Geospatial Multistate Archive and Preservation Partnership (GeoMAPP)

Best Practices for Archival Processing for Geospatial Datasets

November 2, 2011
version: 1.0 (final)
Best Practices for Archival Processing for Geospatial Datasets

Table of Contents

1. Introduction ........................................................................................................................................1
   1.1 The Open Archival Information System (OAIS) Model .................................................................2

2. Storage Architecture .........................................................................................................................6
   2.1 Staging Repository ......................................................................................................................6
   2.2 Preservation Repository ............................................................................................................7
   2.3 Access Repository ...................................................................................................................7

3. Archival Processing ..........................................................................................................................8
   3.1.0 The Ingest Process: ...............................................................................................................10
   3.1.1 Contributor Prepares files ...................................................................................................10
   3.1.2 Contributor Creates/Verifies geospatial metadata ...............................................................11
   3.1.3 Contributor Runs Virus Check on Files ...............................................................................11
   3.1.4 Contributor Packages Files for Transfer .............................................................................11
   3.1.5 Contributor Transfer Files ..................................................................................................12
   3.1.6 Contributor Acknowledgement Received .............................................................................12
   3.2.0. Quality Assurance ...........................................................................................................12
   3.2.1 Archival Processor makes a security copy of the Submission Information Package ............13
   3.2.2 Archival Processor runs virus check on files .....................................................................13
   3.2.3 Archival Processor validates that the data files arrived intact and are consistent with the data files the contributor submitted .................................................................14
   3.2.4 Archival Processor runs virus check on files .....................................................................13
   3.2.5 Archival Processor unpacks the files from the package .......................................................14
   3.2.6 Archival Processor validates dataset composition -- all files included in each dataset (format specific), dataset can be rendered ..........................................................16
   3.2.7 Archival Processor validates the geospatial metadata meets the archival organization’s preservation guidelines and/or policies .........................................................20
   3.2.8 Archival Processor opens the dataset using appropriate GIS tool to verify that data is provided with the dataset .................................................................................22
   3.2.9 Archival Processor sends acknowledgement to geospatial data contributor .....................24
   3.3.0 Generate Archival Information Package (AIP) ....................................................................24
   3.3.1 Archival Processor creates unique identifier for the new digital object being preserved ....26
   3.3.2 Archival Processor saves a copy of the original geospatial metadata (file) ..........................27
   3.3.3 Archival Processor Identifies the Preservation Master file(s) and format. Optionally, reformats/transforms received digital assets to promote long-term preservation ..........................27
   3.3.4 Optional: Archival Processor updates the geospatial metadata file to populate missing metadata fields - using tools such as Esri’s ArcCatalog to auto-populate these fields .............................29
   3.3.5 Optional: Archival Processor update the geospatial metadata file to document the transfer of ownership to the Repository, and the Repository’s processing step(s) in preparing the dataset for preservation, and add ISO Theme Keywords ..................................................................................30
   3.3.6 Archival Processor establishes fixity value(s) or mechanism for the digital asset(s) to be maintained in the AIP .................................................................................................33
   3.3.7 Archival Processor creates the Archival Metadata record for the digital asset. Populates descriptive, technical, and administrative metadata for the digital record ........................................34
Best Practices for Archival Processing for Geospatial Datasets

3.3.8 Archival Processor creates the AIP and add it to the preservation repository ........................................ 37
3.3.9 Archival Processor Creates Security Copy(ies) of the Preservation Master record .................................. 38
3.4.0 Generate Description Information Package (DIP) .................................................................................... 39
3.4.1 Archival Processor Creates Access/Use Copy .......................................................................................... 39
3.4.2 Archival Processor creates Access Aid Derivatives ............................................................................... 40
3.4.3 Archival Processor creates Repository Catalog Entry and Finding Aid .................................................. 42
3.4.4 Archival Processor adds records to the Repository Digital Collection .................................................. 45
3.4.5 Archival Processor registers geospatial datasets with National GIS Inventory Systems ....................... 50
3.4.6 Archival Processor verifies catalog, Digital Collection, and Inventory access ..................................... 51
3.4.7 Archival Processor Updates Archival Metadata Record with “related” resources .................................... 52

4. Conclusions ................................................................................................................................................. 52

APPENDIX A: Quality Assurance Process Workflow Checklist ................................................................. A-1
APPENDIX B: Process steps to create Archival Information Package (AIP) .................................................. A-2
APPENDIX C: Process steps to create Dissemination Information Package (DIP) ........................................ A-3
APPENDIX D: Summary of Key Questions for Repository ........................................................................... A-4
APPENDIX E: Esri ArcCatalog - Auto-update Metadata Fields ................................................................. A-7
APPENDIX F: Digital Repository Platforms ................................................................................................. A-10
1. Introduction

The GeoMAPP project’s primary goal is to “ensure the long term preservation of archived geospatial content.” In order to preserve geospatial content, the GeoMAPP partners explored techniques and technologies to process geospatial data for preservation and access after they are transferred to the archives or preservation organization. Based on the combined experience of the GeoMAPP partners, this paper presents a suggested processing workflow for archival organizations processing geospatial datasets. While the general workflow is based on the Reference Model for an Open Archival Information System (OAIS), it includes special considerations for processing geospatial datasets. The first part of the paper offers a brief introduction to OAIS, as it provides the general framework and basis for an archival processing workflow. In the second part of the paper, a high level storage architecture is presented that supports the multi-phase archival processing workflow. The third part of the paper is the major focus, and offers a geospatial archival processing workflow based on the OAIS model, and highlights special considerations and activities when processing geospatial assets for preservation and access. Representative archival processing details from the GeoMAPP partners are also included to illustrate how many of these tasks were approached or accomplished.

A generalized version of the life cycle of a geospatial dataset from its creation to its final destination in the archival repository for long term preservation is illustrated in Figure 1 below. For many states, the geospatial archival repository will reside in the state archives but it may, alternatively, reside in other institutions such as a state library or a university. Many states have established, or are establishing, a centralized Geographic Information System (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 older datasets. The state archival repository (hereafter, simply referred to as the Repository) typically has the mandate and mission to preserve state historical records and provides a logical location to house, preserve and provide access to these historic geospatial datasets. In the GeoMAPP partner states, the Repositories receive datasets from their respective clearinghouse organizations, as illustrated below. Additionally, the Repository may accept submissions directly from state agency or local government geospatial data producers. In either case, the contributor is responsible for preparing and transferring the datasets to the Repository. Once received by the Repository, the archival organization has a series of tasks to perform to verify, preserve, and provide access to the geospatial datasets, as highlighted by the red box in Figure 1, and will be described in more detail in Part 3.

2 In Kentucky, North Carolina and Utah their state archives hold the geospatial preservation responsibility, while in Montana, the Montana State Library holds the geospatial preservation responsibility under their mandate to preserve state publications.
3 Kentucky, North Carolina and Utah have established GIS clearinghouses that are separate from their state archival organizations. The Montana State Library hosts the GIS clearinghouse and also is responsible for geospatial data preservation for the state of Montana.
1.1 The Open Archival Information System (OAIS) Model

The Reference Model for an Open Archival Information System (OAIS)\(^5\), illustrated in Figure 2, depicts a conceptual model for the preservation environment. The model includes the major roles involved in the use and management of archival information, the major model components, and a high level process flow for how information is ingested, managed and accessed through an archival management system. The OAIS Blue Book provides a descriptive overview of the model and its constituent parts. The GeoMAPP team will not attempt to exhaustively explain the model. Rather, the following provides a general overview of the model with some key excerpts from the \textit{OAIS Blue Book} to highlight key aspects of the model. Utah’s APPX-based Archives Enterprise Manager (AXAEM) system model corresponds to the OAIS model, as is the North Carolina State Archives’ (NCSA) and the Kentucky Department for Libraries and Archives’ (KDLA) general approach to processing and managing their archival collections.

The model depicts three primary roles that interact with the OAIS archive:\(^6\)

- **Producer** - contributes the information (files, digital data objects) to be preserved.
- **Management** - sets the overall OAIS policy and manages those policies within the OAIS environment.
- **Consumer** - interacts with OAIS services to search, access, and acquire preserved information of interest.


\(^6\) Ibid. p. 2-2.

v1.0 11/2/2011
**OAIS Information Package:** The OAIS Information Package is the conceptual model for the container of information that is being preserved as well as its associated descriptive and preservation information for the object that is being preserved.7

The Information Package is a conceptual container of two types of information:8

1. **Content Information** - the information which is the original target of preservation
   Consists of:
   a. **Content Data Object** - the Physical Object or Digital Object being preserved.
   b. **Representation Information** - information provided that makes the Content Data Object understandable to the Designated Community.

2. **Preservation Description Information** (PDI) - information that applies to the Content Information, and is needed to preserve the Content Information as well as understand the environment in which the Content Information was created. The Preservation Description Information is divided into four subcategories:9
   a. **Provenance** - describes the source of the Content Information, who has had custody of it since its origination, and its history (including processing history).
   b. **Context** - describes the purpose of the information and how it relates to other information outside the Information Package. For example, it would describe why the Content Information was produced, and it may include a description of how it relates to another Content Information object that is available.
   c. **Reference** - provides one or more identifiers, or system of identifiers, by which the Content Information may be uniquely identified.
   d. **Fixity** - provides a wrapper or protective shield that protects the Content Information from undocumented alteration. For example, it may involve a checksum of the Content Information of a digital Information Package.

**Packaging Information:** The OAIS model defines three variants of the Information Package that represent the evolution / alterations of the information package as it progresses through its life cycle from ingestion to long-term preservation in the repository, as well as consideration for a variant of the information that users will access. The packaging information represents “[t]he information that is used to bind and identify the components of an Information Package.” As an example, “it may be the ISO 9660 volume and directory information used on a CD-ROM to provide the content of several files containing Content Information and Preservation Description Information.”10

The Content Information and PDI are viewed as being encapsulated and identifiable by the Packaging Information.11

---

9 Ibid. p. 2-6.
10 Ibid. Glossary.
11 Ibid. p. 2-5.
The OAIS model identifies three Information Packages variants to distinguish between the Information Package that is submitted, the Information Package that is archived (preserved), and the Information Package that is disseminated (used for access). Refer to Figure 2 to see how these flow and are transformed through the system.

1. **Submission Information Package (SIP)**
   “The package that is sent to an OAIS by a Producer. Its form and detailed content are typically negotiated between the Producer and the OAIS. Most SIPs have some Content Information and some Preservation Description Information (PDI), but it may require several SIPs to provide a complete set of Content Information and associated PDI to form an AIP. A single SIP may contain information that is to be included in several AIPs. The Packaging Information will always be present in some form.”

2. **Archival Information Package (AIP)**
   “Within the OAIS one or more SIPs are transformed into one or more Archival Information Packages for preservation. The AIP has a complete set of Preservation Description Information for the associated Content Information.” The conceptual basis for the Preservation Master.

3. **Dissemination Information Package (DIP)**
   “In response to a request, the OAIS provides all or part of an AIP to a Consumer in the form of a Dissemination Package. The DIP may also include collections of AIPS, and it may or may not have a complete PDI. The Packaging Information will necessarily be present in some form so that the Consumer can clearly distinguish the information that was requested.” The conceptual basis for the Access or Use copy.

**OAIS Functional Entities:** The OAIS reference model identifies six high-level functional entities for preservation systems. Refer to Figure 2 to see the relationships between the functional entities.

1. **Ingest** digital materials (objects, files, etc.) - accept SIPs from Producers, and prepare the contents for storage and management within the archive. Ingest functions include:
   - receiving SIPs
   - performing quality assurance on the digital files submitted in the SIPs
   - generating an AIP which complies with the archive’s data formatting and documentation standards
   - extracting Descriptive Information from the AIPs for inclusion in the archive database
   - coordinating updates to Archival Storage and Data Management

   There are two general models for transferring data to the archive:

   - **push model:** content contributor deposits digital objects into the preservation system. There are varying level of sophistication for how users can deposit materials into the system ranging from a simple file copy to a common shared drive, or a more sophisticated graphical user interface and sequence of forms or screens the user completes to submit his materials.

   - **pull model:** preservation organization pulls digital objects into the preservation system. There are a variety of alternatives the archives may choose to pull information ranging from manually selecting items and copying them in, or developing automated harvesting applications that will crawl the organization’s Intranet or external Internet in search of materials to ingest into the archives.

---

12 Ibid. p. 2-7.
13 Ibid.
14 Ibid.
15 Ibid. p. 4-1.
2. Archival storage of digital materials (objects, files, etc.) - provides the services and functions for the storage, maintenance and retrieval of AIPs. Archival storage functions include:

- processing SIPS to produce AIPS from the Ingest process and adding the AIP digital artifacts to permanent storage, and possibly storing the SIP
- managing the storage hierarchy
- refreshing the media on which archive holdings are stored
- performing routine and special error checking
- providing disaster recovery capabilities
- providing AIPs to Access to fulfill orders

In considering the requirements for storing digital archival materials, with the potential for mechanical, software, and human failure, multiple copies of the preservation master should be retained (the old adage - lots of copies keep stuff safe), in case the Preservation Master becomes lost, damaged, or corrupted. Optimally, these copies should be maintained on different storage repositories, and even geographically distributed to minimize the risk of any single point of failure. To minimize the potential wear and tear on the Preservation Master, a separate copy designated as the “Access Copy” will often be derived from the Preservation Master. The Access Copy may be a physical duplicate of the Preservation Master, or it may be a derivative form. For example, a repository may choose to make a compressed image Access Copy from the uncompressed Preservation Master to speed online access, and minimize disk usage in the access repository. An organization will likely incorporate standard IT backup procedures for the Access Copies, but in theory, if an Access Copy becomes lost or corrupted, it should be able to be reproduced from the Preservation Master. For patrons who request the full, uncompressed version of the file, the archival organization can make an “on-demand” copy from the Preservation Master, or one of its duplicates. This uncompressed copy may become a part of the Access system to be able to accommodate future requests, or it may be discarded and reproduced at the next request.

3. Data management - often implemented through a database management system, provides the services and functions for populating, maintaining, and accessing both Descriptive Information, which identifies and documents archive holdings, and administrative metadata used to manage the archived materials. Data management functions include:

- administering the archive database functions (maintaining schema and view definitions, referential integrity).
- performing database updates (loading new descriptive information or archive administrative data).
- performing queries on the data management data to generate result sets (service search requests).
- producing reports from these results sets.

4. Administration - provides the services and functions for the overall operations of the archive system. Administration functions include:

- soliciting and negotiating submission agreements with the Producers.
- auditing submissions to ensure that they meet the archives’ standards and policies.
- maintaining configuration management of system hardware and software.
- system engineering functions to inventory, report on, and migrate/update the contents of the archive.
- responsible for establishing and maintaining archive standards and policies, providing customer support, and activating stored request.

5. Preservation planning - provides the services and functions for monitoring the environment of the OAIS and providing recommendations to ensure that the information stored in the OAIS remains accessible to the Designated User Community over the long term, even if the original computing environment becomes obsolete. Preservation planning functions include:

- migrate content.
- develop recommendations for archive standards and policies.
• designs IP templates for submission.
• develops migration plans.
• develops software prototypes and test plans to enable implementation of Administration migration goals.

6. **Access** of digital objects - provides the services and functions that support Consumers in determining the existence, description, location and availability of information stored in the OAIS, and allowing Consumers to request and receive Information products. Access functions include:
- receive requests.
- apply controls to limit access to specially protected information.
- coordinate execution of requests to successful completion.
- generate responses (DIPs, result sets, reports).
- deliver responses to Consumers.

### 2. Storage Architecture

Ideally, the Repository should establish three different storage repositories (or logical repositories) to support preliminary processing, preservation, and access to preserved digital materials. The repositories depicted in Figure 3 may be housed in a single or multiple physical repositories, and use the same or differing technologies (e.g. a Repository may use a digital preservation solution for the preservation repository and Esri’s GeoPortal solution for the access repository). Additionally, as the preservation repository does not have the same interactive use requirements as the access repository, it could be implemented with slower (or offline) storage alternatives.

