professional and the archivist should work together to define the geospatial metadata elements required for submission of the dataset to the archival organization. The GIS professional will have the technical understanding of the various metadata fields and their implications. The archivist can advise which fields will promote the long-term preservation and access of the geospatial dataset. To inform your discussion you may want to refer to the GeoMAPP white paper, Utilizing Geospatial Metadata to Support Data Preservation Practices. You may also want to engage your state’s geographic council or policy-setting organization to assist in the definition and promotion of the metadata guidelines. The State of Montana established a metadata standard

---

(1200.XS4\textsuperscript{19}) for all geospatial data created by the state agencies and all data submitted to the Montana GIS Portal. The Montana GIS Portal offers a guideline document that illustrates the use of this metadata standard.\textsuperscript{20} These documents serve as useful examples.

The geospatial repositories such as the GIS clearinghouse and the archival organizations will need to determine if they will establish any policies regarding acceptance of geospatial datasets that are not adequately described. They may choose not to accept GIS data with incomplete metadata and hold noncompliant transferred datasets in a “conditional” folder (see Figure 7 and GeoMAPP Interim Report: 2007-2009 for a summary of Kentucky data transfer process). Or, you may decide to have clearinghouse staff populate the missing metadata before transferring it to the archival organization.

### 2.4.3 Digital Object Considerations for Data Transfer and Preservation - Naming Conventions

Identify naming conventions for datasets and dataset files to be transferred.

Many of the geospatial datasets will be stored as collections of files on a file system. The GeoMAPP team recommends that your team adopt a self-describing or self-documenting naming convention as part of the planning phase. The name could include key descriptive aspects of the dataset such as the geographic area or extent, the name or theme, and the time period of the content.

Consider a file naming convention such as:

```
Geographic-area-or-extent_Dataset-name-or-theme_Year_Month
```

For example: Wake_Parcels_2005_09 or Shellfish_Growing_Areas_2008_08 or

```
<Location><Theme><Year>
```

and use “CamelCase” where each initial letter of each word is capitalized

For example: MTLandCover2010 or MTCensusBlock2000

Some GIS software uses the name of the dataset as the basis for the file name, so similar naming conventions may be applied for dataset naming.

File names should not contain blank spaces or special characters as some software packages are sensitive to these characters and may have trouble opening them. In place of the blank space, use an underscore or a dash, as noted in the example above. You may also need to be aware of the file name character length, as well as the total absolute path name character length, as some software packages, such as backup software, have limitations on the maximum number of characters. Check with your local IT professional to see if your systems or software experience these types of constraints, as it can significantly impact your file naming strategy.

Lastly, the GIS and archival staff should agree to a file name convention, and consider engaging your statewide geographic coordination bodies to promote the adoption of a file naming standard. This file naming convention should be used consistently. Your team should communicate the naming scheme out to the geospatial data creators either through the clearinghouse or through your statewide geographic council. Receiving files that comply with suggested file naming conventions can reduce the GIS clearinghouse and/or archival file clean-up and preparation efforts.

### 2.4.4 Digital Object Considerations for Data Transfer and Preservation - Data Organization

Another important planning consideration is how to organize the data for transfer and preservation. Archival best practices recommend that the arrangement of archived materials reflect the original order and arrangement as they are received, as that arrangement can provide insight into the creator and/or the creator’s work processes. In the GeoMAPP project we encountered two general organization approaches, largely driven by the method used by the GIS clearinghouse:


1. Datasets that were managed **individually**, either in Esri shapefile or Esri geodatabase formats, and
2. Datasets that were managed **collectively** in an Esri geodatabase.

The following sections look at these two general organizational methods, and present planning considerations that may also influence the structure of your archival repository, your directory structure, and if relevant, your database.

2.4.4.1 Data Organization - Individual Datasets

*Establish dataset organization structure (database, file system, and associated organization scheme) for the archival repository.*

After identifying the geospatial formats that will be preserved, the GeoMAPP team recommends you define an organizational structure for the datasets for your digital repository. The organizational structure could be dependent upon the data retention strategy and the form of the datasets being retained, such as:

- individual geospatial assets (e.g., individual shapefiles or a single layer in a geodatabase format)
- a collection of geospatial assets (e.g., multiple layers in a geodatabase, or an orthophoto comprised of numerous individual tiles).

For archival organizations that will be managing individual datasets, the GeoMAPP team recommends that the first level of organization be based on the general type of geospatial asset being stored:

- Vector Digital Data
- Raster Digital Data (assets such as aerial imagery, orthophotography, and elevations)
- Projects
- Digitized Maps

To aid in consistency of archived data categorization, the GeoMAPP team also recommends that vector digital data can be categorized based on the ISO 19115 – Geographic Information Metadata topic categories or themes, as the GIS community is already familiar with these categories, and other GIS repositories and inventory systems use this general organizing method.

<table>
<thead>
<tr>
<th>Topic Categories (a-im)</th>
<th>Topic Categories (in-u)</th>
</tr>
</thead>
<tbody>
<tr>
<td>biota</td>
<td>inlandWaters</td>
</tr>
<tr>
<td>boundaries</td>
<td>intelligenceMilitary</td>
</tr>
<tr>
<td>climatologyMeteorologyAtmosphere</td>
<td>location</td>
</tr>
<tr>
<td>economy</td>
<td>oceans</td>
</tr>
<tr>
<td>elevation</td>
<td>planningCadstre</td>
</tr>
<tr>
<td>environment</td>
<td>society</td>
</tr>
<tr>
<td>farming</td>
<td>structure</td>
</tr>
<tr>
<td>geoscientificInformation</td>
<td>transportation</td>
</tr>
<tr>
<td>health</td>
<td>utilitiesCommunications</td>
</tr>
<tr>
<td>imageryBaseMapsEarthCover</td>
<td></td>
</tr>
</tbody>
</table>

The NSGIC established the RAMONA (Random Access Metadata tool for Online National Assessment) for the national GIS Inventory, and defined a categorization scheme based on the ISO 19115 topic categories. The inventory defines a series of data layers which are categorized into data categories that then map to the ISO topic categories. In addition, NSGIC designated a subset of the data layers as Framework layers. This subset includes key geospatial asset attributes such as state, county, city, town and village boundaries; as well as theme attributes

---


22 National GIS Inventory. [http://www.gisinventory.net](http://www.gisinventory.net)
such as hydrography and watershed boundaries; digital orthophotography and land cover imagery; and airports and railroad lines.

The following table offers an excerpt of the Ramona Data Categories and Layers as documented by the Wisconsin Land Information Clearinghouse\(^{23}\) which North Carolina used as the basis for its directory organization:

<table>
<thead>
<tr>
<th>Framework Data Category</th>
<th>Data (sub)layer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boundaries</td>
<td>American Indian Reservation</td>
</tr>
<tr>
<td></td>
<td>Cities/Towns/Villages</td>
</tr>
<tr>
<td></td>
<td>Civil Township or Equivalent</td>
</tr>
<tr>
<td></td>
<td>Counties/Parishes</td>
</tr>
<tr>
<td>State</td>
<td></td>
</tr>
<tr>
<td>Imagery/Base Maps/Earth Cover</td>
<td>Digital Orthophotography/Orthoimagery</td>
</tr>
<tr>
<td></td>
<td>Land Cover</td>
</tr>
<tr>
<td>Inland Waters</td>
<td>Hydrography</td>
</tr>
<tr>
<td></td>
<td>Watershed Boundaries</td>
</tr>
<tr>
<td>Transportation</td>
<td>Airports &amp; Airfields</td>
</tr>
<tr>
<td></td>
<td>Railroad Lines</td>
</tr>
<tr>
<td></td>
<td>Roads/Street Centerlines</td>
</tr>
</tbody>
</table>

Figures 3 and 4 offer two examples from the GeoMAPP states of North Carolina and Utah, respectively, organizational schemes for the individual vector datasets using the topic categories and layers. In North Carolina, the GIS Clearinghouse organized the datasets into this foldering structure prior to submitting to the North Carolina State Archives, so the Archives was simply preserving the organization structure of the submitted files, which streamlines the subsequent file manipulation the Archives may need to perform as it is processing and arranging the received datasets.

---

The Utah State Archives uses the categories as the basis for organizing their prototype ftp download site where users can access and download datasets, as seen in Figure 4.

The Utah State Archives and the Montana State Library are implementing database-driven solutions to manage the “preservation masters” in their archival collections of GIS datasets. Therefore, the directory hierarchies for their preservation repositories are shallower and hold less descriptive characteristics within the directory structure. Figure 5 illustrates the MSL folder structure. The Dark Archive folder is the designated directory where the Montana State Library will be storing their preservation masters. Datasets that are being regularly updated will be stored in the Series folder, under the relevant framework subcategory, datasets created as part of specific discrete projects will be stored in the Collections folder, while all other datasets will be placed in a subfolder corresponding to their year.
However, the Montana State Library still plans to use the topic category as a key descriptive metadata field for each vector dataset, and will use the topic category as a fundamental access mechanism through the search function. One benefit of this approach is that some datasets can be associated with multiple topic categories. In the file-based organization method, a dataset will be stored in only one topic category subfolder, while in the database-driven solution, a single dataset could be assigned to multiple topic categories, and appear in the search results-set for any of its related topic categories.

Using the topical categories that are recognized by the GIS community as a basis of organizing the directory hierarchy or characterizing datasets in a database promotes consistency across GIS applications, which can ultimately enhance both preservation and access (see GIS Inventory24 for a nationwide example of the benefits of common dataset categorization scheme).