![Conceptual Archival System Storage Architecture](image)

Figure 3. Conceptual Archival System Storage Architecture

Ideally, the staging repository will be located on a physically separate machine from the preservation and access repositories. However, if you do not have technical infrastructure to support three physical repository systems, the GeoMAPP team suggests establishing separate folders within a single repository (staging, preservation, access), and establishing directory and file level access controls to restrict access based on an individual’s role within the archival processing process. For example, you may grant read/write access to the “staging” repository (or folder) to contributors and archival processing staff, while restricting read/write access to the preservation and access repositories (folders) to only the archival processors and relevant management staff. Separating the staging area from the preservation and access repositories also minimizes the threat of miscreant software infecting your valuable preservation and access digital assets.

#### 2.1. Staging Repository

As datasets (Submission Information Package) are submitted / ingested into the Repository they will be deposited into the archival Staging repository. This should be the only repository that allows write access from users outside of the archival organization. The Staging repository will also be the “working” area for the archival processor to process and prepare the received datasets and archival metadata for deposit into the Preservation repository for long-term preservation (Archival Information Package), and create the necessary access derivatives for end user access (Dissemination Information Package). Ideally, the staging area will also support some sort of designated quarantine space to hold items that have not been processed and/or have been identified as infected or at risk.

v1.0 11/2/2011
2.2. Preservation Repository
The preservation repository holds the Preservation Master and archival metadata for the preserved datasets. In order to ensure ongoing data authenticity and reliability, security on the Preservation repository should be tightly controlled. Ideally, access (both read and write) to this repository will be limited to a small number of personnel with archival management responsibility. Optimally, any changes to the Preservation Master files should be tracked to maintain provenance and chain of custody. The Repository should also set up a mechanism to create at least one “security copy” for each dataset stored in the Preservation repository to protect against the unexpected or inadvertent loss of the Preservation Master. There should be minimal access to the Preservation Master, as access often introduces the opportunity to damage the file. The preservation system needs some sort of service or mechanism to establish and track the fixity, and verify the integrity of the preservation master and its security copy(ies). Ideally, the preservation repository will also include services to perform periodic integrity checks, and automatically restore or recover Preservation Masters (or their security copy(ies)) in the event a file becomes corrupted.

2.3. Access Repository
The Access repository holds the digital assets made available to end users, and thus requires read-access from outside the Repository organization. Employees processing the collection will require write access to the access repository in order to load the access copies and access derivatives of the datasets. The Access Copy does not necessarily need to be the same digital artifact as the Preservation Master. For example, in the interests of optimizing storage and data transfer resources, a compressed version of the Preservation Master may be used as an Access Copy (e.g. the Preservation Master for an orthophoto may be the uncompressed TIFF file, while the derived compressed MrSID format is used for access). Users requiring the uncompressed datasets could receive it on a request-by-request basis. There really is no need to retain a “security copy” of the access version of the datasets as they should be able to be reconstituted from the Preservation Master. However, lost or damaged access copies are more likely to be restored through the IT department’s file recovery processes than be regenerating the access copy within the archival system.

More sophisticated digital preservation software environments may not even create a separate physical access repository, but may instead, through programmatic interfaces, dynamically construct the Access image of a dataset based on information held in the Data Manager and the Preservation Master in response to a user request. In this case, it is technically not the end user that is accessing Preservation Master but the archival repository program that is accessing the Preservation Master on behalf of the end user, so the Preservation Repository security continues to be tightly controlled. For this “on-demand” model, however, there may be some lag-time perceived by the end user as the system produces the access copy. Therefore, the end user response-time should be assessed to ensure that the access copy is being rendered within a reasonable amount of time, or that the end user interface properly sets the expectation that there may be some wait time while the system generates the access copy. A user interface widget such as a progress status bar provides the end user with a helpful visual indicator of the relative time needed to wait for the request to be fulfilled.
3. Archival Processing

The Geospatial Archival Processing Workflow

The above diagram offers a generalized archival processing workflow, based on the OAIS model, to ingest, archive, and make accessible geospatial data from archival organizations. Specific details of how the above steps will be implemented will be largely affected by:

a. The **technical architecture** the Repository employs for the preservation of and access to geospatial information -- including the computing and storage hardware, the network, the software tools and applications, and the data management mechanism.

b. The **geospatial formats** that archives ingest and/or choose to use as the basis for preservation and/or access.\(^{16}\)

As part of the archival processing planning, the organization should determine its strategy regarding the formats that will be used for preservation and access, and any expected format transformations that will be necessary as part of the archival processing process. For archiving geospatial data, this should be part of the archives’ larger

---
\(^{16}\) Refer to the GeoMAPP Geospatial Data File Formats Reference Guide for a listing of several common geospatial data file formats. 
http://www.geomapp.net/docs/GeoMAPP_Geospatial_data_file_formats_FINAL_20110701.xls

v1.0 11/2/2011
media type preservation plan that documents how formats such as documents, audio, image, and video will be stored for preservation, and what formats will be used for end user access, which may not necessarily be the same as the preservation format.\(^\text{17}\)

Repositories could develop a Data Type Preservation Plan (inspired by earlier media type preservation plans), as illustrated in the following table, and then develop specific archival processing transformation details for each of the file formats:

<table>
<thead>
<tr>
<th>Data Type</th>
<th>Data File Formats</th>
<th>Preservation Format(s)</th>
<th>Access Format(s)</th>
<th>Transformation Tool</th>
</tr>
</thead>
<tbody>
<tr>
<td>Geospatial Vector</td>
<td>Esri Shapefile consisting of multiple files: .\shp* (shapefile) .\shx* (shape index) .\dbf* (dBASE database) .\sbn, \sbx .\prj* (project) .\shp.xml* (geo metadata) * required, * recommended</td>
<td>shapefile .\csv (text ver)(^\text{19})</td>
<td>GeoPDF GeoSpatial PDF zip of shapefile collection</td>
<td>Esri ArcCatalog TerraGo</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>zip</td>
</tr>
<tr>
<td>Geospatial Vector</td>
<td>Esri ArcInfo Coverage (hierarchical multi-file structure)</td>
<td>shapefile? ? what is lost?</td>
<td>GeoPDF GeoSpatial PDF zip of shapefile collection</td>
<td>Esri ArcCatalog TerraGo</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>zip</td>
</tr>
<tr>
<td>Geospatial Vector</td>
<td>Esri ArcInfo Export Format for Coverage (.\e00) [single ASCII file]</td>
<td>shapefile? ? what is lost?</td>
<td>GeoPDF GeoSpatial PDF zip of shapefile collection</td>
<td>Esri ArcCatalog TerraGo</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>zip</td>
</tr>
<tr>
<td>Geospatial Metadata</td>
<td>(should be packaged with related dataset)</td>
<td>XML</td>
<td>XML (text) XML HTML (styled)</td>
<td>Esri ArcCatalog USGS Metadata Validator (Outline view)</td>
</tr>
<tr>
<td>Orthophotos</td>
<td>.\tif (image file) .\tfw (world file) .\xml (geo metadata)</td>
<td>.\tif (image file) .\tfw (world file) .\xml (geo metadata)</td>
<td>.\sid (MrSID) .\sdw (MrSID world) .\aux (MrSID projection) XML, HTML</td>
<td>LizardTech GeoExpress</td>
</tr>
<tr>
<td>Geospatial Raster</td>
<td>Esri ArcInfo GRID Esri ArcInfo ASCII GRID grid saved in Esri .\e00</td>
<td>Esri ArcInfo ASCII GRID</td>
<td>zip</td>
<td>Esri ArcCatalog Conversion Toolbox</td>
</tr>
</tbody>
</table>


\(^\text{18}\) Refer to Archivematica’s wiki for example of a Media type preservation plan that addresses several common digital media types. Retrieved 10/27/2011 from: http://archivematica.org/wiki/index.php?title=Media_type_preservation_plans

\(^\text{19}\) You may want to create a simple text, comma separated values (.\csv) version of the .\dbf file to preserve the data in a simple text file format.
3.1.0 The Ingest Process:
The Ingest Process encompasses the means and/or mechanism for a geospatial content contributor to submit datasets to the archives for preservation.

There is a wide spectrum of technical approaches for ingesting files ranging from a simple manual copy of files to a commonly accessible shared network drive or a portable disk device, to a sophisticated web-based ingest application that presents a content submission form to the contributor, a means to transfer and upload files (either individually or as a package) to the archival organization, and also automatically perform a virus check on the transferred files. The ingest application may also include other advanced ingest processing capabilities such as verifying the presence of required geospatial metadata elements in the geospatial metadata file, packaging individual files for transfer, identifying and approving the file formats being submitted against a list or acceptable formats, or verifying that the submitted dataset file collections are “complete” (e.g. logic could be developed to ensure that the shapefile dataset folder includes its minimum required files).

However, regardless of the sophistication of the ingest process, there are a series of steps that, at a minimum, should be completed.

3.1.1 Contributor Prepares files
Archival best practices recommend that archived materials are organized in the fashion that reflects the originator’s use of the materials. However, as geospatial datasets can be contributed from numerous agencies, institutions, and organizations, the GeoMAPP team recommends that a common naming convention and organizational structure be established to enable consistency that will ultimately facilitate the accessibility of the geospatial datasets. Fortunately, there is already a categorization scheme for geospatial datasets established as well as GIS clearinghouse and inventory systems established that can inform the archival arrangement. In addition, as datasets may be individually downloaded, we recommend a descriptive naming convention be applied to the dataset files in order to further aid access. More details on the suggested naming convention and organizational approaches can be found in the GeoMAPP “Best Practices for Geospatial Data Transfer for Digital Preservation” document.

There are a variety of strategies for identifying the granularity of the submitted information. The information, hereafter referred to as the “unit” can be a single file, a collection of files that comprise an individual dataset, a collection of files that are logically related in some fashion, a (geo)database that comprises several individual datasets, or a batch transfer. The archival process and tools are essentially agnostic with regard to the unit that is selected. However, the unit may ultimately serve as the basis for the Archival Information Package (AIP), and possibly be visible through the access interface as the collection and/or record series archival grouping level.

Archival provenance advises that the Repository preserve the organizational structure of the submitted materials. However, depending on the contributor’s data organization and packaging, the archival processor will need to determine if the AIP will reflect the composition and structure of the submitted unit. As an example, a Repository may receive a quarterly transmission of all of the vector datasets superseded during the prior calendar quarter. A question arises - is the submission “unit” the individual datasets in the submission, or the entire collection of datasets submitted? Ideally the contributor will have organized the datasets into their appropriate GIS categories (e.g. data categories or layers for vector datasets) or series. The Repository could create a single AIP that contains all of the datasets - as they were submitted, or create a separate AIP for each individual dataset. Alternatively, a Repository may receive a transmission of a collection of datasets and associated project files for a geospatial project. In this case, as all of the files are logically related, the ingested package should serve as the basis for the AIP, and may be stored in a project series within the agency or municipality’s digital archival collection responsible for that project.

---


v1.0 11/2/2011
3.1.2 Contributor Creates/Verifies geospatial metadata

Aside from file format, metadata is one of the most important contributors to the long-term preservation, access, and utility of preserved geospatial datasets. The Federal Geographic Data Committee’s (FGDC) Content Standard for Geospatial Metadata (CSDGM) and the ISO North American Profile (19115) metadata standards include descriptive, technical, provenance, and administrative metadata that will support the long-term preservation of a geospatial dataset. The metadata may be embedded in the geospatial file, such as with the feature classes in an Esri geodatabase, or more ideally for preservation purposes, in a separate xml (text) file as is seen with the Esri shapefile. The state geographic governing bodies or the Repository may establish policies regarding requisite metadata elements. 22

The content contributor is often responsible for the creation of the geospatial metadata. Regardless if the contributor is the content creator or an intermediary, such as a GIS clearinghouse, the content contributor is responsible for verifying the completeness of the metadata (file) that accompanies the geospatial dataset intended for preservation. Each Repository will need to establish its own policies regarding required metadata, and how geospatial datasets will be managed that do not conform with the minimum metadata guidelines. In the absence of any guidance, the contributor can defer to the FGDC guidelines identifying the mandatory metadata fields.23

3.1.3 Contributor Runs Virus Check on Files

To ensure that the files are clean and free from viruses, a virus check should be run on the files to be submitted. In less sophisticated content submission approaches, the content contributor may need to manually run a virus checking utility on the files she is submitting. In more sophisticated ingestion applications, the ingest program may be able to run the virus check on behalf of the contributor. If the program finds an infected file, it should notify the contributor (and archival processor) and require the contributor to resubmit a new (clean) version of the infected file. Given the lag in virus software’s ability to address the latest vulnerabilities, files should be quarantined for a time, for example 30 days, to give the virus software an opportunity to catch up with the most recent threats that might be manifested in the ingested files.

3.1.4 Contributor Packages Files for Transfer

As individual geospatial datasets are often comprised of multiple files, the GeoMAPP team recommends that some sort of packaging utility be used to ensure that all of the files are transferred as a single unit. Any file ingest/submission program the Repository offers will, likewise, be required to accommodate the submission of multiple files as a single submission, and to maintain the organizational structure of the submitted files. Generally, vector datasets are organized based on the layer categories. For categorical organization examples, please refer to the GeoMAPP Geospatial Data Transfer document. If an interactive, forms-based submission interface for individual datasets is employed, the contributor should be prompted for the necessary categorization details such as the series name, layer name, county, year, etc. (the agreed-to organizational hierarchy should be reflected in the prompts presented to the contributor when submitting datasets).

Common packaging utilities include zip and tar. However, these tools do not provide an intrinsic mechanism to validate that a package, when transferred, arrives unaltered. It is imperative that the Repository have a method and tools to validate that the received digital assets have not been altered through the file transfer process. 24


24 Tools that generate hash values such as md5 hash generator and the Library of Congress BagIt tool can be used to validate the files’ contents post-transfer. BagIt also has the benefits of producing a manifest of the bag and an option to verify the validity of the bag’s contents. The hash values could also potentially be used within the preservation system as a basis for ensuring the ongoing integrity and authenticity of the digital assets. Refer to the GeoMAPP Geospatial Data Transfer white paper for more information. To be published by Dec 2011. Check: http://www.geomapp.net/publications_categories.htm
3.1.5 Contributor Transfer Files

The final step will be the actual transfer of the files from the contributor to the archival organization. As noted above in the storage architecture section, it is recommended that the files be transferred to a dedicated staging area (refer to diagram on page 5) where the archival processors can perform their file transfer verification tasks, and prepare the files for the archival and access repositories. There is a large variety of options available to transfer files including network-based transfers (e.g. manually ftp'ing files, copying files to a commonly accessible network share drive, or a specially developed upload program that’s built into the ingest application), or non-network-based transfers (e.g. copying files to a portable disk device that will be delivered to the archives). Acceptable mechanisms should be identified and communicated to data contributors, especially as due to network constraints, there may be limits to the size of files that may be transferred or time of day files may be transferred via the network. More details on the planning and selection of the file transfer mechanism can be found in the GeoMAPP Geospatial Data Transfer document.

Administrative metadata that should be recorded as part of the ingest process includes:

- Contributor contact information who submitted the file (Name, Agency/Institution, Address, phone, email)
- Description of data files being ingested
- Reason data files being submitted to archives
- Date files (digital assets) ingested (optional: time)
- Number of (data) files ingested.
- Names of the (data) files ingested.
- Size of the collection of files ingested.
- Validity value for the data files (e.g. hash value)
- Format (high level format e.g. documents, photos, audio, video, GIS, GIS vector dataset, GIS raster)
- Access restrictions (default: public record, no restrictions)
- Submission Agreement identification (if one is established)

3.1.6 Contributor Acknowledgement Received

The contributor should receive an acknowledgement from the archival organization whether the dataset he submitted was acceptable and is moving forward to preservation, or was rejected, along with the reason for rejection.

Key Questions for the Repository with regard to the Ingest process:

- What mechanism will contributors use to deposit datasets?
- What geospatial metadata are contributors required to include in the geospatial metadata file?
- How should dataset files and datasets be named for transfer to the Repository?
- How should dataset files be organized for transfer to the Repository?
- How will datasets be packaged for transfer to the Repository?
- How will dataset transfers be validated (e.g. checksum, bag verifyvalid)
- How are datasets transferred to the Repository?
- Do any of the datasets have copyright or access or use restrictions that must be addressed as the datasets are transferred to the Repository?
- How will datasets be handled that do have restrictions? Where will these records be put in the archival organization structure? How will the restrictions be documented and monitored/tracked (if time-based)?

3.2.0. Quality Assurance

With the Repository’s responsibility to hold authentic, reliable records, a critical first step in archival processing is to validate the successful physical transfer of the files (at the bitstream level), and to validate the content (to the extent that’s reasonable) of the submitted data. As recommended above, the Submission Information Package should be deposited into a separate staging area that’s dedicated to verifying and processing received datasets.
The archival processor then performs the necessary validation and verification tasks such as validating the integrity, completeness, and correctness of the data transfer, and confirming that the transferred dataset conforms to organizational, content, and functional guidelines. The following offers an example Quality Assurance process, and the steps will be described in more detail in the following sections:

1. Make a copy of the ingested data so that an original copy of the ingested files is maintained in the event that there are changes made to the files during the Quality Assurance testing.
2. Run a virus checker on received files to ensure no viruses have been transferred.
3. Validate that the data files arrived physically intact and are consistent with the data files the contributor submitted (e.g. check hash values of received data files and verify with transmitted files’ hash values).
4. Validate that the datasets are organized into the agreed to folder structure.
5. Validate that all necessary files have been provided - based on the identified dataset format. Functionally verify the dataset is accessible and renders reasonably.
6. Validate that the geospatial metadata meets the preservation organization’s guidelines and/or policies.
7. Verify the presence of the data in the dataset.
8. Send an acknowledgement to the contributor when completed quality assurance check on the submitted dataset (accept / reject with reason).

Note: To perform the functional and data validation on a submitted geospatial dataset, the archival processor will likely need access to a GIS software product; for example use Esri’s ArcCatalog and Esri’s ArcMap to verify Esri-specific data formats such as shapefiles and geodatabases. Esri also offers a free ArcGIS Explorer.25 The USGS offers a free viewer called DLGV32.26 While the viewer functionality is free, options to “save” require purchase of the product. The USGS offers free metadata tools to view, validate, and save as HTML.27

3.2.1 Archival Processor makes a security copy of the Submission Information Package
During the course of inspecting, verifying, and validating the submission package the archival processor might intentionally or unintentionally alter the files submitted to the archives. Therefore, the archival processor should create a security copy prior to completing any quality assurance activities, so that the original preservation package is preserved for additional archival processing.

3.2.2 Archival Processor runs virus check on files
To ensure that the files are clean and free from viruses, complete a virus check on the submitted files. In less sophisticated content submission approaches, the archival processor may need to manually run a virus checking utility on the files received. In more sophisticated ingestion applications, the ingest program may be able to run the virus checker program on behalf of the contributor, and notify the contributor and archival processor when a file has been found to be infected, and require the contributor to resubmit a new version of the file found to be infected.

When infected files are identified, they will need to be logged and the contributor notified for some sort of remediation. In the case of the sophisticated ingest process, the contributor may receive a success / failure message directly from the ingest application, and be immediately informed of a file rejected due to virus infection, and be prompted to resubmit. In Repositories with more manually-oriented processes, the archival processor will need to notify the contributor of the acceptance or rejection of the submitted files.

The virus checking of geospatial datasets is further complicated due to their multi-file composition. If one file in a dataset collection is found to be infected, then the entire dataset should likely be rejected. Therefore, there will need to be a mechanism to manage the geospatial dataset as a logical unit to accept all or reject all when an individual file is found to be infected.

26 United States Geological Survey. “USGS Digital Data Viewer: dlgv32 Pro.” http://mcmcweb.er.usgs.gov/drc/dlgv32pro/. dlgv32 Pro is a limited feature version of the commercial software - Global Mapper (http://www.globalmapper.com/). Global Mapper v13.00 is able to open shapefiles produced from ArcGIS v9.x, but is only able to open geodatabase files from ArcGIS v10.x or later.
North Carolina runs a virus checker manually on the datasets it receives. The Utah State Archives’ AXAEM system is planning to include virus checking as part of their programmatic file ingest process.

### 3.2.3 Archival Processor validates that the data files arrived intact and are consistent with the data files the contributor submitted

As described in the GeoMAPP Geospatial Data Transfer document, the archival organization and geospatial contributors should collaborate to determine how transferred files will be packaged, and how they will be validated once transferred, as the archival processor will likely be dependent upon the same techniques/technologies/tools that the contributor used to generate the content-level validation key for the files. The archival processor will want to 1) verify that all files have been received, optimally against a manifest that was created when the files were packaged, 2) verify that there are no extra files added to the package, and 3) verify that the files have arrived intact and unaltered through the transfer process. The BagIt tool and its `verifyvalid` option can perform all of these checks.

If the archival processor has received a data submission via a portable disk device, the GeoMAPP team recommends that the archival processor first copy the package of files into the archives’ staging area prior to running the data integrity verification tools so that all inter-device transfers have been completed, thus minimizing the opportunity for any further data transfer oriented corruption. If the package does not pass verification, the archival processor should evaluate the nature of the error (e.g. a missing file, a fixity error on a particular file indicating it has changed). The archival processor may then again copy the package from the transfer disk, and re-invoke the verification tool. If it fails a second time, the archival processor can then run the verification tool on the package on the transfer device to see if it is also corrupted, and if so, will then need to contact the contributor to resubmit the problem file(s).

As part of the initial validation process, the archival processor may choose to employ file format checking / metadata extraction utilities such as JHOVE, DROID (Digital Record Object Identification), National Library of New Zealand Metadata Extractor (NZNZ) or FITS (File Information Toolset) (which wraps several tools). However, these offer limited support for geospatial datasets. JHOVE2 has been enhanced to identify shapefiles.

The Utah State Archives’ AXAEM system currently uses the checksum to validate that a file has been reliably transferred. They also store the checksum in the database for future reference. APPX is going to explore supporting BagIt as part of its ingest process.

### 3.2.4 Archival Processor unpacks the files from the package

Once the transfer package has been verified to be free from viruses and the physical transfer package has been validated as unaltered, the archival processor can unpack the files from the package. The same tool used to create the package will be required to unpack the files. The most common tools are zip and tar, and more recently, the Library of Congress BagIt tool. If BagIt was used to package the files, the archival processor can simply copy the files out of the “bag” folder into the designated processing area folder. This may be configured as another folder in the Staging area (refer to diagram on page 5).

Each archival organization will need to determine how it wants to manage the storage of the originally transferred files (the SIP), and whether they will need to be retained in their originally transmitted packaged format, whether the transmitted files will be retained separately from their transmittal package, or whether the transmitted files will not be retained at all once the Preservation Master has been produced.

The archival organization will also need to determine how it will record file transmittals. BagIt has the benefit of automatically producing a transfer summary log and manifest file that documents the inventory of all of the files in the package. The GeoMAPP team recommends keeping these log files, or at the very least, recording the actions taken during the transfer. This documents the provenance of the files as well as the validity of the transfer should the files’ integrity be challenged. Depending on the specific mechanics of your transfer process, at the time
the Preservation Master is added to the Preservation Repository, new fixity values may need to be computed for all of the files in the dataset which will then become part of the ongoing archival record for that dataset.

Note that the Windows Operating System Windows Explorer has been enhanced to be able to open zipped packages. So, it is now possible for archival processing staff to access the contents of zipped packages without having to purchase and install the zip application (though there may be backward version compatibility issues that may still require zip to open older zip packages).

**Key Questions for the Repository with regard to unpacking the submission package (SIP):**

- What tools can be used to pack and unpack files?
- How to handle received datasets that are not named according to Repository submission guidelines?
- How will the Repository document the processing steps when re-organizing or re-naming dataset files.
- How will the lineage of a renamed dataset in the Preservation Repository be traced back to the originally ingested dataset?

3.2.5 Archival Processor validates that the datasets are organized into the agreed to folder structure and follow naming conventions.

As recommended in the GeoMAPP Geospatial Data Transfer document the GIS and archival organizations should discuss and come to agreement on a general organizational structure for the datasets, and the GeoMAPP team recommends that submitted datasets should conform to these guidelines. Repositories may choose to leverage this same organizational structure in the arrangement of the datasets within the archival repositories, as well as use it to facilitate data access by using it as the basis for a browse/navigation organizational structure for end users. The GeoMAPP team, comprised of both GIS and archival professionals, agreed to use the RAMONA GIS layer categories as a basis for organizing the datasets for both storage and retrieval. This structure offered a navigation structure that is consistent and familiar to the GIS professionals who are likely to be retrieving archived datasets. In addition, the category names are relatively descriptive (e.g. Boundaries, Economy, Transportation, Inland Waters), making them accessible to non-GIS professionals as well.

The GeoMAPP team also agreed to file naming conventions for the datasets, as the dataset file name will often serve as the initial access point for the end user. Please refer to the GeoMAPP Geospatial Data Transfer document for recommendations and guidance on establishing file naming conventions for geospatial datasets.

![Figure 5. Datasets organized in Folders and Datasets organized Layers in Esri Geodatabase](image)

**Note:** Different geospatial format types (e.g. vector datasets vs. raster orthoimagery vs. geodatabases vs. projects vs. digital maps) will likely have different categorization schemes. For example, the GeoMAPP

---

teams organized vector datasets into folders based on their RAMONA categories, organized orthoimagery into folders based on their geographic location (e.g. counties), and organized geodatabases based on their snapshot capture date (e.g. Q1 2011).

Datasets may be delivered to the Repository in a variety of organizational structures and using a variety of naming approaches. The Repository will need to define as part of its submission policies the actions it will take when received datasets do not conform to the naming and/or dataset organizational guidelines. The Repository may choose to re-arrange the received datasets into the Repository’s organization structure in order to facilitate both ongoing management and access.

Depending on the geospatial data file format, the GIS categorical organization structure may be manifested as folders in a file system, as the GeoMAPP team suggests to organize shapefiles, or the categories may be embedded as layer-oriented feature classes within an Esri geodatabase. In the latter case, the archival processor will need to use Esri ArcGIS tools to open and inspect the geodatabase to view the organizational structure of the datasets and the geospatial metadata. Likewise, the archival processor will need to use the ArcGIS tools to perform any re-organization of the datasets stored when they are held within a geodatabase.

**Key Questions for Repositories with regard to the organization and naming verification:**
- How to handle received datasets that are not organized based on Repository submission guidelines?
- How to handle received datasets that are not named according to Repository submission guidelines?
- How will the Repository document the processing steps when re-organizing or re-naming dataset files.
- How will the lineage of a renamed dataset in the Preservation Repository be traced back to the originally ingested dataset?

**3.2.6 Archival Processor validates dataset composition -- all files included in each dataset (format specific), dataset can be rendered**

Geospatial datasets are comprised of numerous files, and the loss of a requisite file in the dataset file composition can disable the ability to render and access the entire dataset. Like other digital file formats, many of the geospatial formats are proprietary and vendor-specific, and are only accessible by that vendor’s tools. Fortunately, many of the GIS vendors seem to be implementing import functions that allow them to import foreign geospatial file formats, but even then, a tool from the original vendor would be required to manipulate the dataset in its native format.

**Shapefiles**

The shapefile is an example of a commonly used geospatial dataset format. Esri originated the shapefile format and they published the technical specification.\(^29\) Now several non-Esri tools are available to render and manipulate shapefile datasets. The specification designates three required files:

- *.shp* - The main file that stores the feature geometry.
- *.shx* - The index file that stores the index of the feature geometry.
- *.dbf* - The dBASE table that stores the attribute information of features.

However, for preservation purposes, there are four other files that GeoMAPP recommends also be included:

- *.shp.xml* - The file that stores the geospatial metadata.
- *.sbn* and *.sbx* - The files that store the spatial index of the features.
- *.prj* - The file that stores the coordinate system information.

While the Esri specification does not designate the metadata file (*.shp.xml) as a required file, for archival preservation, it is beneficial to have a standalone, text-oriented geospatial metadata file accompany the dataset.

---


v1.0 11/2/2011
One method of processing ingested shapefiles would be to check which shapefile component files are included / not included. If one of the required files is missing, the dataset could be rejected and returned to the contributor. In addition to a file-based inspection, the dataset should be opened and viewed using a GIS tool. The archival processor can perform a simple verification by viewing the graphical rendering of the dataset, viewing the data in the dataset, and viewing the dataset’s geospatial metadata. Figure 6 shows an example of viewing a shapefile dataset using Esri ArcCatalog.

![Figure 6. View shapefile graphic to verify dataset renders, as well as look at the dataset data and geospatial metadata (example: ArcCatalog v 9.3)](image)

**NOTE**

ArcCatalog has an option to enable/disable the automatic insertion of missing technical metadata, and will update the metadata last-modified date when the **dataset is opened**. To preserve the originally received geospatial metadata file, GeoMAPP recommends disabling this configuration option when initially inspecting a dataset using ArcCatalog, as illustrated below in Figure 7.

---

30 Note: The Esri Metadata Editor tool user interface has changed significantly with Esri ver. 10.0.
Select: Tools -> Options -> [Metadata]

Figure 7. Disabling auto-insertion of metadata in Esri ArcCatalog v 9.3

**Geodatabases**

More complex geospatial formats, such as Esri’s file geodatabase have a complex folder and file structure and serve as containers for multiple datasets. For these types of formats, in order to verify the information, archival processors will have to open and inspect the datasets using the associated GIS tool (e.g. use Esri’s ArcCatalog to inspect a geodatabase). Once the geodatabase is opened, the individual datasets, stored as feature classes in the case of the geodatabase, can be inspected in much the same way, by functionally verifying that you can access the graphical rendering, verify that the dataset is populated with data, and viewing the dataset’s geospatial metadata. Figure 8 shows an example of viewing a geodatabase that contains several layers using Esri ArcCatalog.

Figure 8. View geodatabase to verify that the dataset renders, as well as look at the dataset data and geospatial metadata (example: ArcCatalog v 9.3)
Orthoimagery

Orthoimagery datasets are often decomposed into a collection of smaller individual “tile” files. Each tile is a self-contained unit, but for preservation purposes, the Repository will want to receive the entire collection of tiles. Additionally, the tiles may be available in both uncompressed (e.g. *.tif with *.tfw world files) and compressed (e.g. MrSID) file formats. Compressed versions of the orthoimages are commonly produced to reduce disk consumption. GeoMAPP recommends that Repositories should receive both versions, if possible, and use the uncompressed files for the Preservation Master and the compressed versions for the Access Copy. Additionally, the imagery contributor may produce access aids such as a shapefile index that provides a graphical representation of each tile on the grid mapped to its associated image file name, or an image-oriented mosaic that provides a stitched together view of the entire collection of tiles.

When inspecting an orthoimage, 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. It will likely be difficult for the archival processor to determine, simply through a visual inspection of the imagery files, if all of the imagery files for the dataset have been provided. Also, due to the large number of tiles that are included in an orthoimage dataset, it is generally not feasible for the archival processor to open and view every image file. GeoMAPP recommends a sampling rate of 5%, and employing either a random sampling or regular sampling (e.g. every 10th tile) method of tile selection. Figure 9 shows an example of viewing a single tile image in an orthophoto dataset using Esri ArcCatalog.

![Figure 9. View orthophoto to verify that the dataset renders, as well as look at the geospatial metadata](example: ArcCatalog v 9.3)

Project Files

Geospatial projects may contain, in addition to the datasets, associated project planning and design files which may be supplied as Microsoft Word, Excel files, Project, Adobe PDF, Visio, etc. files. It would be beneficial to the long term preservation of the project if the contributor provided an index and description of the general composition of the collection of project files to supplement the archival record. Figure 10 shows an example of a directory for a geospatial project, illustrating the variety of project documents and dataset files it might contain.
Key Questions for the Repository with regard to the dataset composition and rendering verification:

- What types of datasets will the Repository be accepting? This will impact the GIS tools the archives will need to have on hand to view/verify the datasets.
- What will be the sampling rate and sampling method for inspecting and viewing orthoimage image files?

3.2.7 Archival Processor validates the geospatial metadata meets the archival organization’s preservation guidelines and/or policies

From a long-term preservation perspective, the geospatial metadata is one of the most valuable components of the geospatial dataset. In defining the geospatial metadata standard, the Federal Geographic Data Committee (FGDC) identified a subset of required metadata elements that serve as a common guideline for all GIS developers. The FGDC metadata covers a wide range of areas that are of interest to the archival organization including:

- title,
- geospatial dataset creator,
- geospatial dataset description,
- geospatial dataset publisher,
- geospatial dataset publication date,
- date(s) of the geospatial data,
- geospatial dataset citation guidance,
- access and use constraints,
- history of the processing steps completed to produce the dataset,
- technical information about the spatial referencing system for the dataset,
- descriptive information about the data fields in the dataset,
- data sources for the data in the dataset,
- technical information about the software used to create the dataset and/or the software recommended to access/render the dataset,
- distributor and ordering information, and
- geospatial metadata update date and contact information.