In organizing imagery assets, such as statewide orthophotos for example, the GeoMAPP team found it useful to organize these files based on county. Government organizations may invest in recurring flyovers every five or ten years to be able to capture imagery of the evolving landscape, so creating a subcategory based on year may also be useful. Orthophotos may also be captured at multiple resolutions, or in black-and-white and color, which also suggests additional subcategories. See Figure 6 for an example directory organizational structure for a collection of orthophotos.

As in the vector example above, if these assets are being managed in a database, the directory structure will likely be flattened, and the descriptive characteristics, such as county, year and scale may be represented as attributes in the database rather than as subfolders in a directory hierarchy.

---

24 GIS Inventory. [http://gisinventory.net/](http://gisinventory.net/)
Metadata Note:
To facilitate the proper categorization of datasets for both the GIS clearinghouse and the archives, the GeoMAPP team recommends that the geospatial creator populate the <themekt> (theme keyword thesaurus) and <themekey> (theme keyword) metadata attributes with the Ramona / ISO 19115 Topic Categories:

- ISO 19115 Topic Category
- Transportation
- Railroad Lines

This provides useful categorical information for long-term preservation, plus provides useful metadata for more sophisticated archival applications that can then either directly index the metadata file on behalf of the dataset, or extract the metadata to populate a database that supports attribute-based navigation and/or search.

2.4.4.2 Data Organization—Bulk Geodatabase Snapshots

Alternatively, you may choose to build your archival repository through a regularly scheduled bulk-“snapshotting” export approach. The bulk-snapshot captures all digital assets at a particular point in time. In this case, you might create separate sub-directories for each snapshot time. Additionally, narrower snapshots may be captured that retrieve a subset of the assets based on geospatial type, such as raster or vector. Figure 7 illustrates Kentucky Department for Libraries and Archives’ (KDLA) repository holding several quarterly snapshots of their geodatabase. Inside the geodatabase, they have organized the feature classes according to the ISO data categories and layers.

Figure 7. Geodatabase snapshots (Kentucky) a) folder structure view, b) Esri ArcMap view
Your repository may also include datasets that have access restrictions. Either the data is confidential by statute or it may be closed for some period of time. Figure 7 illustrates a file organization approach used by the KDLA. Datasets that are available for general public access are placed in the open folder, while datasets with restricted use policies are placed in the closed folder. The conditional folder is a holding area for datasets that do not comply with the submission requirements, such as not providing all requisite metadata.

The geodatabase snapshot is valuable for when the geodatabase includes information about the relationships between the various datasets. But it does have the disadvantage of storing multiple copies of the same dataset across multiple snapshots, and therefore consumes more storage than is the case in the individual dataset capture approach.

### 2.4.5 Dataset Verification

*Identify the tools, methods, and policies for initial dataset verification.*

Prior to transferring the dataset to the archival repository, you should conduct a verification of the submitted datasets to verify that it is free of viruses, satisfies your metadata policies, and is functionally accessible. In this section, you’ll identify the tools used for virus checking, inspecting metadata for compliance with your policies, and functionally verifying the dataset. You will also want to consider how you will notify dataset submitters when their datasets are accepted or rejected.

While identifying the tools and methods for verifying your datasets is important, another critical planning step is defining your policies regarding the required metadata content. This is another key collaborative planning activity between the GIS and archival professionals, as the GIS professionals can share the purpose of each of the metadata fields, and the archivists can identify those files that are important for preservation. It may also be beneficial to involve your statewide geographic council, as these guidelines could impact GIS professionals across the state.

**Virus Check**

Identify the virus check software and the virus checking procedure that will be followed by both the dataset submitter in preparation for transfer, and the dataset receiver upon receipt of the newly transferred datasets. Also, consider the storage configuration for the archival organization related to the virus check process. It is recommended that files be held in an isolated pre-processing area for a period of time (e.g., 30 days) prior to running the virus checking software to give the virus checking software time to be updated with the most recent virus resolvers that could have infected the files on their date of submission.

**Validation of Geospatial Metadata**

Identify the process and/or tools that will be used to validate the geospatial metadata. The metadata may be visually inspected and compared against a metadata policy, or inspected and validated with tools. The FGDC suggests several tools, many free, to check the compliance of the metadata record per the CSDGM guidelines. While these tools cannot check the richness of the metadata content, they can verify that key geospatial metadata fields have been populated. To facilitate presentation and visually-based inspection, there are several style sheets available to transform the xml to a more human-readable html or text version. GIS tools such as the USGS Metadata Validation Service and GIS products such as Esri’s ArcCatalog offer several style sheets to display the metadata in a variety of viewing styles.

The GeoMAPP team explored the USGS Metadata Validation Service which is based on the Metadata Parser utility. To view or validate a metadata file, you simply pick a metadata file from your local system through the

---

25 Alternatively, instead of “open” you may want to use terms such as “unrestricted” or “public,” and instead of “closed” you may want to use “restricted.”


27 USGS. “Geospatial Metadata Validation Service.” Available at: http://geo-nsdi.er.usgs.gov/validation/

v1.0 12/1/2011
web-based interface, and select the [Validate] button. The results page offers options to view the metadata in its original XML text format, or in a more user-friendly view that’s been formatted with an XSL style sheet (e.g., Outline and Text options). It checks CSDGM conformance aspects such as:

- conditionality (i.e., are all mandatory elements populated?).
- data formats.
- element domains

The Metadata Parser cannot evaluate the quality of the text entered in the metadata fields. Therefore, you’ll likely need to perform a visual inspection of the metadata, to review the descriptive contents of the metadata fields.

The Metadata Parser tool was found to be informative for flagging missing metadata fields, though the output may be difficult to wade through. Unfortunately, the Metadata Parser appeared to generate some false errors, so its use would require a thorough item-by-item review of each error and the associated line entry in the metadata file.

The GeoMAPP team also explored Esri’s ArcCatalog (ver. 9.3) for visually inspecting the metadata. ArcCatalog was useful for visually inspecting the metadata file, both to review the presence of required fields, as illustrated in Figure 8, as well as to review the quality of the actual metadata content, especially the descriptive fields.

![Identification Information](image)

**Figure 8. View of metadata that indicates missing FGDC required fields**

However, you should be aware that ArcCatalog (ver. 9.3) includes an option to automatically populate missing values for some metadata fields that can be determined from properties in the dataset. This may actually be beneficial from an archival perspective as it may produce a more robust geospatial metadata file. However, it has the disadvantage of overwriting your originally received metadata file, which is also an important archival artifact. If you are aware of this side-effect, make a safe copy of the original metadata file (e.g., shp.xml -> shp.xml.ORIG) prior to opening the dataset in ArcCatalog, and then take advantage of the ArcCatalog’s auto-update feature. Alternatively, you can disable the metadata auto-update feature, and review and evaluate the metadata file based solely on its original content.

Figure 9 illustrates how to disable the auto-update feature in ArcCatalog (ver. 9.3), select Tools from the main menu bar, and then select Options. On the [Metadata] tab, of the Options window, uncheck BOTH

---


v1.0 12/1/2011
the “Creating metadata” and “Updating metadata” checkboxes that enable the automatic creation of metadata.

![Esri ArcCatalog (ver. 9.3) interface to disable auto-update of the metadata file](image)

**Figure 9.** Esri ArcCatalog (ver. 9.3) interface to disable auto-update of the metadata file

**Functional Verification of Datasets**

Identify the process and tools for the archivist to functionally verifying the received datasets. The data submitter and the data receiver should use similar tools and techniques. The GeoMAPP partner teams used ArcCatalog to verify individual shapefiles, individual feature classes in an Esri geodatabase, and orthophotos and used them to view:

- The preview graphic of the dataset, to verify that the dataset renders graphically,
- The shapefile properties, to verify the dataset has a projection,
- The attributes table, to verify that the dataset data is populated,
- The geospatial metadata, to verify that geospatial metadata is complete, based on your policy.

Fortunately, the GIS tools should not be too difficult for the archivists to learn to use to perform basic inspection functions. There are also some free GIS viewer tools available. Esri offers a free ArcGIS Explorer, and the USGS offers a free viewer called dlgv32. While the viewer functionality is free, options to “save” require purchase of the product.

**2.4.6 Dataset Packaging Mechanism for Transfer and Post-transfer Validation**

*Identify the packaging mechanism for transferring datasets (e.g., BagIt, zip, tar).*

Datasets are usually comprised of several files, and sometimes extremely large files. Therefore, your geoarchiving team needs to agree on mechanisms for packaging and transferring the geospatial datasets to the archival organization. File packaging options the GeoMAPP teams explored included tar, zip, and BagIt. These tools provide differing approaches to transferring collections of files. zip and tar simply package all of the files into a single file “package” that may be transferred. Once transferred, the files are extracted from the package using the

---


v1.0 12/1/2011
same tool (e.g., zip or tar). BagIt, alternatively, creates a “logical” package where the collection of files are simply put into a standard Windows folder. BagIt creates its package-like structure by creating a data folder which holds all of the folders and files to be packaged. The original folder hierarchy and file arrangement are maintained. In addition, BagIt also generates useful administrative text files that include: 1) a manifest file recording all of the files in the bag and each file’s checksum value, and 2) a file that records the file count for the bag, and the date and time the bag was created. These administrative files are of special interest for archivists who need a documentary record for the transferred files. To transfer a bag, the entire BagIt hierarchical folder structure must be transferred (compared with transferring a single file when using either zip or tar). The bag is “unpacked” on the receiving system by simply copying the files out of the bag’s data subfolder using the standard Windows file management functions (see Figure 10).