Lack of technical geospatial metadata can negatively impact the long-term accessibility and management of the geospatial dataset, as it offers preservationists and users guidance on what software to use to access and use the dataset, and important technical geospatial reference information that is important to combining several datasets together for analysis. To ensure the continuing accessibility of the dataset, the technical geospatial metadata is also an important aspect for future dataset migration planning.
In the case of the shapefile, the geospatial metadata may be packaged in a standalone .xml file, which can be inspected through either a text editor or with a web browser (e.g. Internet Explorer, Firefox). Esri’s ArcCatalog provides a useful metadata viewing function, and offers several xml style sheets that provide a more user-friendly formatted presentation of the geospatial metadata, as illustrated below in Figure 11.

![Figure 11. View shapefile geospatial metadata to verify all required fields are populated (example: ArcCatalog)](image)

As was noted on page 17, if using Esri ArcCatalog, to inspect your metadata, the GeoMAPP team recommends that you initially disable the automatic metadata update feature while initially inspecting the metadata for completeness.31

There are also some geospatial metadata checking tools available to validate the metadata file. One is the metadata parser available through the USGS Geospatial Metadata Validation Service website.32 This has the benefit of generating a report documenting errors encountered based on the xml specification for the geospatial metadata file. Tools such as this could be incorporated in the Repository’s preliminary processing.

While the Esri specification does not designate the metadata file (.shp.xml) as a required file, for geospatial preservationists, the geospatial metadata file is an imperative accompaniment to the dataset for long term preservation. If datasets are submitted to the Repository and do not include a standalone geospatial metadata file, the archival processor may be able to use the geospatial tools to generate the accompanying geospatial metadata file with tools such as Esri ArcCatalog or the USGS metadata parser.

**Key Questions for Repository with regard to the geospatial metadata verification:**

- What is the minimum metadata that the Repository will require for accepting geospatial datasets?
- How will archives validate / evaluate the completeness of the geospatial metadata? (e.g. visual inspection of the .xml? visual inspection using something like ArcCatalog to provide a user-friendly display? visual inspection with USGS metadata validator Outline view?)
- How will the Repository deal with missing metadata? Use a tool such as Esri ArcCatalog to autopopulate? Archival processor manually fills in? Automated process? (Given the nature of the geospatial datasets and their metadata, there will be fields that the Archival processor will just not have the ability to complete (e.g. data sources for the data, process steps to produce the geospatial dataset)).
- What is the Repository’s policy with regard to retaining the original geospatial metadata file as part of the archival record -- though it will not reflect the Repository ownership?

---

31 See **Appendix E** for a list of attributes Esri ArcCatalog can automatically update.

3.2.8 Archival Processor opens the dataset using appropriate GIS tool to verify that data is provided with the dataset.

The archival processor should check the data table and confirm that it is populated. The contributor is responsible for the accuracy and authenticity of the data within the dataset submitted to the Repository, and should include the citation for the data sources in the metadata file.

![ArcCatalog - Arcinfo - H:ArcGISDatasetsNCS_AirportsMain.shp](image)

Figure 12. View attribute data in the dataset (example: ArcCatalog)

Also, if special numeric or textual codes are used for data values, these codes should also be documented in the Entity and Attribute Information section of the geospatial metadata file, as seen in the following U.S. Census Bureau TIGER/Line Shapefile metadata for county information. The codes without their meanings may adversely impact the ability to use the data.

Figure 13 below illustrates an example of a coded field in the dataset data attributes that is described in the accompanying metadata file, in this case, referencing a U.S. Census Bureau defined code scheme, and Figure 14 provides an example of the data attribute code documented directly in the accompanying metadata file.
Maintaining and monitoring the fixity of the datasets after they have been ingested serves to reinforce the integrity of the data itself going forward.

**Key Questions for Repository with regard to the geospatial dataset data verification:**

- What will the Repository’s strategy be when inspecting the dataset data?
3.2.9 Archival Processor sends acknowledgement to geospatial data contributor

At the completion of the Quality Assurance checks an acknowledgement should be sent to the data contributor notifying him whether the submitted data passed or failed the Quality Assurance process, and if failed, and suggestions regarding what needs to be corrected to provide an acceptable submission. While some aspects of the Quality Assurance process, such as the virus check, could be automated, several of the steps will require an archival processor to manually open and check the dataset and/or its accompanying files. However, the archival processor could be assisted by the use of a workflow system that provides the task list, and allows the archival processor to record the pass/fail status of each step in the Quality Assurance process. The workflow could then be programmed to send an automated response to the contributor at the completion of the workflow. The workflow log could also serve as documentary evidence of the Quality Assurance steps the archival processor completed when the geospatial dataset was submitted, as well as the final status of the Quality Assurance check. Alternatively, a simple electronic document version of the checklist could be produced, and saved as an administrative document in the AIP.

Quality Assurance Process Workflow Checklist

<table>
<thead>
<tr>
<th>Step #</th>
<th>Step Description</th>
<th>Pass/Fail</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Make copy of submitted digital package</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.2</td>
<td>Run virus check</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.3</td>
<td>Validate package arrived unaltered</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.4</td>
<td>Unpack files from package</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.5</td>
<td>Validate dataset file names and dataset names comply with naming convention and or organized compliant with guidelines</td>
<td>Fail may just be an indicator that the archival processor will need to do some re-arrangement when preparing the AIP</td>
<td></td>
</tr>
<tr>
<td>2.6</td>
<td>Validate dataset composition and that dataset successfully renders</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.7</td>
<td>Validate geospatial metadata is complete to the Repository’s specifications</td>
<td>Fail may just be an indicator that the archival processor will need to do some auto-populating when preparing the AIP</td>
<td></td>
</tr>
<tr>
<td>2.8</td>
<td>Verify presence of dataset data</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.9</td>
<td>Send acknowledgement to geospatial data contributor with status.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

3.3.0 Generate Archival Information Package (AIP)

With the completion of the ingest quality assurance tasks and acceptance of digital materials, the archival processor will next prepare the digital materials for long term storage and deposit them into the preservation storage repository. This entails preparing the Preservation Master image of the digital asset(s) and then packaging them with their related archival administrative metadata to produce the Archival Information Package (AIP). The archival processor’s goal with this process is to produce a robust digital object that ensures the security of, at a minimum, the information held in the digital object, and at best, the original presentation and user experience involved in interacting with the digital object.

Figure 15 provides a conceptual depiction of the relationship between the originally ingested geospatial dataset object and its manifestation in the archival information package (AIP). The AIP will contain the original dataset object (its component file(s)), and its associated archival metadata that is produced through the archival processing process. As seen in the following illustration, the geospatial metadata may be “harvested” to populate the relevant sections of the Archival Metadata record. The following sections will describe the creation and processing of this Archival Information Package.

33 Harvesting may be performed either via manual data entry/data mapping from the source files to the archival management system, or through utility programs that can parse the source metadata files and automatically populate the metadata values in the archival management system.
Figure 15. Producing Archival Information Package from SIP - conceptual illustration

Note: Figure 15 illustrates an AIP that contains one single dataset. Depending on the Repository’s technical and organization strategy, an AIP may contain multiple datasets, for example an AIP may represent an Archival “Series” of related datasets.

The steps entailed include:
1. Creating the unique identifier for the new digital object being preserved.
2. Saving a copy of the original geospatial metadata file
3. Identifying the Preservation Master file(s) and format. Optionally, reformatting received digital assets to promote long-term preservation.
4. Optionally, updating the geospatial metadata file to populate missing technical metadata fields -using tools such as Esri’s ArcCatalog to auto-populate these fields.
5. Updating the geospatial metadata file to document the transfer of ownership to the Repository, and the Repository’s processing step(s) in preparing the dataset for preservation.
6. Establish fixity value(s) for the digital asset(s) to be contained in the AIP.
7. Create the Archival Metadata record for the digital asset. Populate descriptive, technical, and administrative metadata for the digital record.
8. Adding the AIP to the preservation repository.
9. Creating security copy(ies) of the Preservation Master record.

Note: depending upon your technical implementation, the order of these steps may be altered, but the goal at the end of this phase is to have a packaged digital object that is accompanied with:
- its associated fixity information - to ensure the long term integrity of the digital files comprising the digital object, and
- its associated archival metadata which is necessary to manage the digital object over its life time in the archival repository.
3.3.1 Archival Processor creates unique identifier for the new digital object being preserved

The archival processing process produces numerous artifacts that are all related to the digital object that is being preserved, including the originally ingested digital file(s), a possibly reformatted version to ensure long-term preservation, and the various access derivatives that may be produced to facilitate the end user’s access to the information held in the digital object. A significant design decision for an archival / preservation repository is the mechanism it will use to maintain the relationships between these various digital assets.

One method is to utilize a unique identifier to represent the entire digital object being preserved, including all of its constituent related resources. Through the course of the archival processing of a digital object, all of the artifacts consumed and produced could then be linked together through this unique identifier, as illustrated below in Figure 16.

Figure 16. Example Information Model Utilizing Unique Identifier to Represent Digital Object Being Preserved

There are a variety of ways that a Repository may choose to instantiate and implement this unique identifier. Technical selection of an approach will depend upon the domain the organization wishes guarantee uniqueness to apply ... whether at just the organizational level or external to the organization. Database products, such as Oracle, have the ability to generate a unique identifier which can be used internally as the key to join several records together. There are also a variety of open source unique identifier generator utilities.34

Key Questions for Repository with regard to the unique identifier:

- What mechanism will the Repository use to associate the various related digital artifacts associated with the archival processing of geospatial datasets?
- What will the general information architecture be to manage the digital objects and the metadata throughout the archival processing life cycle?

34 For an example, see the UUID Generator. Retrieved 10/27/2011 from: http://sourceforge.net/projects/uuidgenerator/
3.3.2 Archival Processor saves a copy of the original geospatial metadata (file)

Geospatial datasets are descriptively augmented by a substantive metadata standard. The datasets contributed to the GeoMAPP partner states were described using the Federal Geographic Data Committee (FGDC) metadata standard. The metadata can be embedded within the dataset and/or produced in a companion geospatial metadata .xml file. Unlike the typical document, image, and audio formats, the geospatial metadata standard goes well beyond the basic descriptive and technical metadata elements. The geospatial metadata represents the state of the geospatial dataset at the time it was submitted to the Repository, and includes descriptive, constraint, technical, lineage, data definitions, data citations, contact, distributor and ordering information metadata. It can also include process steps completed to produce the dataset, a data dictionary, and (hopefully) data source citations for the data part of the dataset.

There is also the ongoing challenge of dataset developers creating the geospatial metadata. In the case of Esri ArcCatalog - in the absence of the developer creating metadata, ArcCatalog can assist in at populating several of the metadata fields based on the dataset’s properties. Esri can populate technical dataset information that can be extracted from the geospatial dataset itself, such as the bounding coordinates, and the coordinate system name and coordinate system values. These metadata values, decoupled from the physical geospatial dataset in an ASCII text readable format can enhance the long-term preservability of the datasets.

The geospatial metadata file often serves as a fundamental access aid for GIS distributors (e.g. clearinghouses, inventory systems), documenting the distributor, the distributor’s access and use constraints based on the policies of the distributing organization, and distribution fees. Therefore, as the geospatial dataset is transferred to the Repository, an updated metadata file may need to be produced that reflects the Repository’s role as the distributor of the dataset. The Repository should also document its ingest and processing of the dataset as a Process Step in the metadata file. If a Repository chooses to produce a Repository-specific version of the geospatial metadata, it will be critical to retain the originally received metadata to preserve the original historical record. You may consider the Repository-specific metadata file as yet another dataset derivative produced to facilitate access to the dataset.

Therefore, an early archival processing step should be to make a copy of the original geospatial metadata file. We recommend that you establish a common naming convention that reflects the original status of these files. The geospatial metadata file will generally reflect the dataset name with an .xml extension (e.g. 1992_Congressional_Districts.shp.xml). An example naming convention might be to insert “ORIG” into the name prior to the .xml extension (e.g. 1992_Congressional_Districts.ORIG.shp.xml). You should maintain the .xml extension so that the file continues to be identified as an xml file. Some systems use the .shp.xml file extension pattern to identify this file as a shapefile-related xml file, therefore, this part of the file naming pattern should be left intact.

**Key Questions for Repository with regard to the original geospatial metadata file:**

- ✔ Will the Repository make updates to the geospatial metadata (file) to reflect the transfer to the Repository? If not, then it’s not necessary to make an original copy of the metadata file .... though it’s probably still not a bad idea, as the ArcGIS ArcCatalog tool may be configured to automatically update the geospatial metadata file in the dataset, and the unaltered original metadata file would be lost.

- ✔ What naming convention will the Repository employ to indicate the original file?

3.3.3 Archival Processor Identifies the Preservation Master file(s) and format. Optionally, reformats/transforms received digital assets to promote long-term preservation

A critical step in the preservation process is identifying the Preservation Master file(s) and how it (they) will be produced from the Submission Information Package. In the simplest case the file(s) submitted in the Submission Information Package will map directly to the Preservation Master.

---

35 See Appendix E for a list of attributes Esri ArcCatalog can automatically update.
In some cases, the submission package might include both uncompressed and compressed versions of the geospatial dataset. Following archival best practices, the uncompressed version of the geospatial dataset should be the basis for the Preservation Master. This distinction and preference may also be documented on your Preservation Data Type Plan. Uncompressed variants are preferred as the Preservation Master in the event the dataset needs to be migrated to a new format, and uncompressed formats contain the full set of data and information for the dataset, whereas compressed datasets often result in some sort of data loss through the compression process.

As many geospatial datasets exist in complex formats that may be waning or offer limited access with current tools, archival processing may involve transforming the datasets to a more sustainable data file format to enhance the long-term preservability of the geospatial dataset. This is where it will be helpful to have your Preservation Data Type Plan in place, to streamline the decision to reformat, as well as identify the target format to reformat the dataset to, and what tools and/or processes to execute to transform the dataset. If the Repository chooses to reformat the geospatial dataset, it should retain the original dataset, as it is unknown if future transformations will be necessary, and if those subsequent transformations will be most effective sourced from the original dataset format or the reformatted version. Over the course of a geospatial digital object’s life, as the software technology continues to evolve, the geospatial digital object may require several reformattings. The reformatting process should also be recorded in the archival metadata record, and even in the geospatial metadata record, to maintain

Figure 17. Decision tree branch to handle the many varieties of geospatial data file formats

the lineage and manipulations performed on the geospatial digital object over the course of its life cycle. Having a unique identifier to associate both the original and reformatted digital materials will also aid in maintaining the relationship between the original and reformatted digital objects. A user or process should be able to trace through to the original digital object from the reformatted digital object, or conversely, be able to trace through to the reformatted digital object from the original.

Example: The shapefile offers an opportunity for a transformation to promote long-term preservability. The dataset data attributes are stored in a dBASE .dbf format, which may be transformed to a comma-separated values text-based file. At this time, Excel 2007 is able to read .dbf formatted files, and can be used to convert them to a .csv formatted file. In the event the shapefile dataset or the individual .dbf file cannot be read, the data that was held in the dataset would be accessible independent of any GIS or database application.

Each of the GeoMAPP Partners selected a slightly different approach when designating their Preservation Masters, largely based on the differences in how geospatial data was submitted to each archival institution.

- KDLA receives quarterly geodatabase snapshots from their GIS clearinghouse, and retains their Preservation Master in the geodatabase format. Shapefiles are extracted from the geodatabase, retained as Preservation Masters, and individually deployed through the KDLA e-archives application, based on DSpace.

- NCSA receives shapefile datasets from their GIS clearinghouse when they are superseded by a newer version of the dataset, and the collection of files that comprise the shapefile are retained as the Preservation Master.

- Utah State Archives receives shapefile and geodatabase versions of the individual layers, and retains both formats in its Preservation Master. A geoPDF is also received and retained as a dissemination copy, as are separate metadata files in HTML and XML form.

Key Questions for Repository with regard to the creating the Preservation Master:

- What are the acceptable formats for the Preservation Master?
- What are the mappings between expected submitted formats and their companion Preservation Master format? (refer to your Data Type Preservation Plan)
- Will compound datasets be decomposed into their individual datasets (e.g. will individual feature classes in a geodatabase be extracted as individual shapefile datasets; will individual raster images be exported separately)?
- What tools and processes will be employed to transform a geospatial dataset received in one format to another format to produce the Preservation Master?

3.3.4 Optional: Archival Processor updates the geospatial metadata file to populate missing metadata fields - using tools such as Esri’s ArcCatalog to auto-populate these fields

The archival processor will need to inspect the geospatial metadata file and determine if it is missing any required metadata fields, and if the available GIS tools can assist in populating these fields. Examples of some geospatial fields include:

- Identification Information -> Spatial Domain (e.g. Bounding Coordinates)
- Identification Information -> Native Data Set Environment: operating system and GIS software application name and version
  
  **Example:** Microsoft Windows XP Version 5.1 (Build 2600) Service Pack 3; ESRI ArcCatalog 9.3.0.1770
- Spatial Data Organization Information -> Direct Spatial Reference Method (controlled vocabulary that identifies whether the dataset is a “Point”, “Vector”, or “Raster”. This field could be extremely useful as a field to search against for future processing and/or access.
- Spatial Reference Information -> Coordinate system name and related information

---


v1.0 11/2/2011
The following table lists some of the technical geospatial fields that Esri ArcCatalog can automatically populate.

<table>
<thead>
<tr>
<th>Field</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Horizontal coordinate system definition</strong></td>
<td></td>
</tr>
<tr>
<td>Projected Coordinate system name **</td>
<td>NAD_1983_StatePlane_North_Carolina_FIPS_3200</td>
</tr>
<tr>
<td>Geographic coordinate system name **</td>
<td>GCS_North_American_1983</td>
</tr>
<tr>
<td><strong>Planar</strong></td>
<td></td>
</tr>
<tr>
<td>Grid coordinate system name **</td>
<td>NC: State Plan Coordinate System 2983</td>
</tr>
<tr>
<td></td>
<td>Utah: Universal Transverse Mercator</td>
</tr>
<tr>
<td></td>
<td>Maine: Universal Transverse Mercator</td>
</tr>
<tr>
<td>plus several other geospatial factors **</td>
<td>e.g. Lambert conformal conical</td>
</tr>
<tr>
<td><strong>Planar coordinate information</strong></td>
<td></td>
</tr>
<tr>
<td>Planar coordinate encoding method **</td>
<td>coordinate pair</td>
</tr>
<tr>
<td>plus several other geospatial factors **</td>
<td></td>
</tr>
<tr>
<td><strong>Geodetic model</strong></td>
<td></td>
</tr>
<tr>
<td>Horizontal datum name **</td>
<td>North American Datum of 1983</td>
</tr>
<tr>
<td>Ellipsoid name **</td>
<td>Geodetic Reference System 80</td>
</tr>
<tr>
<td>Semi-major axis **</td>
<td>6378137.000000</td>
</tr>
<tr>
<td>Denominator of flattening ratio **</td>
<td>298.257222</td>
</tr>
</tbody>
</table>

** Indicates fields that Esri ArcCatalog can automatically populate

Key Questions for Repository with regard to the updating missing technical metadata fields:

- What are the technical metadata fields that the Repository will require?
- What tools, techniques, or strategies will the Repository employ to populate the missing technical fields?
- Note: One possible option is to reject a dataset at submission time that does not have the requisite technical metadata fields populated ... especially as these are especially crucial to the long-term preservability of the geospatial dataset.
- What updates or modifications will the Repository make on the geospatial metadata file?

3.3.5 Optional: Archival Processor update the geospatial metadata file to document the transfer of ownership to the Repository, and the Repository’s processing step(s) in preparing the dataset for preservation, and add ISO Theme Keywords

The Repository should determine its policies with regard to the fields in the metadata file that reflect the contributor’s ownership or policies, and how these may be different from the Repository’s information and polices, and how the Repository’s policies will be conveyed for users who access these datasets - especially as the geospatial metadata file is the primary vehicle for conveying this information. There are a couple of possible alternatives the Repository could explore including:

1) documenting the Repository’s policies in the archival metadata outside of the geospatial metadata file that could then be surfaced in the finding aid or catalog entry presentation, and/or

2) updating the geospatial metadata file itself.

The advantages of documenting the policy in the archival metadata are that the Repository probably already has these processes, tools and presentation methods in place to record and display at least use and access policies for its archival materials. The disadvantage is that once a user downloads the dataset it will be de-coupled from the archival interface, and the dataset metadata conveying the contributors’ policies will then be the only information the dataset user sees. With the second option of updating the geospatial metadata file, when the user downloads the dataset from the Repository, the dataset metadata will indicate to the user that the ownership has changed.
The Repository could choose to update the entire geospatial metadata file to reflect the Repository’s ownership, but this will likely be infeasible from a manpower availability and effort perspective. However, a Repository with a savvy Information Technology Services shop might be able to produce a program that automates a standardized update to the geospatial metadata file -- at least for those fields that would be common across all datasets such as the **Point of Contact** and **Distributor**. Alternatively, the Repository could pursue a “lightweight” strategy and place a boilerplate statement in the **Abstract** field that indicates that the dataset ownership has been transferred to the Repository, and that the user should consult the Repository for the use/access constraints, point of contact, distributor, citation information, etc. (e.g. This dataset was accessioned by ‘xyz’ Repository on mm/dd/yyyy. The only changes to the metadata file are in the Abstract, Themes, and adding a new Processing Step. Consult the ‘xyz’ Repository entry for the dataset for citation, contact, and access/use constraint information which may supersede the information and policies conveyed within the metadata file. The original metadata file has been retained for historical reference and is available as name_ORIG.shp.xml). As an alternative to embedding the statements within the metadata file, the Repository may also present its policy statements in its access interface.

The Repository should also add a **Processing Step** that records the ingestion of the geospatial dataset into the Repository, and note what changes it made to the geospatial metadata file. The Repository may also note in its Processing Step that it has retained the originally received geospatial metadata file, provide the name of the original file, and how end users can access the original geospatial metadata file.

However, if datasets are being stored as feature classes within an Esri geodatabase, the geospatial metadata is embedded within the geodatabase tables and is not accessible as an independent file. For geodatabase configurations, the archival processor may export independent metadata files for each of the feature classes through Esri ArcCatalog, verifying that the metadata auto-update feature is disabled. These individual metadata xml files could then be packaged with the collection of geodatabase files for archival storage.

The following are examples of some constraint statements in the geospatial metadata that may not reflect the Repository’s use and/or access policies, or may reflect policies that are obsolete based on the passage of time:

**Use Constraints:**
The use constraints documented in the geospatial metadata file may not reflect the Repository’s use policies.

**Example 1: 1992 NC Congressional Districts (NC CGIA)**
Acknowledgement of products derived from this data set should cite the following: The source of the 1992 Congressional voting districts data is the North Carolina Corporate Geographic Database. Earlier versions of this dataset may exist. The user must be sure to use the appropriate data set for the time period of interest. While efforts have been made to ensure that these data are accurate and reliable within the state of the art, CGIA cannot assume liability for any damages or misrepresentation caused by any inaccuracies in the data or as a result of changes to the data caused by system transfers.

**Example 2: NC Shellfish Growing Areas 2009_04 (NC CGIA)**
This data consists of shellfish classifications valid through 04/15/2009. Please contact the Shellfish Sanitation and Recreational Water Quality Section (SS+RWQS) for information on this dataset.

**Example 3: NC Natural Heritage Element Occurrences Polygons 2006_08 (NC CGIA)**
These data are intended for research or planning projects that will contribute to better protection for the ecological features involved. Due to its dynamic nature, this data becomes outdated very quickly. The Natural Heritage Program must be contacted before each use of the data set to ensure data currency. Acknowledgement of products derived from this data set should cite the following: The source of the Natural Heritage Element Occurrences data is the North Carolina Corporate Geographic Database. Earlier versions of this dataset may exist. The user must be sure to use the appropriate data set for the time period of interest. While efforts have been made to ensure that these data are accurate and reliable within the state of the art, CGIA cannot assume liability for any damages or misrepresentation caused by any inaccuracies in the data or as a result of changes to the data caused by system transfers.
Example 4: SGIC93_BIOSCIENCE_Habitat_Moose (Utah AGRC)

The Utah Automated Geographic Reference Center has adopted the following spatial data disclaimer to be explicitly included or referenced in all geospatial data, mapping products, and services created or hosted at AGRC including the contents of State Geographic Information Database (SGID).

'This product is for informational purposes and may not have been prepared for, or be suitable for legal, engineering, or surveying purposes. Users of this information should review or consult the primary data and information sources to ascertain the usability of the information. AGRC provides these data in good faith and shall in no event be liable for any incorrect results, any lost profits and special, indirect or consequential damages to any party, arising out of or in connection with the use or the inability to use the data hereon or the services provided. AGRC shall not be held liable for any third party's interpretation of data provided by AGRC. AGRC provides these data and services as a convenience to the public. Furthermore, AGRC reserves the right to change or revise published data and/or these services at any time.'

Furthermore, it is the official policy of the AGRC:

- that the adopted disclaimer be used on all hard copy maps produced from geospatial data, and that the date and source of the data be included on the map;

- that spatial data producers be allowed to extend the adopted disclaimer with additional language further defining the limits of their liability;

- that a more robust disclaimer may be used in conjunction with any and all geospatial data published on the Internet, on a separate page preceding access to the data, with an accept/reject option for users;

- that standardized metadata be included with any distribution of all geospatial data; and

- that the disclaimer above may be used as a blanket disclaimer for documents containing a number of small maps.

Access Constraints

The access constraints documented in the geospatial metadata file may not reflect the Repository’s access policies.

Example 1: NC Shellfish Growing Areas 2009_04 (NC CGIA)

Please contact the Shellfish Sanitation and Recreational Water Quality Section (SS+RWQS) for information on this dataset.

Example 2: NC Natural Heritage Element Occurrences Polygons 2006_08 (NC CGIA)

Although present in the Corporate database, the element occurrence data remains under the supervision of the Natural Heritage Program. Permission MUST be given by the Natural Heritage Program for each different use of the data. Although more significant ecological features are lost because of ignorance and accident than because of exploitation or intentional destruction, publication of locations of many species increases the risk of deliberate damage to them. To minimize this risk, elements are not identified in the CGIA database. Points are labeled only as plant, animal, or natural community occurrences. If a particular occurrence is of interest, the Natural Heritage Program must be contacted to determine the identity of the element and its significance to the project concerned. There may be a charge for Natural Heritage Program services in providing the data, in addition to CGIA fees. A preapproval letter must be sent from CGIA to the source agency before release of this data can occur. Contact Production Services at (919)733-2090.
Distributor:
The Repository effectively becomes the distributor of datasets which it accessions, however, having a record of the original distributor will be useful for maintaining lineage and provenance of the dataset.

Example 1: 1992 NC Congressional Districts (NC CGIA)
NC Center for Geographic Information and Analysis

Example 2: Utah: SGID93_BIOSCIENCE_Habitat_Moose (Utah AGRC)
Automated Geographic Reference Center

Example 3: Kentucky (KYGeoNET)
Kentucky KYGeonet: retains the originating distributor (e.g. Wildlife Management Areas)
Originator: KY Dept of Fish & Wildlife Resources
Distributor: KY Fish & Wildlife Information System

Example 4: Kentucky Legislative Districts (KYGeoNET)
Originator: Division of Geographic Information
Distributor: Division of Geographic Information

Another best practice for the geospatial metadata is the inclusion of the associated ISO 19115 Topic Category\(^\text{39}\) as a keyword thesaurus and keyword field. This provides a consistent method of description through a geospatial-community-controlled vocabulary. These offer a categorical scheme that is familiar to GIS professionals, and the Topic Category could also be used as a basis for search indexing for the access interface. Also, if this field is consistently entered and assigned, it could be used to drive a category-based navigation model for the user interface.

\[
<\text{theme}>
  <\text{themekt}>\text{ISO 19115 Topic Category}\</\text{themekt}>
  <\text{themekey}>\text{boundaries}\</\text{themekey}>
</\text{theme}>
\]

North Carolina State Archives is exploring a light-weight geospatial metadata management approach, by adding a Processing Step and archival processor contact information that documents the Archives’ addition of the dataset to its collection, and if not already provided, adding the theme keyword metadata reflecting the topic category of the dataset.

Utah and Kentucky are not yet modifying their received geospatial metadata files.

**Key Questions for Repository with regard to the updating the geospatial metadata file:**

✔ What is the Repository’s strategy for conveying that the geospatial dataset ownership has transferred to the Repository within the metadata file?

✔ What updates or modifications will the Repository make on the geospatial metadata file?

### 3.3.6 Archival Processor establishes fixity value(s) or mechanism for the digital asset(s) to be maintained in the AIP

One of the most important archival responsibilities is ensuring the integrity of the digital assets that are in its care for long-term preservation. A fixity value needs to be computed for each digital item, and then periodically re-computed to ascertain that the physical file contents have not been modified (either through mechanical or software failure, or human intervention). Therefore, a baseline fixity value needs to be computed and saved for each file that is to be managed by the Repository.

There are a variety of techniques, tools, and strategies a Repository may pursue, but in essence the fixity values need to be stored and be accessible for future validation checks for each of the files comprising the Preservation Master and Preservation Security copies that make up each geospatial dataset “object.” Note that as the geospatial dataset is usually a complex, multi-file collection that there will be multiple fixity values associated with a single geospatial digital “object,” so even if you have a method in your existing archival database for recording fixity values, it may need to be augmented to support the geospatial multi-file scenario.

The Repository can choose to develop its own utilities for computing the baseline fixity values and develop its own data store (e.g. database) to store the fixity values for each dataset file, or incorporate them into its archival database. In addition to recording the fixity values, the Repository needs to establish the processes and tools to:

a) periodically re-compute the fixity values on the files and compare them with the baseline,

b) establish a replication mechanism to make a security copy(ies) to be used when a damaged file needs to be restored,

c) compute the fixity values on the replicas to make sure they are in sync with the “master”, and

d) replace damaged files as they are identified.

The Repository might consider using the BagIt tool to assist with the establishment and monitoring of the fixity values, as in the course of creating a bag the BagIt tool computes and records checksums for each of the items put into the bag. A replica of the bag could be created as the Security Copy, and then periodically, bag verifyvalid could be run on both the Preservation Master as well as the Security Copy(s) to verify that the files in the bag are still intact.

Alternatively, there are storage solutions available that embed fixity baseline calculations, file replication, fixity checking, and remediation when a file is found to be corrupted. In the case of the latter, the Repository would not likely need to explicitly maintain the fixity values themselves in the AIP or the archival database, but should document in the archival metadata for the AIP the technical mechanism used to maintain and monitor the fixity values. If a Repository chooses to create their security copies on tape-based media, which are usually stored off-line, then the tapes must be accessible in order to perform periodic fixity checks on the security copy, and/or restore a damaged Preservation Master file.

North Carolina is exploring the open source ADAPT Audit Control Environment (ACE)40 Audit Manager system to manage the fixity of their files, while Utah is recording checksum fixity values in their AXAEM database, which can be later used as the basis for validating the files’ integrity.

Key Questions for Repository with regard to establishing digital asset fixity:

✓ What method will the Repository use to compute, store, re-compute, and compare fixity values for both the Preservation Master and Preservation Master Security Copy(ies)?
✓ Will the Repository store the fixity values for the geospatial dataset files within the archival metadata?
✓ Will the Repository store the fixity values for the geospatial dataset files within the AIP itself?
✓ How would someone perform an audit check to view and/or validate that a file’s integrity is intact?