Figure 10. BagIt folder structure: (top) Contents of a bag labeled “20101216_Adhoc_NCOneMap”; (middle) the bag’s data folder holding four folders of geospatial themes of datasets; (bottom) the geospatial dataset files in the Biota theme folder

If you used zip to package the files for transfer, the receiver would likewise be required to use the zip utility to extract the files from the zip archive. With BagIt, no special utility is required on the receiver’s system to extract the folders and files from the bag. This offers a potential benefit for archivists looking to minimize their reliance on special file formats for file transfer and ingest processing.

The most critical aspect of the file transfer is validating that the received files are equivalent to the transmitted files, that is, confirming that the files have not been corrupted or data lost in the course of the transmission. Using tools that generate checksums for each of the files before and after they are sent is one technique to validate the transmitted files arrived at their destination unchanged. However, as geospatial datasets comprise numerous files,

BagIt Code Download: http://sourceforge.net/projects/loc-xferutils/
it may be burdensome to manually calculate, store, and validate checksums on the (potentially large) collections of files that comprise each dataset for both the originating and receiving systems (though a script may be written).

One approach for doing bit-level validation of transferred files is to first package the dataset files into a single “archive” file using zip or tar, and then compute the single checksum value for the single zip/tar file prior to transfer. The receiver then verifies the checksum of the zip or tar file upon receipt. Checksum tools that offer file-oriented checksum mechanisms include: Checksum Tool, MD5summer, and md5deep. BagIt’s verifyvalid option provides an automated mechanism for validating the checksum on each of the files in the “bag,” alleviating the user from having to compute, save, and validate checksums manually for each of the individual files. Plus, BagIt offers the benefit that the verifyvalid option can be re-run at a later time to re-validate the bit-level integrity of the collection of files. This can be especially useful in the archival scenario where files need to be periodically checked to make sure that they have not been changed or corrupted over time.

Maintaining bit level data integrity over time is another important preservation consideration, as media can experience gradual decay, or individual bits of storage may be lost due to media degradation (known as bit rot). Preservation best practices recommend that a bit-level fixity archival metadata record be recorded and retained to validate future data integrity at the bit level. An archivist needs to identify a bit-level validation mechanism, such as a checksum. The archivist should also record in the archival metadata the validation mechanism and the fixity data (e.g., MD5summer, and a checksum value) so that the file fixity may be continually re-validated in the future. Depending on your storage and arrangement strategy for your archived datasets, you may want to consider using the fixity value generated from your file transfer as the basis for monitoring the data integrity of the preservation master dataset. This also reinforces that the file’s authenticity has been retained from transmitter to archival storage.

In the event a file is found to be corrupted, it should be restored from its security copy. This is why archival best practices recommend that each preservation master have at least one security copy. In the case of a validation failure in either file, it may be restored from its companion. Advanced storage management systems can automate file integrity checks and repairs in the event a problem is found, so you should check with your IT partner to determine the best method for tracking file integrity and addressing file integrity issues as they are encountered.

Kentucky, Montana, North Carolina, and Utah explored BagIt for packaging datasets for transfer. Montana bagged the multiple files that comprise a single dataset, transferred the bag, and then validated the transfer at the bag level, essentially validating the single dataset. North Carolina packaged a collection of datasets into a single bag. The North Carolina file validation process was streamlined as the entire collection of transferred datasets could be validated with a single invocation of the BagIt command and its verifyvalid option. However, this approach does not allow for future validation capabilities at the individual dataset level, which is necessary for ensuring the long-term bit-level integrity. With this mass transfer approach, a separate fixity record needs to be produced for each file in the preservation master for each dataset. In the absence of an automated file-based checksum solution, BagIt can be used to re-package each dataset individually. These individual dataset bags can then be re-validated at a later time using the verifyvalid option. With its command-line invocation, scripts can also be developed to invoke the BagIt utility, which can be useful to run regularly scheduled fixity checks.

The GIS and archival organization will need to work together to figure out the technical mechanism and method to package and validate the integrity of the transferred files. The archival organization will have the additional task of determining the method used to monitor the ongoing integrity of the datasets once they are loaded into the archival repository. These may be the same or different methods.

---

34 MD5summer Tool: [http://www.md5summer.org/](http://www.md5summer.org/)
35 md5deep Tool: [http://md5deep.sourceforge.net/](http://md5deep.sourceforge.net/)
37 Even in the case that the storage management system can monitor and repair file integrity issues, a fixity value will still be required for archival purposes, to validate the authenticity of the file at a later date. The fixity value can serve as a sort of digital “finger-print” of the file that can be used to validate the authenticity of the file at a later date. So, even if you do not need to record the fixity value for ongoing file maintenance, you will still likely need to record a fixity value for ensuring the authenticity of the digital asset.
2.4.7 File Transfer Size Requirements

*Estimate the size of files that will be transferred.*

Geospatial datasets can be comprised of a collection of files that together consume megabytes of space (or even terabytes as may be the case for orthophotos). A key preparation step is to assess the sizes of the geospatial datasets you will be transferring. The sizes of the datasets that you want to transfer may impact your available file transfer medium options, as well as the time of day you may transfer your files. Your options for transferring your files may depend on technical infrastructure details such as your network infrastructure and network bandwidth availability. Internal network transmission restrictions may not allow the transfer of files that exceed a certain size, and local agency policies may also influence the time of day that data may be transferred (e.g., large network data transfers may be restricted to “off” hours).

Having your geospatial inventory in hand, and estimates of the sizes of the geospatial datasets will help prepare you for conversations with your information technology partners to determine if and when you can transfer geospatial dataset packages electronically through your network.

2.4.8 File Transfer Mechanism

*Identify the mechanism and medium for transferring files sets.*

There are essentially two options for transferring files:

- Transferring files via the network,
- Transferring files to a portable storage device that will be physically moved and attached to the target computer or server.

Transferring files via the network is the most convenient mechanism for both the GIS clearinghouse submitter and the archival organization receiver. However, due to network restrictions (e.g., file size limitations, security restrictions), your organization may be restricted from using the network to transfer large geospatial data packages. This option may be acceptable, however, for smaller datasets.

*Establishing storage for data transfer*

In addition to determining how you will transfer your files, the GeoMAPP team recommends that you allocate storage for data preparation in both the GIS clearinghouse’s and archive organization’s repositories. The GIS clearinghouse will likely need a staging area to verify, prepare and package the datasets for transfer. Similarly, the archival organization will need a staging area to receive the files, validate the files’ integrity was maintained through the transfer, perform virus checks, and functionally verify the datasets prior to processing and depositing them into the preservation repository. For the archival organization, this storage space should be dedicated to supporting the preliminary data transfer validation tasks, and ideally, isolated from the preservation repository to minimize the chance to transfer any viruses to the preserved data. This is in addition to the storage space dedicated to the preservation master, preservation master security copy(ies), and use/access copies of the dataset.

*Pushing datasets vs. Pulling datasets*

Data will generally be submitted (“pushed”) to the archival organization by the data submitter, or retrieved (“pulled”) by the archival organization from an agreed to distribution location. However, there may be special cases where the archives does not have a formal arrangement with an organization, and the organization publishes important datasets to its website. Additionally, the organization may not contribute its datasets to the state’s GIS clearinghouse. In this case, the archives may choose to informally pull datasets available on a website without the explicit participation of the organization or data creators. This is generally not a recommended best practice, except when the web is the only available source for these datasets, and they are identified as preservation-worthy, and are at risk for being superseded or lost. The archives should prepare an inventory of websites and datasets of interest to the archives. The archives should also define how geospatial metadata will be reviewed and how any deficiencies may be addressed, and identify the frequency that datasets will be pulled from these websites. If possible, the archives should attempt to engage the website owner, and engage them as a regular geospatial data submitter, so that datasets are submitted to the archives using its regular processes.
2.4.8.1 Network-based Transfer

When conducting electronic transfers via the network, you should conduct preliminary planning, and possibly testing, with your information technology partners to determine what and when you can transfer files across the network, the characteristics of the performance of your network transfers, and what tools you will use to conduct those transfers.

Collaborate with your Information Technology Partner

As we advised earlier in section 2.4.1, you should already have established communication with your information technology team. To conduct network-based file transfers, you will likely need assistance from your information technology partners to establish network connectivity, especially if you are transferring files across different networks, and you may want to discuss your options with regard to your file transfer tools selection. They will also be able to inform you of any restrictions regarding the time of day you may transfer files, and support you in configuring appropriate security settings within the network, such as configuring network firewalls.

Network bandwidth, network access, and time of day to transfer

In order to electronically transfer files across the network you must determine the network bandwidth (i.e. size of the network pipe) available for transferring data. You should also identify the appropriate time of day to transfer so as not to impact the productivity of the other network users. If files are being transmitted to the archival organization from outside of the archives’ network, you may also have to negotiate firewall permissions or other network security mechanisms to allow transfer of files between the organizations’ networks. Depending on your information technology department’s policies and procedures, this can take some time to work through all of these details.