3.3.7 Archival Processor creates the Archival Metadata record for the digital asset. Populates descriptive, technical, and administrative metadata for the digital record

In order to make the geospatial dataset accessible and to be able to manage it for the long-term, an archival metadata record needs to be created to accompany the geospatial dataset. In the OAIS model, this information

would be stored and managed in the Data Management module. This record will describe and manage the geospatial “object”, which might be comprised of numerous files. Archival metadata goes beyond the basic descriptive information, and provides:

a) Technical information about the object being managed, including the software application name and version used to create and/or access the digital object, and in the case of geospatial data, technical characteristics such as the projection and coordinate systems that are necessary in order to be able to render the geospatial dataset;

b) Administrative information to manage the digital object ranging from accession/ingestion information, use rights, access constraints, the provenance or history of the object, ongoing record of any changes made to the object after it’s been ingested by the Repository (e.g. initial processing, or future format transformations or migrations), information about the contributor, and if applicable, the submission agreement; and

c) Descriptive information to facilitate access to the object being managed, including title, author, date of record, keywords, etc.

As the geospatial dataset is comprised of numerous files, the archival metadata record should, at a minimum, store the “manifest” of the geospatial dataset that lists all of the files that comprise the dataset, or a pointer to a manifest file. In addition, the Repository may also want to record in its archival database the basic descriptive (name, size, date created, date modified), and technical metadata (type of file or format, application name and version) for each of the files that comprise the geospatial dataset.

As geospatial datasets often are accompanied with a rich companion metadata that aligns with many of the important archival metadata fields (e.g. title, date, abstract, and purpose can comprise the description, originator can be used for author, keywords can be used as the basis for subject headings), the Repository may choose to populate the archival metadata record by harvesting the geospatial metadata record. However, it probably is not necessary to replicate the entire geospatial metadata record into the archival metadata record, and in some cases the original geospatial metadata value may not represent the appropriate value for the archival record (e.g. use or access constraints in the original geospatial metadata will not likely reflect the Repository’s use and access statements). Therefore, the Repository may want to be selective with regard to which geospatial metadata fields are extracted into the archival metadata record. The descriptive fields that map to the Dublin Core metadata fields are a good start, and could then be expanded to any additional fields that aid in access such as the bounding coordinates. Then, the archival processor should consider extracting fields that aid in the access and manageability of the dataset, such as fields from the Identification Information section of the geospatial metadata (e.g. the geospatial layer should be identified as a subject field), and the geospatial metadata standard name and version. End users may be interested in exploring other metadata fields prior to downloading the dataset, and rather than replicating all of the geospatial metadata in the archival metadata record, the Repository can simply publish the geospatial metadata file through its the archival access interface. If the Repository will not be providing online access to the geospatial datasets or the geospatial metadata, then the Repository may want to consider replicating all of the geospatial metadata in the archival record and making that available through its archival search interface.

At this time, there are no real standards regarding what metadata actually should comprise the archival metadata record. The GeoMAPP team offers a proposed archival metadata dictionary organized based on the OAIS Preservation Metadata Taxonomy in the GeoMAPP “Archival Metadata Elements for the Preservation of Geospatial Datasets.” white paper.41 The metadata dictionary is intended to be flexible enough to accommodate a variety of digital formats, not just geospatial datasets. In developing this data dictionary, the archival metadata models of several preservation initiatives were consulted including:

• CEDARS,\textsuperscript{42} - offers a geospatial preservation perspective,
• CIESIN GER\textsuperscript{43} - offers a geospatial preservation perspective,
• the Stanford National Geospatial Digital Archive Project\textsuperscript{44} - offers a geospatial preservation perspective,
• the National Library of New Zealand,\textsuperscript{45} and
• the National Library of Australia.\textsuperscript{46}

The Defense Technical Information Center\textsuperscript{47} also published a preservation metadata model intended for the long-term management of digital materials, which documents a comprehensive preservation metadata model which, likewise, accommodates a variety of digital formats. This strategy allows a Repository to develop a single archival metadata model to manage the variety of digital formats that will likely comprise their digital holdings. It identifies a core set of common metadata elements that will apply across all archived assets, and then offers digital format-specific metadata fields to apply based on the format of the digital object being archived. This model offers the flexibility of extensibility in formats, as new formats can be easily added to the model, and isolates individual formats from the complexities of all of the digital formats.

The archival metadata record can be populated either manually and/or through technology. However, even if an automated archival metadata assignment program is created, it is likely that some manual intervention will still be required to complete the metadata record ... if for nothing else, to perform a quality assurance inspection. There are a variety of tools such as PRONOM,\textsuperscript{48} JHOVE/JHOVE2,\textsuperscript{49} DROID,\textsuperscript{50} the New Zealand Metadata Extractor (NLNZ),\textsuperscript{51} and the wrapper FITS\textsuperscript{52} that offer varying degrees of functionality and reliability in identifying data formats and extracting metadata. These will offer limited utility when creating the archival metadata record for geospatial objects. JHOVE2 can now recognize the shapefile format, and PRONOM recognizes .shp as an ESRI Arc/View Shapefile (or 3D Studio Shapes), and .adf as the ESRI Arc/Info Binary Grid format, which may be useful for the data format identification part of an automated archival metadata process. However, they currently offer no services for extracting the geospatial metadata from the geospatial metadata (.xml) file. Therefore, additional custom programming will be required to implement a utility that automatically extracts metadata from the source files and populates the archival metadata record.

The Utah State Archives conducted some testing with JHOVE2. They found useful its overall ability to determine that a group of records in a folder comprised a shapefile, and to report:

• the number of files in the group,
• the file names for each of the files,
• the file format(s) for each of the files, and
• the file size for each of the files.

JHOVE was also able to extract the attribute fields from the dBASE .dbf file, as well as the number of data records (these make up the data attribute table for the dataset), and some other information from the shapefile header. Alternatively, the attribute information could be extracted from the geospatial metadata file, so an organization could have an implementation choice.

However, when presented with the files individually, the relationship between the files was lost, and JHOVE2 could no longer recognize the shapefile construct. It could identify, based on PRONOM lookup, that the .shp file was either a 3D Studio Shape or a shapefile, but did not run the shapefile module. For more information on the Utah State Archives’ JHOVE testing, refer to the GeoMAPP project Final Report.

The Utah State Archives is in the process of implementing the APPX-based Archives Enterprise Manager (AXAEM) system to create and manage their geospatial records. It offers an extensive archival metadata schema, and APPX has developed a metadata extractor that populates the archival metadata record directly from the JHOVE2 output and the geospatial XML file. The AXAEM interface also includes options for designating the various archival asset renditions, such as whether the record references the Original (Preservation Master), a Preservation Copy, or an Access Copy (see left screen capture in Figure 18).

Figure 18. AXAEM interface for creating archival metadata

**Key Questions for Repository with regard to the updating the geospatial metadata file:**

- What technical mechanism will be used to record and store the archival metadata?
- What metadata elements comprise the archival metadata record?
- What geospatial metadata fields will be extracted and included in the archival metadata record?
- Will the Repository create a modified geospatial metadata file that
  
  a) has missing technical metadata fields filled in?
  
  b) reflects transfer of ownership of the geospatial dataset to the Repository organization?
- How will the Repository record its processing steps in the archival metadata record (e.g. dataset format transformations, geospatial metadata modification or customizations)?
- Will the fixity record for each of the files be recorded as part of the metadata record?

**3.3.8 Archival Processor creates the AIP and add it to the preservation repository**

Once the geospatial datasets and accompanying archival metadata records have been prepared, the digital object representing the Preservation Master can be “packaged” and deposited into the archival repository. What constitutes the Archival “Record” will be defined by each Repository. However, at a minimum, it should include the original (and reformatted, if applicable) dataset including the original collection of data files and the originally received geospatial metadata file, plus the Repository’s customized version of the geospatial metadata file, if applicable. Additionally the Repository may want to store the record of the transfer for chain-of-custody recording purposes. If using BagIt, this may take the form of storing, for example, the BagIt bag-info.txt, manifest-md5.txt and tagmanifest-md5.txt files as part of the archival record for the Preservation Master.

The technical physical manifestation of the Archival Information Package will depend upon each Repository’s technical implementation. This could entail adding data to a database and/or adding files to a file system.
designated for archival use. PREMIS\textsuperscript{53} and METS\textsuperscript{54} can be used to document the package specification. Packaging tools such as BagIt or zip could also be used to create a package-type object that contains the file-related artifacts for the geospatial dataset object being preserved. BagIt has the benefits of producing a documented manifest of the contents of the package, and recording checksums for each of the items in the bag, and offering the \texttt{bag verifyvalid} utility that can be interactively run at any later time to verify that the bag’s files are still intact and have not been modified.

The Repository can choose the composition of its Archival Information Package. The simplest strategy is to create an AIP for each geospatial dataset, and the archival metadata will directly represent the individual dataset. Alternatively, the Repository may choose a lower level of granularity, for example, store a collection of datasets that are represented as an Archival Series. However, in this case, the Archival Metadata record will not necessarily be harvested directly from the geospatial metadata, but might be defined by the archival processor.

Ideally, to minimize the opportunity for data corruption, human error, or data loss, access to the Preservation Master should be minimized. Therefore, as the AIP is added to the preservation repository, the appropriate access controls should also be configured for the data object(s) stored in the repository. The Repository manager should define access policies for the AIP and its associated digital artifacts, and then work with the Information Technology department to determine how these policies will be implemented within the capabilities and constraints of the selected archival storage solution hardware and software. It may require the establishment of file-based access controls and/or application-based access controls. GeoMAPP suggests that the Repository manager establish rather restricted access to the Preservation Master and Security Copies especially considering their uniqueness and associated fundamental value. In some cases, the Preservation Master may effectively be the “last copy” of a digital artifact, and thus must be treated with great care and consideration, especially with regard to digital access.

**Key Questions for Repository with regard to the updating the geospatial metadata file:**
- How will the Repository structure and store its Archival Information Packages?
- Will the AIP contain a single dataset or a collection of dataset?
- Will the Repository use a special packaging mechanism (e.g. METS, zip, BagIt)?
- Who will have access to the Archival Information Packages? What type of access (read, write)?
- What access policies and rights will be configured for the AIP? Who can access it? For what purpose?

### 3.3.9 Archival Processor Creates Security Copy(ies) of the Preservation Master record

With the ever-present risk of the loss of digital data due to either hardware, software, or human error, one or more security copies of the Preservation Master record should be created and retained. In the event that the Preservation Master is corrupted or lost, it may be recovered from a security copy.

There are a variety of media and storage options available for storing the security copy ranging from disk to tape, and online to offline storage alternatives. Optimally, the security copies should be stored in a geographically distributed location. Consideration must be made regarding how the ongoing integrity of the security copy(ies) will be monitored and maintained, especially for offline implementations. Additionally, storage solutions are becoming available that embed features offering physical security, replication, and repair.

**Key Questions for Repository with regard to the creating and managing the Security Copy(ies):**
- What will the mechanism(s) to create a security copy(ies), and on what type of media format(s) will the security copy(ies) be stored?
- What is the mechanism and frequency that the security copy(ies) will be validated?
- What is the process for restoring the Preservation Master in the event that it is lost or corrupted?
- What is the process for restoring a Security Copy in the event that it is lost or corrupted?

---


v1.0 11/2/2011
3.4.0 Generate Description Information Package (DIP)

After the archival processor has prepared and deposited the AIP, he will next prepare the Dissemination Information Package (DIP) that provides the access point for end users. Figure 19 provides a conceptual illustration of how the DIP may be derived from the AIP, including the descriptive metadata as well as additional related resources that facilitate access.

![Figure 19. Dissemination Information Package produced from Archival Information Package](image)

The process of preparing the DIP includes:

1. Preparing an access copy of the preserved digital artifacts, plus the additional archival processing tasks related to the traditional arrangement and description including creating access metadata, and incorporating the DIP into the Repository’s access information organizational structure,
2. Creating derivatives that facilitate access to the digital artifact,
3. Adding entries to the Repository’s catalog, creating / updating the finding aid,
4. Adding the digital object to the Repository’s online access application, if available, and
5. Optionally, consider publishing the geospatial dataset to a national GIS inventory system (e.g. Geo.Data.gov,\(^{55}\) or the Ramona GIS Inventory\(^{56}\)).

3.4.1 Archival Processor Creates Access/Use Copy

To minimize access to the Preservation Master, an access (or use copy) of the digital geospatial dataset is created. This will be placed in a system and repository that is accessible by end users, or surrogates retrieving the geospatial datasets on behalf of the end users (e.g. a Reference Librarian for a Repository that does not offer a publicly accessible digital collection may be able to retrieve datasets from the archival repository). The use copy may be a duplicate of the Preservation Master, offering the geospatial dataset in the same format as the Preservation Master, or, alternatively, to optimize the storage utilization for the access solution, the use copy may


be a compressed version of the geospatial dataset (e.g. orthophotos in compressed MrSID format require significantly less disk space (e.g. 1"=100’ map grid, Pixel size = .5 ft on ground: TIFF tile: 6000 KB vs. MrSID tile: 306 KB). The compressed version of the file may adequately serve the needs of most of the patrons, however, there may be some patrons who require the full-featured, high-resolution original data formats. The Repository should have a process in place to support these special requests, and this may be a rare instance of needing to access the Production Master. However, to minimize the access to the Production Master going forward, the Repository may want to add the high resolution variant as part of the Dissemination Information package as it is requested - to minimize future access to the Preservation Master. The Repository will need to work with its geospatial data contributors to determine if the contributors will produce and provide the compressed geospatial datasets as part of their submission package, or if it will be the Repository’s responsibility to produce the compressed variants. If the responsibility falls upon the archival processor, then he will require access to and training on the appropriate software tools to perform the conversions. If the Repository is automating their file transformation process, they may want to consider producing the alternate (compressed) access versions at the same time any other format transformations are being executed.

Kentucky evaluated the vector feature classes included in its geodatabase snapshots and extracted, as shapefiles, individual feature classes that have not recently experienced significant change, and then made these available through the KDLA e-archives online collection. North Carolina is publishing the shapefiles it receives to the North Carolina Digital Collections online collection. North Carolina also explored processing some orthophotos, and made the compressed MrSid files available. For its prototype access point, Utah used a web-accessible ftp site, and made both the geodatabase and shapefile formats available for download. This is being superseded by a web-based search engine that has been tied to files ingested into AXAEM, which offers download capabilities.

**Key Questions for Repository with regard to the creating and managing the Access/Use Copy:**

- What format will be used for the Access Copy? Same format at Preservation Master? An alternate, compressed format?
- What tool(s) are necessary to create the Access/Use Copy?
- Can you make arrangements with the data owner to create the Access Copy format at the same time he creates the original dataset (e.g. orthoimages - include in the vendor contract to provide both TIFF and MrSID formats for each tile).

### 3.4.2 Archival Processor creates Access Aid Derivatives

To facilitate end users’ search and retrieval tasks, and considering the significant size of some of the datasets, the archival processor may produce additional geospatial dataset derivatives that may assist users in evaluating the relevance of the geospatial dataset to their search needs without having to download and access the entire geospatial dataset through a GIS software application. Potential access derivatives include:

- **Geospatial PDF**
  - **TerraGo GeoPDF** - provide a PDF-accessible version of the dataset. Adobe extended the PDF specification so that geospatial PDFs provide access to the geospatial coordinates in the dataset. Esri ArcMap includes an export option to produce a Geospatial PDF from a geospatial dataset. TerraGo has developed their own PDF-compatible version. TerraGo offers a GeoPDF creation utility that integrates directly into the GIS tools such as Esri’s ArcMap allowing the export of one or more layers to a GeoPDF file. The following illustrates examples of PDF views for point and polygon vector datasets, and an orthoimage tile. Publishing these PDFs in the Repository’s access interface allows end users to visually preview the dataset prior to deciding to download the dataset. Figure 20 shows geospatial PDF renderings for three types of geospatial datasets including point (vector), polygon(vector), and orthoimage (raster).

---


• **Create HTML version (FGDC style) of the Repository’s customized geospatial metadata file.** The geospatial metadata file holds an extensive collection of information about the geospatial dataset. In addition to the descriptive information, that can be a significant assistance to a researcher to determine if the dataset is relevant to her search needs, the metadata file should also define the collection of data attributes that comprise the underlying dataset. The geospatial metadata file is produced in XML which is not a very convenient format for human comprehension. However, the FGDC has defined an XML stylesheet that offers a much more user-friendly view of the metadata file. Esri’s ArcCatalog and the USGS metadata parser utility offer converter utilities to generate an HTML version of the geospatial metadata file. Publishing an HTML version of the geospatial metadata file in the Repository’s access interface allows end users to preview and inspect the geospatial metadata prior to deciding to download the dataset. Figure 21 shows the base XML version and styled presentation of the geospatial metadata file.