Method of Transfer and Tools

If you determine that you will be able to transfer files electronically across the network, you will also need to identify what software will be used to transfer the files. A really convenient mechanism is to configure a network share drive. With appropriate network access privileges, both organizations can access a common shared drive by using the familiar Windows file browser to copy files from one system to the other. However, organization and network security policies may restrict this capability. ftp (file transfer protocol) is also a common command-line tool to transfer files from one computer to another across the network, and tools are now available that provide a graphical user interface front end for conducting ftp-based file transfers. We recommend you use ftp’s secure form to safeguard your file transfers. 38 Once again, we advise that you work with your IT department to identify the best tools to perform the actual data transfer.

GeoMAPP States Network-based Transfer Methods

Each of the three GeoMAPP states transferred some of their GIS data by pushing it across the network, and worked closely with their IT partners to establish access and troubleshoot technical issues. Universally, the IT groups of each state determined that the ideal time to transfer large amounts of data was after business hours.39 In this way, it would not consume the available network bandwidth and adversely affect productivity during normal working hours. In addition, both the GeoMAPP data submitter, receiver, and IT closely monitored the transfer to see if any problems were experienced.

The GeoMAPP teams used shared network drives that both the GIS submitter and the archivist receiver had read and write access. However, they chose different folder strategies to receive the files at the archives. In Kentucky and Utah, the GIS group deposited their files, via the network, to the “preservation” folder of the geoarchives. North Carolina, alternatively, established a “staging” folder on a network-accessible share drive into which the North Carolina GIS group deposited their files. Once the North Carolina GIS staff deposited data in this folder they sent a notification via e-mail to the archivist. The North Carolina archives staff then accessed the files in the “staging” folder and began the ingest procedure, from which datasets would be processed and prepared and then transferred to the “preservation” folder. Montana’s GIS clearinghouse and archival responsibilities are co-located within the Montana State Library, and share the same local area network (LAN) infrastructure. Montana transferred files through the LAN from the GIS clearinghouse repository to a submission folder on the shared

39 For example, North Carolina conducted a test to transfer 395 GB of data that took several hours to transfer.

v1.0 12/1/2011
network intended for archival preparation and processing. The files were then transferred to a folder in the permanent archival repository, and the integrity of the transferred files was validated. Montana also conducted a wide area network (WAN) transfer to a secure ftp site hosted by KDLA. The WAN transfer required some initial configuration setup by Kentucky and Montana staff to establish permissions and access into Kentucky’s secure network.

Kentucky and Utah also used Microsoft’s robocopy, a command-line utility to copy files. robocopy offers the benefits of handling network interruptions, and preserving file ownership, time stamps and audit information on transferred files.


2.4.8.2 Transfer via a Removable Storage Device

Due to a lack of shared network access, the size of the dataset package, and/or network restrictions your organization may be prohibited from transferring files across the network. As an alternative, the GIS clearinghouse may need to download the packaged datasets to a removable or portable disk, deliver the disk to the archival organization, and the archivist will then upload the packages to the archival system’s storage repository.

In this case, you will need to select a portable storage device that will have sufficient storage capacity and technical connectivity for both the GIS clearinghouse and the archives. Prior to transferring datasets, it is recommended you perform a validation test that confirms the GIS clearinghouse can successfully connect and write to the storage device, and that the archives can, likewise, successfully connect the storage device to their computer, and access and retrieve the files from the device. You may need assistance from your IT partner to connect the device and/or download new drivers for your computer to properly access the device.

2.4.9 Identify Archival Submission Policies

As part of your planning you should define policies regarding the submission of geospatial datasets.

In developing your policies, you may consider questions such as:

- Who is authorized to submit geospatial datasets to the archival organization?
  - Will datasets be accepted from only the GIS clearinghouse?
  - Can state agencies, municipalities, and local governments (data creators) submit datasets directly to the archival organization?
  - Who will be considered the submitter? The GIS clearinghouse (if applicable) or the original GIS creator?
- How does the data transfer fit into the archival organization’s accessioning requirements and process?
  - Will donor agreements be required?
  - Are retention schedules in effect?
- What triggers the transfer of datasets to the archival organization?
  - Will datasets be transferred when received by the GIS clearinghouse, or when modified (the GIS clearinghouse may add value to the received datasets), or when superseded with an updated version from the GIS creator?
  - Will datasets be transferred from the GIS clearinghouse based on a scheduled frequency (e.g., monthly updated datasets, quarterly snapshots)?
  - For organizations where the GIS clearinghouse and the archival organization are co-located, will the datasets be transferred to the archival organization upon arrival to the clearinghouse, or will they be archived when superseded or on a defined schedule?


v1.0 12/1/2011
• Is there a geospatial metadata policy or are there guidelines that define which geospatial metadata fields are expected to be completed for archiving?
  • What happens if the archival receives a dataset with incomplete metadata?
• What dataset formats will the archival organization accept?
• What dataset preparations will the GIS clearinghouse conduct prior to transferring datasets to the archival organization?
  • e.g., provide a format transformation service to prepare datasets for transfer to the archives (e.g., extracting feature classes from Esri geodatabases as shapefiles).
  • e.g., ensure that shapefile datasets are configured with projections.
  • e.g., ensure that geospatial metadata is complete.
• What is the packaging strategy for transferring the datasets to the archival organization?
  • Will a delta approach be used in which the GIS clearinghouse collects for transfer only those datasets that have been changed or superseded with newer datasets since the last transfer? The datasets could be transferred as they are superseded, or scheduled on a quarterly basis to reduce and regulate the transfer process.
  • Alternatively, will a monolithic approach be employed, that takes a regularly scheduled snapshot of all datasets, and transfers the entire dataset collection to the archival organization? Keep in mind that this technique will require greater storage capacity over the long run as datasets will be duplicated across the snapshots.
• Once datasets are loaded into the preservation repository, will special (GIS) software be required to access and/or manage the archival organization’s preservation master? 41
• Once datasets are accessioned by the archival organization, are they available for the public? Or will there be any restrictions imposed (e.g., confidentiality, copyright, embargo period) on the datasets that are transferred to the archives?
• Once datasets are made accessible, will special software be required to use the archive’s access copy?
  • What type of derivatives can be created that minimize the need for a patron to use special software (e.g., geospatial pdf of a shapefile, user-friendly HTML version of the geospatial metadata)

2.4.10 Create a Written Document for Data Transfer to the Archival Organization
Create a written document that describes the processing for the preparation for transfer, transfer, and immediate validation processing of received geospatial datasets.

• Establish a schedule documenting the transfer of ownership of datasets (may take the form of a formal archival schedule, or a memorandum of understanding).
• Develop a written specification that describes how the submitting organization will prepare, arrange, package and transmit geospatial datasets. Include details regarding the use of any special software or tools.
• Develop a written specification that describes how the datasets will be received and validated by the archival organization. Include details regarding the use of any special software or tools.
• Document the organizational structure (folder structure) of the datasets for transfer. We recommend that the directory organizational structure for the datasets to be transferred reflect the archival arrangement as we found that this significantly reduced the archival ingest processing time and effort by eliminating the need for the archivist to assess and re-organize the datasets in preparation for archival storage. We also recommend that you consider organizing your datasets based on the ISO or Ramona categories to promote dataset organizational consistency, as these are general categories that are in common use in many GIS repositories and can aid access and discovery.


v1.0 12/1/2011
3. Role-based Tasks for Geospatial Data Transfer

While the preceding section focused on the planning and preparation for your geospatial data transfer, the following section provides a role-based summary of the key data transfer-related tasks for:

1. Geospatial Data Creators
2. GIS Clearinghouse Professionals (and/or geospatial data submitters of geospatial data to the archival organization)
3. Archivists

Refer to the section that most closely aligns with your role and responsibility in supporting the transfer of geospatial datasets to the archival organization.

3.1 Geospatial Data Creators Tasks that can Facilitate Data Transfer and Preservation

The geospatial data creators can be a significant contributor in the preparation of geospatial data to be transferred to an archival organization for long term storage and preservation, most significantly through their development of the geospatial metadata, and secondarily, through their foldering organization, and their file and dataset naming practices. The statewide GIS coordinating council, the GIS clearinghouse and/or the archival organization can schedule communication sessions with the geospatial data creators to encourage and guide them in preparing geospatial datasets that are populated with rich and descriptive metadata, as the creators are the best qualified to create the geospatial metadata. Additionally, they may be instructed on naming conventions adopted for dataset transfer. Alternatively, the GIS clearinghouse can rename files upon receipt to comply with the prescribed naming conventions.

When engaging third party contractors to create geospatial data (e.g., orthophotos), you may consider including as part of the contract terms:

- metadata requirements and acceptable level of description detail
- dataset and file naming requirements

3.2. GIS Clearinghouse Geospatial Data Transfer Tasks (or Submitter of geospatial data to the Archival Organization)

The submitter of the geospatial datasets has several tasks that will facilitate the overall data transfer process. Cross-organizational planning and coordination, definition of submitter processes, and selection of technologies and tools to transfer the datasets have been described as part of the planning process above. The following provides a brief summary of the key tasks the dataset submitter completes in the transfer of geospatial datasets to the archival organization:

1. Establish a working-area repository to prepare geospatial files prior to their transfer to the archival organization. At a minimum, you will need repository space for at least two additional copies of the original geospatial dataset files:
   - a “work-in-progress” copy that will be manipulated in preparation for data transfer (e.g., metadata cleanup, renaming of files to match file naming conventions, organizing the files in the anticipated directory structure)
   - the copy of the “packaged” file that will be transferred

2. Name geospatial dataset files to be submitted according to file naming standards and conventions.

3. Functionally verify that the dataset content can be accessed.
   - This will require some sort of GIS software. The following provides suggestions for verifying a few geospatial format types:
• Vector Datasets (point, line, polygon)
  A shapefile will generally represent a single layer, and will store point, vector or polygon features. For individual shapefiles, each shapefile should be opened and visually inspected. The shapefile graphic representation should render reasonably, whether it consists of points, lines or polygons. You should validate that the dataset file properties appear to be properly populated, especially confirming the presence of the projection. You should also validate that the shapefile data attributes (view data table) are populated for the features represented in the dataset.