![Figure 20. PDFs: Vector dataset - points  Vector dataset - polygons  Raster dataset - orthophoto tile](image)

<table>
<thead>
<tr>
<th><strong>Figure 20. PDFs: Vector dataset - points</strong></th>
<th><strong>Vector dataset - polygons</strong></th>
<th><strong>Raster dataset - orthophoto tile</strong></th>
</tr>
</thead>
</table>

• **Create a zip file containing all of the files in the dataset.** As the geospatial datasets are commonly comprised of several files, the end user’s download will be facilitated by providing them with the collection of files in a single package file. End users will then either need the zip utility to access and unpack the geospatial dataset files, or recent Windows Operating System versions now support accessing zip archives through the Windows File Explorer. Packaging the several geospatial dataset files is a common practice that is shared among the GIS clearinghouses that offer online downloadable geospatial datasets.

• **Create a mosaic or index shapefile for collections of orthoimages.** Orthoimages are commonly decomposed into a collection of individual tiles delivered as individual TIFF and/or MrSID files. Ideally your geospatial contributor should provide you with the shapefile index and/or a mosaic. The mosaic can serve as a useful thumbnail illustrating all of the individual tiles “stitched” together. Figure 22 provides an example of an “index” shapefile and a mosaic that shows a stitched-together collection of orthophoto tiles.

![Figure 21. XML and Styled HTML view of geospatial metadata file](image)

<table>
<thead>
<tr>
<th><strong>Figure 21. XML and Styled HTML view of geospatial metadata file</strong></th>
</tr>
</thead>
</table>

v1.0 11/2/2011
Kentucky produced a zip package for the shapefile. For shapefiles, North Carolina produced 1) a zip package for
the shapefile, 2) a geospatial PDF version of the shapefile, and 3) a styled HTML version of the geospatial
metadata file. For orthophotos North Carolina produced 1) a styled HTML version of the geospatial metadata file,
2) geospatial PDFs for a small subset of tiles, and 3) a zip package containing all of the tiles. Utah produced 1) a
zip package for the shapefile, 2) a TerraGo GeoPDF version of the shapefile, 3) a styled HTML version of the
geospatial metadata xml file, and 4) an image version (.bmp) of the graphical representation of the shapefile.

Key Questions for Repository with regard to the creating and managing the Access Derivatives:

✓ What access derivatives seem appropriate for your datasets?
✓ What tools do you need to produce these derivatives?

3.4.3 Archival Processor creates Repository Catalog Entry and Finding Aid

Repositories commonly manage a catalog that describes their holdings, and serves as a primary access point to
search and identify materials in the Repository’s collections. Traditional archival processing includes adding the
new archival record to the Repository catalog, and creating and/or modifying an archival finding aid that
describes the archival record. The Repository should employ its standard procedures for creating both. As the
catalog is often a common search access point for the Repository’s collection, if it is technically possible, the
Repository may consider making the access aids produced above accessible through the catalog interface. The
development of the finding aid will largely depend upon how the management of the geospatial datasets jibe with
the Repository’s collection organizational strategy, as finding aids typically describe whole collections rather than
individual items that are held within the Repository.

Figure 23 shows the North Carolina State Archives’ Manuscript and Archives Reference System (MARS) catalog
interface illustrating a search for GIS records. In the North Carolina MARS catalog, each individual dataset has
been cataloged. By navigating through the catalog hierarchy, the user can identify individual datasets. At this
time, the North Carolina Catalog only provides the descriptive information for the dataset, and does not provide
direct access to the dataset itself, but you could see how that might be implemented in the catalog detail page.
In contrast, the Utah State Archives catalog record itself\(^{59}\) does not mention individual datasets, but metadata from those datasets will be searchable in their new web interface.

Finding Aids are traditional tools that repositories develop to describe and provide access to the collections in their holdings. A finding aid traditionally describes a collection of materials rather than an individual item, and in addition to describing the item, who submitted it to the Repository, access and/or use restrictions, and preferred citation information, the finding aid will describe the collection and how the collection is arranged.

Figure 24 shows a prototype online finding aid from Utah’s State Archives and Records Service for a collection of county boundary records, which provides links to the FTP site where the patron could download the metadata and dataset. In this example, the finding aid provides access to the individual geospatial dataset.

---

A primary challenge for the Repository with regard to managing geospatial records within the archival domain will be devising the organizational scheme for the collection of geospatial records. The above example crafts an archival series around a particular dataset, and then provides the user with easy access to the progression or evolution of that dataset over time, as the datasets are organized chronologically in the series and associated finding aid. This organization scheme will facilitate a user or researcher who is attempting to conduct temporal analysis over time, which is understood to be one of the common reasons people want to access extant copies of datasets.

Your strategy will largely depend on if and how the user interface design will enable end users to access the individual datasets.
Key Questions for Repository with regard to adding entries to Catalog and creating Finding Aid:

✓ How will geospatial datasets be arranged/organized within your archival collection? What constitutes a Collection? What constitutes a Series?
✓ Will individual datasets be entered into your Repository Catalog? If not, how will users find out about the individual datasets? If yes, how will the geospatial datasets be entered into your Repository Catalog?
✓ What type of finding aid will you create to describe your geospatial datasets?

3.4.4 Archival Processor adds records to the Repository Digital Collection

Ideally, as geospatial datasets are digital artifacts, end users will likely want to be able search, locate, and download geospatial datasets directly from geospatial distributors’ websites. This requires a Repository to design, develop and deploy some sort of digital collection interface. The Repository can implement an access interface as part of their archival solution, or they can employ a digital collection software application platform such as CONTENTdm\(^60\) or DSpace\(^61\) that can serve other types of digital content in addition to the geospatial data.

Alternatively, a Repository may want to deploy a geospatial-specific platform, which delivers geospatial data and geospatial data manipulation tools directly to the end users as they search and retrieve archived geospatial datasets. Map-oriented search interfaces are starting to appear for geospatial portals, and may be an option for archival repositories to implement as the basis of their geospatial access interfaces.

Figure 25 shows early prototypes of the online availability of geospatial datasets through the online digital collections of the North Carolina State Archives (implemented on CONTENTdm) and the Kentucky Department for Libraries and Archives (implemented on DSpace). North Carolina publishes the individual datasets it receives and access derivatives (geospatial PDF, styled HTML of geospatial metadata, zip package of shapefile files) it produces, while Kentucky exports individual feature classes as shapefiles from the ingested geodatabase and publishes a zip package of the shapefile to their Electronic Records Archives, built on DSpace.

Figure 25. North Carolina Digital Collections implemented on CONTENTdm\(^62\)

KDLA e-archives implemented on DSpace\(^63\)


Figure 26 shows a prototype finding aid-style interface that the Utah State Archives has developed to provide access to their online downloadable datasets. Utah’s search interface, still in development, connects Solr-based search software to their AXAEM archival management database. Solr integration offers faceted searching, which is useful for limiting searches to shapefiles or geodatabases, and will offer users the ability to download files and view metadata. Figure 27 shows a prototype of the Utah State Archives search interface.
3.4.4.1 Examples of Map-based Geospatial Dataset Search Interfaces

This section presents some additional examples outside of the GeoMAPP partners’ implementations of some other online digital collections offering a map-oriented search interface for geospatial datasets. (Note: Their inclusion is strictly for example illustrative purposes, and does not imply endorsement from the GeoMAPP project.)

Figure 28. MetroGIS DataFinder Cafe66: 7 County Minneapolis - St. Paul Metropolitan Area

Figure 29. USGS Earth Explorer67 provides a visually-oriented polygonal boundary selection tool.

---

Figure 30. Maine GeoLibrary Geoportal\textsuperscript{68} offers a bounding box geography selector

Figure 31. Maine GeoPortal - Search Results Display on Map: terrain and satellite overlays (hydrants: vector point dataset)

Figure 32. Maine GeoPortal - Search Results Display on Map: terrain and satellite overlays (roads: vector: line dataset)
Key Questions for Repository with regard to adding entries to a Digital Collection:
- What technology platform and software will be the basis for your geospatial digital collection?
- How will geospatial datasets be added to the digital collection?
- What metadata is necessary to describe and/or manage the geospatial record in the digital collection?
- How will the geospatial datasets be organized and arranged/categorized in your digital collection?
- How will users search and access datasets from your digital collection?
- How will the geospatial dataset and access derivatives be presented to your users through the digital collection interface?

3.4.5 Archival Processor registers geospatial datasets with National GIS Inventory Systems

There are two significant GIS national inventory initiatives that Repositories may want to consider registering their datasets to, as these offer nationally-oriented access points to search for geospatial datasets. These inventory systems do not actually hold the geospatial datasets, rather they hold metadata records, which provide information to the end user where the dataset can be retrieved - whether it’s through online or offline access means. These inventories are equivalent in concept to the union catalogs such as WorldCat or the National Union Catalog of Manuscript Collections, offering a one-stop shop to search for geospatial datasets. At this time, these GIS inventories are predominately populated with current datasets, and they would offer an excellent access point for locating archived datasets held by geospatial archival organizations.

The (Ramona) GIS Inventory is produced by the National States Geographic Information Council (NSGIC) as a tool for states to record the availability of geospatial data across the fifty United States. The interface provides a status map which illustrates the status of geospatial data developments. The GIS Inventory provides a form-based metadata template for contributors to input their metadata. At this time, unfortunately, it is not able to import an FGDC-compliant xml metadata record. However, they do encourage contributors to provide a URL to the comprehensive FGDC metadata record. The tool is useful, however, for those organizations that have not yet developed their GIS metadata, as they can simply fill out the form to create a starter metadata record, and then download an FGDC-compliant xml file that contains their supplied metadata as a “starter” metadata file. GIS Inventory users may also, optionally, make their geospatial metadata available for harvesting by Geospatial One Stop, so as a user creates their metadata within the GIS Inventory, it may also be made available to Geospatial One Stop.

Geospatial One-Stop (GOS) was “an intergovernmental project managed by the Department of the Interior (DOI) in support of the President’s Initiative for E-government,” to improve access to geospatial information through “one complete comprehensible portal.” Geospatial One-Stop defined standards to establish common requirements to facilitate data exchange across the web for seven themes of geospatial data: geodetic control, elevation, orthoimagery, hydrography, transportation, cadastral, and government boundaries, and supports the Federal Geographic Data Committee (FGDC) metadata standards. In GOS, users had three options to publish metadata:

1) create the metadata record within GOS using a form-based metadata template,
2) upload an existing geospatial metadata .xml file, or
3) harvest metadata from an existing metadata catalog (e.g. the GIS Inventory or a Clearinghouse catalog).

---

70 Geospatial One-Stop (GOS). http://gos2.geodata.gov/wps/portal/gos As of October 1, 2011, the federal geospatial datasets in Geospatial One Stop have been integrated into to http://www.data.gov. GOS is still included here as an example of registering archived geospatial datasets with a national inventory system. Another alternative is the National States Geographic Information Council GIS Inventory site: http://gisinventory.net/. After October 2011 refer to Geo.Data.gov: http://geo.data.gov/geoportal/

v1.0 11/2/2011
A challenge with registering archived datasets to the national inventories is distinguishing these “archived” datasets from the current datasets. At this time neither of these inventory tools offer a built-in mechanism to identify or label a dataset as superseded or archival. However, you may want to use a direct labeling technique such as embedding “Archive” into the title or the archival organization name, as illustrated in Figure 33, as a mechanism to indicate an archived dataset. Also, given that users are only accessing the metadata record, it will be imperative for repositories to provide some sort of visibility and documentation within the metadata file that the metadata record has been accessioned by an archival organization and is available through the archival organization, rather than the original geospatial dataset producer and/or organization who may have submitted the geospatial dataset to the Repository for long-term preservation. This may be simply addressed by adding a statement to the Abstract metadata field that directs the reader to the archival organization, as described above in the Update Metadata section (3.3.5).

Key Questions for Repository with regard to adding geospatial entries to the National GIS Inventories:

- What National GIS Inventory Systems does your state’s GIS organization already contribute to?
- How will the geospatial metadata file be customized to reflect the dataset transfer of ownership to the archival organization?
- What National GIS Inventory Systems will you register your datasets to?

3.4.6 Archival Processor verifies catalog, Digital Collection, and Inventory access

After the dataset entries and/or datasets have been added to the access systems, they should be tested to verify that they are accessible. This can be as simple as navigating through the catalog interface, issuing some test searches, and verifying that the new items are returned in the search results lists. In the case that the datasets are being made available online, the dataset display interface and dataset download mechanisms should also be tested to verify that the datasets can be downloaded.
Key Questions for Repository with regard to verifying access:

✔ What test cases will you use to validate the access to your newly added datasets?
✔ How will you record the results of your testing?

3.4.7 Archival Processor Updates Archival Metadata Record with “related” resources

As the archival processing potentially produces several derivative resources, the archival metadata record should be updated to record the related resources that have been produced that now exist in the archival system. Basic descriptions of the related resources should likely be included, along with method of producing the derivative (e.g. how was the Geo PDF surrogate for the dataset produced), and the date the derivative was created. The derivatives are not as “at risk” as the Preservation Master, as they should always be able to be re-created from the original dataset. However, to expedite disaster recovery situations, they should probably be backed up using the organization’s standard backup procedures.

Key Questions for Repository with regard to updating the Archival Metadata Record:

✔ What metadata will you record for the derivative data?

4. Conclusions

Archival processing is a complex process, in general, and the inherent complexities of processing digital materials and complex geospatial datasets adds additional levels of complexity to the total overall process. Some of the special considerations regarding the archival processing of geospatial datasets include:

- Geospatial datasets are usually comprised of several files that must all be preserved together.
- There are a multitude of geospatial data formats for both raster and vector datasets, with varying levels of currency. Each geospatial data format may require custom archival processing processes. A Repository will benefit from preparing a Preservation Media Type Plan that identifies the formats the Repository expects to receive, and which of those formats may be reformatted to promote long-term sustainability, as well as which tools will be used to perform that data transformations.
- A comprehensive metadata standard is defined for geospatial datasets, including required fields. However, GIS developers are inconsistent in providing values for the metadata fields. This inconsistency may require intervention on the archival processor’s part to create a geospatial metadata record that can better support the long-term sustainability of the geospatial dataset.
- Special software tools, often vendor-specific, are required to interact with the geospatial datasets. However, derivatives can be produced for some formats that can be accessed with more generally available tools, such as a PDF reader.
- The archival metadata record for the geospatial dataset will require geospatial-specific metadata. However, the total metadata model should be able to be designed in such a way that allows the Repository to support a variety of digital formats.
- To facilitate and enhance access, derivative forms of the geospatial dataset may be produced as part of the archival processing, but they will also need to be managed as part of the total archival record.
- Some of the geospatial datasets, such as orthophotos, will require significant storage space, especially when considering the multiple copies suggested to sustain the long-term preservation of digital artifacts.
- The Repository will need to customize or develop from scratch a digital repository solution, the archival metadata management environment, and the geospatial access solution that allows end users to search, identify, and retrieve geospatial datasets. There are a variety of access methods that could be developed ranging from a traditional catalog search and retrieve interface, to a map-oriented search and retrieve interface. There is also a variety of geospatial functionality that could be offered ranging from basic download of the geospatial dataset, to a full featured geospatial platform that offers geospatial tools to interact with and manipulate geospatial datasets, for example to produce new maps.
Given the complexities in both the process and the geospatial materials being preserved, the Repository will benefit from diagramming and describing their archival processing process. The process is relatively modular, and can then be implemented in an iterative, cumulative fashion. Also, Repositories can begin with a relatively low-tech, manually-oriented process to begin their processing, and can evolve to more sophisticated technological solutions as resources (people, money, time, technology) are available. Regardless of the level of technical sophistication of the solution, however, Repositories will benefit from:

- establishing a sound geospatial data organizational strategy
  - e.g. what constitutes a “collection”, a “series”, and
  - how the datasets should be organized for both preservation storage and end user access - e.g. around the geospatial layers, and

- defining how the geospatial metadata file will be managed
  - e.g. will key missing technical fields be filled in, and
  - will the metadata be updated to reflect the transfer of ownership to the Repository such as access / use constraints, distributor, etc.).