![Figure 11. View shapefile rendering, attribute table, and metadata (e.g., Esri ArcCatalog v. 9.3)](image)

• Raster Datasets / Orthoimagery
  An orthoimagery collection is generally a collection of raster-based digital images. It may consist of orthophotos, which are geographically corrected (rectified) digital images resembling photographs. Digital elevation models are another source of geospatial raster data. File derivatives from Digital Elevation Models illustrate topographic characteristics, such as slope, aspect, hillshade, surface water flow direction, or simple elevation representation. Orthoimagery datasets, which can consist of hundreds or even thousands of individual image tiles, should be visually inspected by randomly selecting a sampling (e.g., 5%) of images. You should view the dataset file properties of the individual image to see that fields such as Data Type, Columns and Rows, Number of Bands, Cellsize, Format, Pixel Depth, Pyramids, and Extent values are populated. Orthoimage access can be enhanced for long-term preservation by creating a companion “index” shapefile that provides a summary and arrangement for the tiles that collectively comprise the orthoimage.
Geodatabases
A geodatabase can contain numerous layers (or feature classes) and standalone tabular data, as well as maintain complex relationships between the layers. For example, multiple layers can share and depend upon a common geographic reference location (topology). Other layers have relationships defined with tabular data making the data more informative.

To confirm the validity of a geodatabase submittal:

1. Validate the geodatabase metadata, if available, per your metadata guidelines.
2. Obtain high-level geodatabase design documentation that lists the objects contained in the geodatabase.
3. Open ArcCatalog and select the geodatabase to view its contents.
4. For objects that have text or geometry (i.e. feature classes, annotation, tables, etc.), confirm that you can view their graphic representations and/or text attributes, and metadata (see Figure 13).
5. For all other objects identified in the design documentation (i.e. relationship classes, topologies, toolboxes, etc.), ensure that they are present in the geodatabase. Verification of these objects may require additional input from the geodatabase designer.

Note: ArcCatalog offers a validating function associated with certain objects in a geodatabase. This GIS data validation function confirms that data’s construction meets certain design specification and identifies any errors. This is not to be confused with the archivist’s task of validating the integrity of the geodatabase itself.
4. Validate the metadata
   a. Use the tool and approach identified during your planning to validate your geospatial metadata.

   b. Validate that the datasets have requisite descriptive and metadata fields populated. This will require a copy of GIS software, or a tool such as the USGS metadata verification tool.

      i. For individual shapefiles, verify that a projection is defined and geographic coordinate system is defined.

      ii. View the metadata .xml file and verify that all of the requisite metadata is provided.

         Note: Orthophoto datasets may be comprised of multiple image resolutions in a single dataset. In this case there should be a metadata file associated with each resolution that describes the capture mechanism, technology, and general geospatial descriptive metadata.

5. Run virus checking on all dataset files prior to submission to archival organization (submitter).

6. Package the geospatial data files with the agreed-to file packaging mechanism. Perform the agreed-to file integrity (bit level) validation (e.g., determine checksum or run BagIt ‘verifyvalid’) prior to transferring the files.

7. Transfer files with agreed-to transfer mechanism.
   This may entail electronically transmitting the files across the network, or copying the files to a removable storage device. If you utilize a program such as robocopy that produces a log file, you should review the reports and/or audit logs to ensure that the data transferred completely and accurately.

3.3. Archival Organization Geospatial Data Receipt Tasks

As the receiver of the geospatial datasets, the archival organization has several tasks to validate and verify the transferred datasets prior to proceeding with the actual archival ingestion and archival processing steps. Cross-organizational planning and coordination, definition of archivist processes, and selection of technologies and tools to receive and validate the datasets have been described as part of the planning process above. The following
provides a brief summary of the key tasks the dataset receiver completes in the transfer of geospatial datasets to the archival organization:

1. Establish the geospatial archives repository **architecture**. Technically, for the data transfer, the archival organization need only allocate repository space to receive and immediately validate the datasets. However, the receiving area will likely be a part of the greater archives repository, so the archival organization may want to consider planning its overall repository architecture as it is doing its initial data transfer storage planning. Based on OAIS guidelines, the archives should contain several areas (see Figure 14 below for an illustration):

   a. **Receiving area.** This is the area, likely part of your larger staging area, where received files will be stored (OAIS Submission Information Package\(^{42}\)). For preservation purposes, retain the original files, and copy them to a secure storage location, possibly in your preservation area. You may want to designate a quarantine area in the receiving area to hold files waiting to be checked for viruses and/or infected files, especially as files may not be checked for viruses immediately upon receipt.

   b. **Staging area / Work-in-Progress area.** This will be a working area to prepare the files for loading into the archival repository. This is also the area where derivative files may be created such as compressed files, .zip files for electronic distribution of datasets, PDFs of shapefiles, HTML files for geospatial metadata, etc.

   c. **Preservation area.** This will be the area where the “preservation” copy of the dataset will be stored. This copy serves as the “master” copy of the dataset (OAIS Archival Information Package). This file should not be compressed. It should not be regularly accessed. The “preservation master” may not be equivalent to the received dataset, as its format may have been transformed, or additional metadata associated with it, for example, metadata that reflects the archival processing steps.

   d. **Preservation master duplicates:** A duplicate of the preservation master should be maintained in case of loss or corruption to the preservation master. If resources allow, you should consider saving additional copies. These copies may be maintained through normal backup procedures (e.g., to tape), through copies to online accessible disk, or through remote storage solutions. Optimally, the duplicates should be stored on different physical servers, networks, etc. to minimize risk due to any single point of failure.

   e. **Access copy:** This is a copy that will be accessed by users (OAIS Dissemination Information Package). Compression techniques may be used to create the access copy to preserve data space in the accessible repository, and speed up download times. The access package may include additional access aids or derivative files, such as a human readable version of the geospatial metadata file, or geospatial PDF/GeoPDF version of the dataset, or a zip package containing all of the files in the dataset.

---


v1.0 12/1/2011
2. Establish repository storage space for dataset receipt processing, archival processing, and archival storage of geospatial datasets and their associated files. As a rule of thumb, consider that you will need five times the space of the geospatial data files to accommodate the:

   a. Originally received geospatial data files – which should be retained,
   b. “Work-in-progress” copy of the received data which will be processed and prepared for archiving,
   c. Preservation master plus at least one copy of the preservation master,
   d. Access copy.

3. Upon receipt of the files, place the files in your Receiving/Staging area.

4. Run a virus check on received files to validate they are clean.

5. Run the appropriate dataset receipt file integrity (bit-level) validation utility to confirm that the files were not corrupted during the transfer.
   e.g., run verifyvalid option of BagIt utility
   e.g., verify checksum on tar or zip file that contains all dataset files

6. Record an inventory of the received geospatial data (either at the geospatial dataset level, or more recommended, a manifest of all of the included files), date received, submitting organization, and submitter.

7. Create a “work-in-progress” copy of the received dataset, to preserve the originally received data from potential alterations. It is possible that your transfer validation tasks may alter, either intentionally or unintentionally, some of the dataset data.

8. Confirm that datasets and dataset files follow the agreed to naming conventions. Adjust names, if necessary, or return dataset back to submitter if names are indecipherable.

9. Functionally verify that the datasets can be accessed. Where feasible and reasonable, the archivist should use similar verification tools and processes as the GIS clearinghouse/submitter.

   This will require some sort of GIS development software. The following provides suggested approaches for functionally verifying a few types of geospatial datasets. The GeoMAPP team used Esri ArcCatalog to functionally verify the shapefiles, orthophotographs (both TIFF and MrSID formats), and geodatabases that were transferred to the state archival organizations.

   a. Vector Datasets (point, vector, polygon)
      A shapefile will generally represent a single layer, and will store point, vector or polygon features. For individual shapefiles, each shapefile should be opened and visually inspected. The shapefile graphic representation should render reasonably, whether it consists of points, lines or polygons. You should validate that the dataset properties appear to be properly populated, especially confirming the presence of the projection. You should also validate that the shapefile data attributes (view the table of attributes) are populated for the features represented in the dataset.
b. Raster Data / Orthoimagery

An orthoimagery collection is generally a collection of raster-based digital images. It may consist of orthophotos, which are geographically corrected (rectified) digital images resembling photographs. Digital elevation models are another source of geospatial raster data. File derivatives from Digital Elevation Models illustrate topographic characteristics, such as slope, aspect, hillshade, surface water flow direction, or simple elevation representation. Orthoimagery datasets, which can consist of hundreds or even thousands of individual image tiles, should be visually inspected by randomly selecting a sampling (e.g., 5%) of images. You should view the dataset file properties of the individual image to see that fields such as Data Type, Columns and Rows, Number of Bands, Cellsize, Format, Pixel Depth, Pyramids, and Extent values are populated. Orthoimage access can be enhanced for long-term preservation by creating a companion “index” shapefile that provides a summary and arrangement for the tiles that collectively comprise the orthoimage. If provided, the companion orthoimage “index” shapefile should also be inspected, to verify that it aligns with the orthoimage tiles.
Figure 16. View raster images and metadata (e.g., Esri ArcCatalog v9.3)

c. Geodatabases

A geodatabase can contain numerous layers (or feature classes) and standalone tabular data, as well as maintain complex relationships between the layers. For example, multiple layers can share and depend upon a common geographic reference location (topology). Other layers have relationships defined with tabular data making the data more informative.