Establishing a strong archival processing foundation will then facilitate the ongoing evolution and development of the total archival solution.
## APPENDIX A: Quality Assurance Process Workflow Checklist

<table>
<thead>
<tr>
<th>Step #</th>
<th>Step Description</th>
<th>Pass/Fail</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Make copy of submitted digital package</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.2</td>
<td>Run virus check</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.3</td>
<td>Validate package arrived unaltered</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.4</td>
<td>Unpack files from package</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.5</td>
<td>Validate dataset file names and dataset names comply with naming convention and or organized compliant with guidelines</td>
<td>Fail</td>
<td>A Fail may just be an indicator that the archival processor will need to do some re-arrangement when preparing the AIP.</td>
</tr>
<tr>
<td>2.6</td>
<td>Validate dataset composition and that dataset successfully renders</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.7</td>
<td>Validate geospatial metadata is complete to the Repository’s specifications</td>
<td>Fail</td>
<td>A Fail may just be an indicator that the archival processor will need to do some auto-populating when preparing the AIP.</td>
</tr>
<tr>
<td>2.8</td>
<td>Verify presence of dataset data</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.9</td>
<td>Send acknowledgement to geospatial data contributor with status.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## APPENDIX B: Process steps to create Archival Information Package (AIP)

<table>
<thead>
<tr>
<th>Step #</th>
<th>Step Description</th>
<th>Pass/Fail</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>Creating the unique identifier for the new digital object being preserved.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.2</td>
<td>Saving a copy of the original geospatial metadata file</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.3</td>
<td>Identifying the Preservation Master file(s) and format. Optionally, reformatting received digital assets to promote long-term preservation.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.4</td>
<td>Optionally, updating the geospatial metadata file to populate missing technical metadata fields - using tools such as Esri’s ArcCatalog to auto-populate these fields.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.5</td>
<td>Updating the geospatial metadata file to document the transfer of ownership to the Repository, and the Repository’s processing step(s) in preparing the dataset for preservation.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.6</td>
<td>Establish fixity value(s) for the digital asset(s) to be contained in the AIP.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.7</td>
<td>Create the Archival Metadata record for the digital asset. Populate descriptive, technical, and administrative metadata for the digital record.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.8</td>
<td>Adding the AIP to the preservation repository.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.9</td>
<td>Creating security copy(ies) of the Preservation Master record.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## APPENDIX C: Process steps to create Dissemination Information Package (DIP)

<table>
<thead>
<tr>
<th>Step #</th>
<th>Step Description</th>
<th>Pass/Fail</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1</td>
<td>Prepare an access copy of the preserved digital artifacts, plus the additional archival processing tasks related to the traditional arrangement and description, which include incorporating the DIP into the Repository’s access information organizational structure.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4.2</td>
<td>Create derivatives that facilitate access to the digital artifact</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4.3</td>
<td>Add entries to the Repository’s catalog, creating / updating the finding aid.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4.4</td>
<td>Adding the digital object to the Repository’s online access application, if available.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4.5</td>
<td>Optionally: Publish the geospatial dataset to a national GIS inventory system (e.g. Geospatial One Stop, or the Ramona GIS Inventory).</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
APPENDIX D: Summary of Key Questions for Repository

Key Questions for Repository with regard to the Ingest process:
- What mechanism will contributors use to deposit datasets?
- What geospatial metadata are contributors required to include in the geospatial metadata file?
- How should dataset files and datasets be named for transfer to the Repository?
- How should dataset files be organized for transfer to the Repository?
- How will datasets be packaged for transfer to the Repository?
- How will dataset transfers be validated (e.g., checksum, bag verifyvalid)
- How are datasets transferred to the Repository?
- Do any of the datasets have copyright or access or use restrictions that must be addressed as the datasets are transferred to the Repository?
- How will datasets be handled that do have restrictions? Where will these records be put in the archival organization structure? How will the restrictions be documented and monitored/tracked (if time-based)?

Key Questions for the Repository with regard to unpacking the submission package (SIP):
- What tools can be used to pack and unpack files?
- How to handle received datasets that are not named according to Repository submission guidelines?
- How will the Repository document the processing steps when re-organizing or re-naming dataset files. How will the lineage of a renamed dataset in the Preservation Repository be traced back to the originally ingested dataset?

Key Questions for Repositories with regard to the organization and naming verification:
- How to handle received datasets that are not organized based on Repository submission guidelines?
- How to handle received datasets that are not named according to Repository submission guidelines?
- How will the Repository document the processing steps when re-organizing or re-naming dataset files. How will the lineage of a renamed dataset in the Preservation Repository be traced back to the originally ingested dataset?

Key Questions for Repository with regard to the dataset composition and rendering verification:
- What types of datasets will the Repository be accepting? This will impact the GIS tools the Repository will need to have on hand to view/verify the datasets.
- What will be the sampling rate and sampling method for inspecting and viewing orthoimage image files?

Key Questions for Repository with regard to the geospatial metadata verification:
- What is the minimum metadata that the Repository will require for accepting geospatial datasets?
- How will the Repository validate / evaluate the completeness of the geospatial metadata? (e.g. visual inspection of the .xml? visual inspection using something like ArcCatalog to provide a user-friendly display? visual inspection with USGS metadata validator Outline view?)
- How will the Repository deal with missing metadata? Use a tool such as Esri ArcCatalog to autopopulate? Archival processor manually fills in? Automated process? (Given the nature of the geospatial datasets and their metadata, there will be fields that the Archival processor will just not have the ability to complete (e.g. data sources for the data, process steps to produce the geospatial dataset)).
- What is the Repository’s policy with regard to retaining the original geospatial metadata file as part of the archival record -- though it will not reflect the Repository ownership?

Key Questions for Repository with regard to the geospatial dataset data verification:
- What will the Repository strategy be when inspecting the dataset data?
Key Questions for Repository with regard to the unique identifier:
- What mechanism will the Repository use to associate the various related digital artifacts associated with the archival processing of geospatial datasets?
- What will the general information architecture be to manage the digital objects and the metadata throughout the archival processing life cycle?

Key Questions for Repository with regard to the original geospatial metadata file:
- Will the Repository make updates to the geospatial metadata (file) to reflect the transfer to the Repository? If not, then it’s not necessary to make an original copy of the metadata file .... though it’s probably still not a bad idea, as the ArcGIS ArcCatalog tool may be configured to automatically update the geospatial metadata file in the dataset, and the unaltered original metadata file would be lost.
- What naming convention will the Repository employ to indicate the original file?

Key Questions for Repository with regard to the creating the Preservation Master:
- What are the acceptable formats for the Preservation Master?
- What are the mappings between expected submitted formats and their companion Preservation Master format? (refer to your Media Type Preservation Plan)
- Will compound datasets be decomposed into their individual datasets (e.g. will individual feature classes in a geodatabase be extracted as individual shapefile datasets; will individual raster images be exported separately)?
- What tools and processes will be employed to transform a geospatial dataset received in one format to another format to produce the Preservation Master?

Key Questions for Repository with regard to the updating missing technical metadata fields:
- What are the technical metadata fields that the Repository will require?
- What tools, techniques, or strategies will the Repository employ to populate the missing technical fields? Note: One possible option is to reject a dataset at submission time that does not have the requisite technical metadata fields populated ... especially as these are especially crucial to the long-term preservability of the geospatial dataset.
- What updates or modifications will the Repository make on the geospatial metadata file?

Key Questions for Repository with regard to the updating the geospatial metadata file:
- What is the Repository’s strategy for conveying that the geospatial dataset ownership has transferred to the Repository within the metadata file?
- What updates or modifications will the Repository make on the geospatial metadata file?

Key Questions for Repository with regard to establishing digital asset fixity:
- What method will the Repository use to compute, store, re-compute, and compare fixity values for both the Preservation Master and Preservation Master Security Copy(ies)?
- Will the Repository store the fixity values for the geospatial dataset files within the archival metadata?
- Will the Repository store the fixity values for the geospatial dataset files within the AIP itself?
- How would someone perform an audit check to view and/or validate that a file’s integrity is intact?

Key Questions for Repository with regard to the updating the geospatial metadata file:
- What technical mechanism will be used to record and store the archival metadata?
- What metadata elements comprise the archival metadata record?
- What geospatial metadata fields will be extracted and included in the archival metadata record?
- Will the Repository create a modified geospatial metadata file that
  a) has missing technical metadata fields filled in?
  b) reflects transfer of ownership of the geospatial dataset to the Repository organization?
- How will the Repository record its processing steps in the archival metadata record (e.g. dataset format transformations, geospatial metadata modification or customizations)
- Will the fixity record for each of the files be recorded as part of the metadata record?
Key Questions for Repository with regard to the updating the geospatial metadata file:
✓ How will the Repository structure and store its Archival Information Packages?
✓ Will the AIP contain a single dataset or a collection of dataset?
✓ Will the Repository use a special packaging mechanism(e.g. METS, BagIt)?
✓ Who will have access to the Archival Information Packages? What type of access (read, write)?
✓ What access policies and rights will be configured for the AIP? Who can access it? For what purpose?

Key Questions for Repository with regard to the creating and managing the Access/Use Copy:
✓ What format will be used for the Access Copy? Same format at Preservation Master? An alternate, compressed format?
✓ What tool(s) are necessary to create the Access/Use Copy?
✓ Can you make arrangements with the data owner to create the Access Copy format at the same time he creates the original dataset (e.g. orthoimages - include in the vendor contract to provide both TIFF and MrSID formats for each tile).

Key Questions for Repository with regard to the creating and managing the Access Derivatives:
✓ What access derivatives seem appropriate for your datasets?
✓ What tools do you need to produce the derivatives?

Key Questions for Repository with regard to adding entries to Catalog and creating Finding Aid:
✓ How will geospatial datasets be arranged / organized within your archival collection? What constitutes a Collection? What constitutes a Series?
✓ Will individual datasets be entered into your Repository Catalog? If not, how will users find out about the individual datasets? If yes, how will the geospatial datasets be entered into your Repository Catalog?
✓ What type of finding aid will you create to describe your geospatial datasets?

Key Questions for Repository with regard to adding entries to a Digital Collection:
✓ What technology platform and software will be the basis for your geospatial digital collection?
✓ How will geospatial datasets be added to the digital collection?
✓ What metadata is necessary to describe and/or manage the geospatial record in the digital collection?
✓ How will the geospatial datasets be organized and arranged/categorized in your digital collection?
✓ How will users search and access datasets from your digital collection?
✓ How will the geospatial dataset and access derivatives be presented to your users through the digital collection interface?

Key Questions for Repository with regard to adding geospatial entries to the National GIS Inventories:
✓ What National GIS Inventory Systems does your state’s GIS organization already contribute to?
✓ How will the geospatial metadata file be customized to reflect the dataset transfer of ownership to the archival organization?
✓ What National GIS Inventory Systems will you register your datasets to?

Key Questions for Repository with regard to verifying access:
✓ What test cases will you use to validate the access to your newly added datasets?
✓ How will you record the results of your testing?

Key Questions for Repository with regard to updating the Archival Metadata Record:
✓ What metadata will you record for the derivative data?
APPENDIX E: Esri ArcCatalog - Auto-update Metadata Fields

The items noted in blue and with an asterisk in the following are geospatial metadata fields that Esri ArcCatalog will automatically update, extracting, where possible, from the dataset’s properties, or from the dataset .shp.xml metadata file.

Identification Information

Citation

Citation Information

Title*
File or table name*
Geospatial data presentation form*

Description

Language of dataset*

Spatial domain:

Bounding coordinates

Left bounding coordinate*
Right bounding coordinate*
Top bounding coordinate*
Bottom bounding coordinate*

Local bounding coordinates

Left bounding coordinate*
Right bounding coordinate*
Top bounding coordinate*
Bottom bounding coordinate*

Native dataset format*
Native dataset environment*

Data Quality Information

Spatial Data Organization Information

Direct spatial reference method*

Point and vector object information:

SDTS terms description:

Name:* 
SDTS point and vector object type*
Point and vector object count*

Spatial Reference Information

Horizontal coordinate system definition:

Coordinate system name:

Projected coordinate system name*
Geographic coordinate system name*
Planar:

Grid coordinate system:
  Grid coordinate system name*
State Plane Coordinate System:
  SPCS zone identifier*
Lambert conformal conic:
  Standard parallel*
  Standard parallel*
  Longitude of central meridian*
  Latitude of projection origin*
  False easting*
  False northing*

Planar coordinate information:
  Planar coordinate encoding method*
  Coordinate representation:
    Abscissa resolution*
    Ordinate resolution*
    Planar distance units*

Geodetic model:
  Horizontal datum name*
  Ellipsoid name*
  Semi-major axis*
  Denominator of flattening ratio*

Entity and Attribute Information

Detailed description:
  Name*

Entity type:
  Entity type label*
  Entity type type*
  Entity type count*

Attribute: (repeated for each attribute in an entity)
  Attribute label*
  Attribute alias*
  Attribute definition*
  Attribute definition source* (for Esri-sourced attributes)
  Attribute type*
  Attribute width*
  Attribute precision*
  Attribute scale*
  Attribute number of decimals*

Attribute domain values
  Unrepresentable domain*

Distribution Information

Standard Order Process:

v1.0 11/2/2011
Digital form:
  Digital transfer information:
    Transfer size*
    Dataset size*

Metadata Reference Information

  Metadata date*
  Language of metadata*
  Metadata standard name*
  Metadata standard version*
  Metadata time convention*
APPENDIX F: Digital Repository Platforms

The following are provided for informational purposes only. No endorsement of any of these technologies or solutions is implied by inclusion in this list.

Archivematica: http://archivematica.org
Archivematica is a comprehensive digital preservation system. Archivematica uses a micro-services design pattern to provide an integrated suite of free and open-source tools that allows users to process digital objects from ingest to access in compliance with the ISO-OAIS functional model. Archivematica is free and open source software.

APPX Archives Enterprise Manager (AXAEM): http://www.axaem.com/
AXAEM is a comprehensive solution applicable to a variety of archives organizations. AXAEM is a content management system designed specifically for state and other archives responsible for managing records and collections. AXAEM is available from APPX Software, Inc. under an Open Source license.
Note: AXAEM implementation documentation from is available from the Utah State Archives at: http://archives.utah.gov/axaem/axaem.html

Carolina Digital Repository’s Curator’s Workbench:
A desktop tool to manage the submission workflow and processing for large collections with custom metadata. The process produces a METS manifest, and the digital repository technology base is Fedora and iRODS Grid. Curator’s Workbench is free and open source software hosted at github.com

CONTENTdm Digital Collection Management Software by OCLC: http://www.contentdm.com/
CONTENTdm is a single software solution that handles storage, management and delivery of your library’s digital collection.

DSpace: http://www.dspace.org/introducing
An Open Archives Initiative (OAI)-compliant open-source software released by MIT for archiving eprints and other kinds of academic content. DSpace preserves and enables easy and open access to all types of digital content including text, images, moving images, mpegs and data sets.

Esri Geoportal Server is a free open source product that enables discovery and use of geospatial resources including datasets, raster, and Web services. It helps organizations manage and publish metadata for their geospatial resources to let users discover and connect to those resources. The Geoportal Server supports standards-based clearinghouse and metadata discovery applications.

Fedora Repository Project: http://www.fedora-commons.org
Fedora (Flexible Extensible Digital Object Repository Architecture) was originally developed by researchers at Cornell University as an architecture for storing, managing, and accessing digital content in the form of digital objects inspire by the Kahn and Wilensky framework. Fedora defines a set of abstractions for expressing digital objects, asserting relationships among digital objects, and linking “behaviors” (i.e. services) to digital objects. The Fedora Repository Project (i.e., Fedora) implements the Fedora abstractions in a robust open source software system.

iRODS (Integrated Rule-Oriented Data System): http://www.irods.org
iRODS, the Integrated Rule-Oriented Data System, is a data grid software system developed by the Data Intensive Cyber Environments research group and collaborators. iRODS management policies (sets of assertions communities make about their digital collections) are characterized in iRODS Rules and state
information. At the iRODS core, a Rule Engine interprets the Rules to decide how the system is to respond to various requests and conditions. iRODS is open source under a BSD license.


Tessella Safety Deposit Box (SDB) provides a platform and services that conform to the Open Archives Information Systems (OAIS) reference model, and allows you to store and preserve critical digital information in a highly reliable, yet accessible manner. SDB provides support for Ingest, Preservation, Access, Data Management and Administration, and Storage.