To confirm the validity of a geodatabase submittal:

1. Validate the geodatabase metadata, if available, per your metadata guidelines.
2. Obtain high-level geodatabase design documentation that lists the objects contained in the geodatabase.
3. Open ArcCatalog and select the geodatabase to view its contents.
4. For objects that have text or geometry (i.e. feature classes, annotation, tables, etc.), confirm that you can view their graphic representations and/or text attributes, and metadata (see Figure 17).
5. For all other objects identified in the design documentation (i.e. relationship classes, topologies, toolboxes, etc.), ensure that they are present in the geodatabase. Verification of these objects may require additional input from the geodatabase designer.

Note: ArcCatalog offers a validating function associated with certain objects in a geodatabase. This GIS data validation function confirms that data’s construction meets certain design specification and identifies any errors. This is not to be confused with the archivist’s task of validating the integrity of the geodatabase itself.
10. Validate the metadata
   a. Use the tool and approach identified in during your planning to validate your geospatial metadata. 
      Note: If using ArcCatalog - you should disable / enable ArcCatalog’s metadata auto-update feature as desired.
   
   b. Validate that the datasets have requisite descriptive and metadata fields populated.
      This will require a copy of the geospatial development software.
      i. For individual shapefiles, verify that a Projection is defined and the Geographic Coordinate System is defined.

      ii. Verify the required metadata is provided by using a geospatial metadata viewing tool, or by directly inspecting the .xml file.
      Note: Orthophoto datasets may deliver multiple image resolutions in a single dataset. In this case there should be a metadata file associated with each resolution that describes the orthophoto capture mechanism, technology, and general geospatial descriptive metadata.

At the conclusion of the file transfer validation and verification, the archivist can proceed with the archival processing steps to ingest, arrange, describe, and create access tools for the newly received datasets.

4. Conclusions

Although Kentucky, Montana, North Carolina, and Utah chose differing formats of datasets to transfer, they shared many general practices that other GIS professionals and archivists can benefit from as they consider approaches for preserving their geospatial datasets. Cross-organizational planning and collaboration is key, along with engaging your information technology professionals as part of your cross-organizational team. Inventorying and appraising your available datasets will help define the overall parameters with regard to what you will be transferring. Defining and documenting how datasets will be transferred, including how they will be organized,


v1.0 12/1/2011
packaged and the transfer tools and mechanisms are important tactical details. And, finally, the archives performs its receipt validation and verification to confirm that the datasets have successfully arrived and have not been inadvertently altered through the transfer process. The geospatial dataset should be functionally verified, confirming that its graphical rendering, its underlying tabular data and its rich geospatial metadata record are all supplied intact.

Once the dataset has successfully completed its journey to the archival organization, the archivists will proceed with their archival processing steps to ensure the long term sustainability of the dataset, as well as its access to future GIS-interested users.
# Appendix B: Process Checklist

**General Project Planning Roadmap:**

<table>
<thead>
<tr>
<th>Step #</th>
<th>Status</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td></td>
<td>Prepare preliminary project plan</td>
</tr>
<tr>
<td>2.</td>
<td></td>
<td>Hold the kickoff meeting with representatives from the GIS and Archives communities.</td>
</tr>
<tr>
<td>3.</td>
<td></td>
<td>Inventory the geospatial data assets relative to the project scope to understand what geospatial data exists.</td>
</tr>
<tr>
<td>3. (opt)</td>
<td></td>
<td><em>Pilot Project Alternative:</em> Inventory geospatial datasets targeted for the data transfer pilot.</td>
</tr>
<tr>
<td>4.</td>
<td></td>
<td>Survey the geospatial community in the state to understand nature of geospatial data development, geographies covered, software and tools being used, file formats being saved, and frequency of dataset updates.</td>
</tr>
<tr>
<td>5.</td>
<td></td>
<td>Conduct a technical infrastructure survey for the GIS clearinghouse and the archival organization to understand the current state of the hardware, software, storage – especially storage capacity, and networks.</td>
</tr>
<tr>
<td>6.</td>
<td></td>
<td>Appraise geospatial datasets for archiving, and identify which datasets are to be transferred to the archives for long-term preservation.</td>
</tr>
<tr>
<td>6.(opt)</td>
<td></td>
<td><em>Pilot Project Alternative:</em> Select the geospatial datasets for an initial “pilot” project dataset transfer from the GIS clearinghouse to the archival organization.</td>
</tr>
<tr>
<td>7.0</td>
<td></td>
<td>Geospatial data transfer planning – collaboratively between the GIS and Archives professionals</td>
</tr>
<tr>
<td>7.0.1</td>
<td></td>
<td>Determine the file formats that will be accepted for archiving.</td>
</tr>
<tr>
<td>7.0.2</td>
<td></td>
<td>Create a file transfer flow chart that illustrates the transfer of datasets from GIS data creators to the archival organization.</td>
</tr>
<tr>
<td>7.1</td>
<td></td>
<td>Engage your Information Technology partners – in both the GIS clearinghouse and the archival organization.</td>
</tr>
<tr>
<td>7.1.1</td>
<td></td>
<td>Engage the Information Technology organization to start obtaining network access, firewall clearances, file system or network security access rights.</td>
</tr>
<tr>
<td>7.1.2</td>
<td></td>
<td>Engage the Information Technology organization for infrastructure planning, storage purchase, and software purchase, installation and configuration.</td>
</tr>
<tr>
<td>7.2</td>
<td></td>
<td>Establish geospatial metadata requirements policy (identify which fields must be filled in) Recommend: Reference FGDC CDSGM specification for required metadata fields. Recommend: The GIS and Archives professionals collaborate to ensure that rich geospatial metadata is provided that promotes the long term preservation of the geospatial datasets.</td>
</tr>
<tr>
<td>7.3</td>
<td></td>
<td>Determine file naming conventions for datasets and dataset files to be transferred Recommend: Establish self-describing naming conventions for datasets and dataset files.</td>
</tr>
<tr>
<td>7.4</td>
<td></td>
<td>Identify file organization structure (database, file system, and associated organization scheme) Recommend: Use the ISO or Ramona categories for the high level organization structure. Recommend: A commonly agreed to file organization for both data transfer and archival arrangement that facilitates the ingestion of the datasets into the archival repository.</td>
</tr>
<tr>
<td>7.5</td>
<td></td>
<td>Identify the dataset verification tools and methods for verifying the datasets and geospatial metadata.</td>
</tr>
<tr>
<td>7.6</td>
<td></td>
<td>Identify file transfer packaging mechanism and validation mechanism (e.g., BagIt with validation, zip or tar with checksum)</td>
</tr>
<tr>
<td>7.7</td>
<td></td>
<td>Identify file transfer size requirements</td>
</tr>
<tr>
<td>7.7.1</td>
<td></td>
<td>Define the storage space requirements for the submitter’s dataset preparation work area.</td>
</tr>
<tr>
<td>7.7.2</td>
<td></td>
<td>Define the storage space requirements for the archives’ dataset receipt staging work area.</td>
</tr>
<tr>
<td>7.7.2.1 (opt)</td>
<td></td>
<td>Define the general storage space architecture and storage space requirements for the archival repository.</td>
</tr>
<tr>
<td>7.8</td>
<td></td>
<td>Identify the file transfer mechanism (e.g., network-based transfer, removable disk)</td>
</tr>
<tr>
<td>7.8.1</td>
<td></td>
<td>Identify file transfer tools (e.g., file copy from one network-accessible file system to another, ftp, robocopy)</td>
</tr>
</tbody>
</table>
| 7.9 | | Identify archival submission policies.  
- Who is authorized to submit geospatial datasets?  
- Are datasets only accepted from the GIS clearinghouse?  
- Can municipalities and local governments submit datasets to the archival organization?  
- What is the trigger for when datasets will be transferred to the archival organization?  
- What is the frequency for when datasets will be transferred to the archival organization? |
<table>
<thead>
<tr>
<th>Step #</th>
<th>Status</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.12</td>
<td>Select geospatial datasets for transfer. For a “pilot” data transfer, select 10-50 datasets that are representative of the variety of data formats targeted for archiving, as well as can serve as an exemplar assortment of the variety of dataset categories (e.g., choose datasets from several framework layers).</td>
<td></td>
</tr>
<tr>
<td>8.1</td>
<td>Gather the datasets to be transferred</td>
<td></td>
</tr>
<tr>
<td>8.2</td>
<td>Functionally verify the datasets – using GIS tools</td>
<td></td>
</tr>
<tr>
<td>8.3</td>
<td>Validate the presence of appropriate / requisite geospatial metadata</td>
<td></td>
</tr>
<tr>
<td>8.3.1</td>
<td>Metadata clean-up, if necessary</td>
<td></td>
</tr>
<tr>
<td>8.4</td>
<td>File naming clean-up, if necessary</td>
<td></td>
</tr>
<tr>
<td>8.5</td>
<td>Organize the datasets to optimize the transfer and archival processing</td>
<td></td>
</tr>
<tr>
<td>8.6</td>
<td>Run virus checker to verify the files have no viruses</td>
<td></td>
</tr>
<tr>
<td>8.7</td>
<td>Package the datasets for transfer</td>
<td></td>
</tr>
<tr>
<td>8.7.1</td>
<td>Verify the integrity of the package prior to transfer</td>
<td></td>
</tr>
<tr>
<td>8.8</td>
<td>Identify the transfer medium (e.g., via network, via portable disk device)</td>
<td></td>
</tr>
<tr>
<td>8.8.1</td>
<td>If via network: identify transfer time of day - if I.T. has imposed transfer restrictions</td>
<td></td>
</tr>
<tr>
<td>8.9</td>
<td>Transfer the files</td>
<td></td>
</tr>
<tr>
<td>9.1</td>
<td>Archival organization dataset receipt processing</td>
<td></td>
</tr>
<tr>
<td>9.2</td>
<td>Unpack the files from the transfer package</td>
<td></td>
</tr>
<tr>
<td>9.3</td>
<td>Record the inventory of the received files</td>
<td></td>
</tr>
<tr>
<td>9.4</td>
<td>Run virus checker to verify the received files have no viruses</td>
<td></td>
</tr>
<tr>
<td>9.5</td>
<td>Make a ‘working’ copy of the originally received files, saving the original files in their original form</td>
<td></td>
</tr>
<tr>
<td>9.6</td>
<td>Validate the metadata</td>
<td></td>
</tr>
<tr>
<td>9.6.1</td>
<td>Validate that all required fields are populated</td>
<td></td>
</tr>
<tr>
<td>9.6.2</td>
<td>Verify that metadata field values seem reasonably complete</td>
<td></td>
</tr>
<tr>
<td>9.7</td>
<td>Functionally verify the geospatial dataset (using GIS tools):</td>
<td></td>
</tr>
<tr>
<td>9.7.0</td>
<td>Enable/Disable metadata auto-update feature</td>
<td></td>
</tr>
<tr>
<td>9.7.1</td>
<td>Visual rendering of the graphic</td>
<td></td>
</tr>
<tr>
<td>9.7.2</td>
<td>Visually inspect the dataset attributes</td>
<td></td>
</tr>
<tr>
<td>9.7.3</td>
<td>Visually inspect the metadata</td>
<td></td>
</tr>
</tbody>
</table>

**Technical Infrastructure Planning:**
Concurrent with the geospatial dataset planning and transfer activities, the technical infrastructure must also be assessed.

<table>
<thead>
<tr>
<th>Step #</th>
<th>Status</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Storage needs in the GIS clearinghouse to prepare and stage data for transfer</td>
<td></td>
</tr>
<tr>
<td>2.</td>
<td>Storage needs in the archives to receive, process, and store preservation and access copies</td>
<td></td>
</tr>
<tr>
<td>3.</td>
<td>GIS Software and tools the GIS submitter and the archives will require for accessing and validating the datasets</td>
<td></td>
</tr>
<tr>
<td>4.</td>
<td>Dataset packaging tools (e.g., BagIt, zip, tar)</td>
<td></td>
</tr>
<tr>
<td>5.</td>
<td>Dataset transfer mechanism</td>
<td></td>
</tr>
<tr>
<td>5.1</td>
<td>Network infrastructure configuration (e.g., firewalls, security on share drives, security on directories, etc.)</td>
<td></td>
</tr>
</tbody>
</table>

**Software Planning**
Also, new software and tools may need to be acquired and/or developed for either the GIS clearinghouse or the archival organization. The archival organization may want to investigate archival storage solutions such as DSpace or Fedora. As the geospatial metadata is defined based on a standard, tools may be developed to automatically extract metadata values to populate the archives’ archival metadata record. The archives may want to develop ingestion tools that enable submitters to contribute geospatial datasets to the archival organization.
Appendix C - Resources

2.1. Kick-off meeting

**RESOURCES:**
- GeoMAPP website Project Documents ([http://www.geomapp.net/documents.htm](http://www.geomapp.net/documents.htm)):
- Geoarchiving Project Plan Checklist (Appendix B)

2.1. Orientation and cross training

**RESOURCES:**
- Iowa Geographic Information Council. An Introduction to Geographic Information Systems Technology ([http://www.iowagic.org/resources/aboutGIS/AN%20INTRODUCTION%20TO%20GEOGRAPHIC%20INFORMATION%20SYSTEMS%20TECHNOLOGY.pdf](http://www.iowagic.org/resources/aboutGIS/AN%20INTRODUCTION%20TO%20GEOGRAPHIC%20INFORMATION%20SYSTEMS%20TECHNOLOGY.pdf/))
- GIS 101 Questions, Archiving 101 Questions (available in GeoMAPP Wiki)
- GeoMAPP website Geoarchiving Glossary ([http://www.geomapp.net/using.htm](http://www.geomapp.net/using.htm))
- Geoarchiving 101: How to engage GIS practitioners in the Archival Discussion (pp. 28-48) ([http://www.geomapp.net/docs/Butch_Alec_BPE_20090902_Final_2.pdf](http://www.geomapp.net/docs/Butch_Alec_BPE_20090902_Final_2.pdf))
- Iowa Geographic Information Council. An Introduction to Geographic Information Systems Technology ([http://www.iowagic.org/resources/aboutGIS/AN%20INTRODUCTION%20TO%20GEOGRAPHIC%20INFORMATION%20SYSTEMS%20TECHNOLOGY.pdf](http://www.iowagic.org/resources/aboutGIS/AN%20INTRODUCTION%20TO%20GEOGRAPHIC%20INFORMATION%20SYSTEMS%20TECHNOLOGY.pdf/))

2.2. Inventories, Assessments and Surveys

**Inventory**

**RESOURCES:**
- GeoMAPP website Project Documents ([http://www.geomapp.net/documents.htm](http://www.geomapp.net/documents.htm)):
  - Geoarchiving Self Assessment ([http://www.geomapp.net/docs/GeoMAPP_GeoArchiving_SelfAssessment_20100914.xls](http://www.geomapp.net/docs/GeoMAPP_GeoArchiving_SelfAssessment_20100914.xls))
- GIS Inventory ([http://www.gisinventory.net/](http://www.gisinventory.net/))
- In addition, numerous states have prepared state-oriented inventory web sites often managed by the state’s GIS clearinghouse.
- GIS Inventory Information Capture:
  - Name of the dataset
  - The name of the data creator of the steward
  - Date of dataset
  - Production date of dataset
  - Format of the dataset
  - Size of the dataset
  - Description of the data
  - Geographic scale of the data
  - Any important dataset attribute information (e.g., orthophotos: resolution)
  - ISO category that the dataset belongs to
  - How frequently the dataset is updated
  - If the data has been or is being continually archived
  - Where is the data being archived
  - Frequency of archival capture
  - How to access the current and/or the archived data.
  - Record retention schedule in place for the dataset?
2.3. Appraisal

RESOURCES:
- GeoMAPP website project documents: [http://www.geomapp.net/documents.htm](http://www.geomapp.net/documents.htm)
- Geospatial Packaging Whitepaper (Work-in-Progress)
- Kentucky State Agency Records Retention Schedule: [http://www.geomapp.net/docs/ky_gis_schedule.pdf](http://www.geomapp.net/docs/ky_gis_schedule.pdf)
- Maine Records Retention Schedule: [http://www.geomapp.net/docs/me_gis_schedule.pdf](http://www.geomapp.net/docs/me_gis_schedule.pdf)
- North Carolina Local Retention Schedule (Draft): [http://www.geomapp.net/docs/local_gis_retentionschedule_items_v1.pdf](http://www.geomapp.net/docs/local_gis_retentionschedule_items_v1.pdf)

2.4. Geospatial Data Transfer Planning

RESOURCES:
- GeoMAPP website Presentations: [http://www.geomapp.net/presentations.htm](http://www.geomapp.net/presentations.htm)
- Geoarchiving 101: How to engage GIS practitioners in the Archival Discussion (p. 31): [http://www.geomapp.net/docs/Butch_Alec_BPE_20090902_Final_2.pdf](http://www.geomapp.net/docs/Butch_Alec_BPE_20090902_Final_2.pdf)
- Footprints in the SAN: Transferring and Archiving GIS Data (p. 6): [http://www.geomapp.net/docs/NC_GeomAPP_BPE_Alec_Druscie_20090903_FINAL2.pdf](http://www.geomapp.net/docs/NC_GeomAPP_BPE_Alec_Druscie_20090903_FINAL2.pdf)
- Kentucky Collaborates in GeoMAPP Project (p. 12): [http://www.geomapp.net/docs/BPE_2009_KY.pdf](http://www.geomapp.net/docs/BPE_2009_KY.pdf)
- Process Flow Diagrams
    - Slide 40: North Carolina
    - Slide 41: Kentucky
2.4.2.1 Geospatial Metadata Requirements

**RESOURCES:**
- GeoMAPP website Project Documents (http://www.geomapp.net/documents.htm)
  - “FGDC and Dublin Core Metadata Comparison, March 2009” (http://www.geomapp.net/docs/MetadataComparison_200903.pdf)
  - Geospatial Archival Metadata Whitepaper (Work-in-Progress)

Geospatial Metadata Compliance Checking

**RESOURCES:**
- FGDC. “What metadata tools are available to assess compliance to the CSDGM?” Retrieved 1/31/2011 from: http://www.fgdc.gov/metadata/geospatial-metadata-tools#csdgmcompliance

2.4.3 Naming Conventions

**RESOURCES:**

2.4.4 Data Organization

**RESOURCES:**

2.4.6 Dataset Packaging Mechanism for Transfer and Post-transfer Validation

**RESOURCES:**
- BagIt download site via SourceForge: http://sourceforge.net/projects/loc-xferutils/
- Checksum tool via SourceForge: http://checksumtool.sourceforge.net/

2.4.9 Identify archival submission policies

**RESOURCES:**
- GeoMAPP website Project Documents (http://www.geomapp.net/documents.htm)
  - “Memorandum of Understanding.” (http://www.geomapp.net/docs/2009_AGRC_Archives_MOU.pdf)

2.4.10 Create a written document that describes the data transfer process

**RESOURCES:**
- GeoMAPP website Project Documents (http://www.geomapp.net/documents.htm)
  - Intrastate Data Transfer Design Outline (http://www.geomapp.net/docs/Intrastate_design_outline_20090515.pdf)
  - North Carolina Intrastate Data Transfer Design and Evaluation (http://www.geomapp.net/docs/NC_Intrastate_Geoarchives_Final_20090914.pdf)
• Utah Intrastate Data Transfer Design and Evaluation
  (http://www.geomapp.net/docs/Utah_Intrastate_Report.pdf)
  (http://www.geomapp.net/docs/GeoMAPP_InterimReport_Final.pdf)
Appendix D - Archives 101: GIS Questions about Archiving

These are common questions that GIS professionals have when first learning about Archives:

1. Who are the primary contacts at the state archives?
   a. Who are the primary contacts for GIS records?

2. Describe the how the Archives are organized (who handles Local, State and electronic records)

3. Describe the change in approach between traditional paper records management and electronic records

4. What is archiving? How is it different than doing a backup?

5. Why do we need to ‘archive’ our geospatial data? Can’t we just back it up?

6. What are the special risks associated with geospatial data?

7. Where should geospatial data be saved for long term preservation and access? (local retention at the agency level vs. retention at the state archives .... this will likely be different from state to state, and even from agency to agency within each state)
   a. Does my state preserve geospatial data?
   b. Whose responsibility is it now to preserve geospatial data?
   c. Whose responsibility should it be to preserve geospatial data?

8. Why should geospatial data be saved at the state archives?

9. Describe the state’s Public Records Law and how does it apply to digital geospatial data?

10. What is provenance? (Good opportunity to tie into FGDC Metadata – Process Steps)

11. What is appraisal?

12. What is a retention schedule?

13. Are there retention schedules in place for geospatial data in our state? If so, where do we find those?

14. Do archivists care about the metadata we (GISers) create? Is it necessary to transfer the metadata at the same time as we transfer the geospatial data?

15. What is administrative/preservation DESCRIPTIVE TECHNICAL metadata?
   a. What geospatial metadata is needed to facilitate long term preservation?
   b. What additional metadata is recorded to facilitate long term preservation?
      e.g., technical metadata that describes software used to create the dataset, checksum metadata to validate the integrity of the digital file, administrative metadata that describes any transformations performed on the dataset to newer software versions.
   c. What geospatial metadata is necessary to facilitate access going forward?
   d. What additional metadata is necessary to facilitate long term access?

16. What is OAIS?

17. Why do you need us to input all of this (FGDC) metadata? It’s a chore, it’s time consuming.
   What is the recommended metadata that needs to be populated for preservation and access purposes?
   Is there a (state) guideline on what geospatial metadata needs to be populated for preservation and access?

18. How will geospatial datasets be transferred to the state archives?

19. Who can submit geospatial datasets to the state archives?
20. What will you do with our datasets once they are transferred to the state archives?

21. What happens to the “complex” geospatial datasets as they are submitted to the state archives? Will all of the complex relationships be maintained?

22. How will people find and access geospatial data once it’s been transferred to the state archives?

23. Will there be ONLINE access to the geospatial datasets once they are transferred to the state archives? (people may just assume that it’s there ... because ‘everything’ is perceived to be online now)
   a. How will the datasets be accessed? (e.g., download zip files of the dataset “source” files)
   b. Will special GIS software be required to view and/or manipulate the GIS dataset?
   c. Are there any alternative derivative formats that allow access that don’t require special GIS software (e.g., geospatial PDF that’s accessible via Adobe Reader)

24. Does the state archives have some sort of ONLINE search interface to find information? (What’s a “catalog”)
   a. How can users search for and locate GIS datasets in our archives?

25. What’s a finding aid? How do people use finding aids?

26. How do finding aids relate to the geospatial dataset archives?

27. What’s a collection, record group, series?

28. How does the geospatial metadata get into the archives system?

29. How does metadata get used / leveraged / transformed as the geospatial datasets move into the state archives repository? (e.g., the title will pass through to be a title field, the originator will likely pass through to a creator or author field, keywords may be transformed to Library of Congress Subject Headings, place keywords may be transformed to Library of Congress Geography Subject Headings)

30. What is a (metadata) ‘crosswalk’?
   a. Is there a crosswalk that illustrates the mappings between various metadata dictionaries in our systems?
   b. Will a crosswalk be created that reflects mappings to the FGDC CSDGM metadata?

31. How much storage is available at the state archives to hold our geospatial datasets?

32. How are this storage and the systems paid for that will house and manage the geospatial archives?

33. What software (vendor, application, platform) is used to store and manage the geospatial archives?

34. What other types of electronic records are being preserved at the state archives?
Appendix E - GIS 101: Archivists Questions about GIS

These are common questions that archival professionals have when first learning about GIS:

1. What does GIS stand for?
2. What are some examples of geospatial datasets?
3. What is the composition of a geospatial dataset? (e.g., the file structure) (need to describe that there are actually several types of geospatial datasets)
4. Who uses geospatial data and for what purposes?
5. What is a GIS Clearinghouse? What is its history?
6. How do people access geospatial datasets in our state today?
7. How do people search for geospatial datasets today? (what are the key search attributes/metadata fields? it’s likely that these same attributes should be surfaced in the Archives’ geospatial search interface)
8. Who are the primary contacts at the GIS Clearinghouse?
9. Who can deposit geospatial datasets into the GIS Clearinghouse repository?
10. Does the GIS Clearinghouse staff perform any sort of processing on the GIS datasets prior to making it available?
11. What’s FGDC metadata? (provide a walkthrough of FGDC metadata example ... and illustrate the extensive metadata fields that are provided as part of the standard related to: the creator, the technical metadata, the lineage capturing the process steps, special access restrictions, etc.)
   a. Which metadata elements are important to identify the geospatial object?
   b. Which metadata elements are important to preserve to ensure future access to the geospatial object?
   c. Is there a (state) guideline on what metadata needs to be populated for geospatial description?
   d. What is the recommended metadata that needs to be populated for preservation and access purposes?
12. What happens if a geospatial file gets disconnected / separated from its metadata file? (e.g., ArcGIS will populate a metadata file with the subset of metadata that is encoded directly into the dataset file)
13. What’s a layer (file)? (geospatial data file representation)
14. What are these ISO Categories?
   a. What’s a framework layer? Why is it called “framework”?
   b. Where do I find a listing of all of the defined ISO Category layers?
   c. How can I see what layer a GIS dataset has been assigned to?
15. What’s a vector dataset? What are some examples in our state? How is it primarily used?
16. What’s a raster dataset? What are some examples in our state? How is it primarily used?
17. How many files make up “one” ‘file’ (geospatial object)
   e.g., shapefiles? What’s a shapefile? How many pieces parts are there?
   e.g., orthophotos? What’s an orthophoto? How many pieces parts are there?
18. What’s a map projection? Why is it important to the integrity and accessibility of the dataset?
19. What’s a coordinate system? Why is it important to the integrity and accessibility of the dataset?
20. What are our state’s plans for capturing orthophotos?

21. What descriptive information (metadata) is associated with all of these types of geospatial data? Are there similarities? Are there differences?

22. What’s a geospatial database / geodatabase?

23. What’s a geospatial project?

24. What’s a geospatial digitized map?

25. What software application(s) (vendor, version, platform) are used in our state to create geospatial datasets?

26. What is RAMONA GIS Inventory?
   Does our state participate?
   If so, who contributes to it in our state?
   Who uses it?

27. What is the status of our state’s development of the GIS Inventory? (http://gisinventory.net/)

28. How frequently are your geospatial datasets changing? being updated? being superseded?

29. Can you provide a brief demonstration of ArcGIS software? (Need to describe that it’s actually several software applications: ArcMap, ArcCatalog seeming to be the most commonly used)

30. Are you keeping older versions of the geospatial datasets as newer, updated versions become available?

31. How do I view a geospatial dataset?
   a. shapefile
   b. geodatabase
   c. orthophoto
   d. project
   e. digitized map

32. How do I view geospatial metadata?
   a. shapefile
   b. geodatabase
   c. orthophoto
   d. project
   e. digitized map

33. How do I create a PDF of a shapefile? (e.g., ArcMap file export to PDF)

34. How do I export geospatial metadata to a user-friendly representation? (e.g., ArcCatalog HTML export)

35. Are there special access restrictions or considerations for any of the geospatial datasets?