Utah's Data Transfer Procedures.

## 1. Background/Introduction

Utah's archiving process began to take form in June 2008. At that time, the Automated Geographic Reference Center (AGRC) entered into a partnership with the State Archives to purchase a new server to be located in the Richfield Utah Data Center and to share the AGRC's server in Salt Lake City Data Center. These two data centers are approximately 2 1/2 hours apart by car. These servers run on the Suse Linux Enterprise Server 10.2 operating system. There is not a set storage capacity at this time. Capacity will be added as needed, with currently a limited storage set for imagery. This server will house all the geospatial vector data and eventually all imagery submitted to the Archives for retention.

Data submitted to the Archives will be placed in a directory on the AGRC's Salt Lake FTP site. Submitted data will go through a validation, explained in Section 3b. When this validation is completed, the data will be ready to be "pushed" down to the Archives' FTP site in Richfield. The open source software RSYNC on the Salt Lake FTP server will be used to transfer the data to the server in Richfield for permanent retention. The transfer will include the utilization of the CheckSum feature contained within RSYNC. AGRC will also be testing BAGit from the Library of Congress for the interstate transfer. The AGRC will be the administrator of both the Salt Lake and Richfield servers, but once the validated data has been pushed to the Richfield server it will be under full control of the Archives.

## 2. Dataset Selection

The GeoMAPP Content Lifecycle and Data Transfer group provided guidelines for the types of datasets to be picked for the archiving demonstrations that would exhibit a broad variety of spatial data. These were:

- Local Government Datasets
- Ortho-imagery
- Centralized datasets (both Framework and Non-Framework data)
- Project Files
- Digitized Maps

Utah's targeted local government data come from two Wasatch Front urban counties: Salt Lake and Davis. These counties have extensive geospatial databases with complete datasets to choose from. The data from both counties are:

- Parcel data sets ~ 200MB each
- Zoning data sets ~ 18MB each
- Municipality data sets ~ 2 MB each

Salt Lake County is part of the USGS 128 metropolitan areas that are required to have 1 ft. ortho-imagery produced every 3 years. Because of this program, the AGRC has multiple years of ortho-imagery for Salt Lake County. Utah's team chose to archive 3 sets of FSA NAIP Imagery of the Salt Lake County, circa 1977, 1990, and 2006 ~ 200-300MB for each year. This ortho-imagery represents three different time periods of growth in the Salt Lake Valley.

The selection of centralized datasets focused on framework datasets common between the partner states and non-framework datasets that were more Utah-centric: unique to a state that has desert and rural themes.
Framework datasets:

- State ownership ~12MB
- Municipalities ~ 4MB
- Centerline Streets ~ 320MB
- NHD 24K Streams ~ 250MB

Non-framework datasets:

- Biota Datasets
    - Southwest ReGAP ~ 1.5MB
    - Water Related Landuse ~ 50MB
- Environmental Datasets
    - Brownfield Projects ~ 2MB
    - LUST Open Tanks ~ 3MB
    - National Priority List ~ 2.5MB
- Inland Waters Datasets:
    - Wetlands ~ 30MB
    - Watersheds ~ 30MB
    - Great Salt Lake Shoreline Flooding ~ 3MB
- Structures Datasets:
    - Hospitals ~ 1.5MB
    - Fire Stations ~ 1MB
    - Police Stations ~ 1MB
- Transportation Datasets:
    - Utah Railroads ~ 3 MB
    - Utah State Fuel Sites ~ 1MB

Project Files: Utah chose to archive a sample project that was unique to Utah, but could be applied to other states for analysis. This project helped to analyze the "drug zone free" law in Utah.

- Drug Zone Free Law Analysis ~>1GB

Scanned/Geo-referenced/Digitized Maps:
Utah has many different sets of scanned maps, but chose to archive a selection of USGS 7.5 quad maps over the Salt Lake County – 22 -- 7.5 Quads ~ 5MB per quad = ~ 110MB

**3. Geospatial Data Transfer Preparation**

**a. Series Schedule Prep and Approval**

Before any geo-spatial data can be submitted to State Archives, there are specific archival tasks to be followed. These are:

- An Agency's State Records Analyst meets with the GIS coordinator of the agency to check if there is a general retention schedule the data could be tied to. General retention schedules are categorized either by government type (county, municipal, etc.) or by specific agency.
- These general retention schedules have been approved by the State Records Committee and a general retention schedule can quickly be tied to an agency's data. If none of general retention schedules fit the description of the data, an agency specific retention schedule is created to provide more detail about specific records.
- This agency specific retention schedule it will include:
    - Information on the data
    - The retention information, including:
        - Timeframe for media migration
        - Timeframe for geospatial format migration
        - Retention schedule of the data
    - All new retention schedules must be approved by the State Records Committee before any data can be tied to them.
    - Once the new schedule is approved, the unique agency retention schedule should be linked to a general retention schedule that categorizes records according to ISO standard category theme name.
        - Linking agency specific data theme names to ISO standard category themes aids in the cross referencing in data research.

**b. Geospatial Data Validation**

The validation process for geospatial data submitted to the State Archives will be the same whether the data comes from the AGRC or if it comes from another State or Local agency. This validation will take place in the Archives' directory on the AGRC FTP site. Data will be submitted or transferred to the Archives' directory using the agency's preferred transfer media or method. After the transfer, the data must pass this validation before it can formally be added to the Archives database.

- Opening the data in ArcGIS to ensure the data file is spatially valid (opens and is readable within a GIS environment.
- Check for a defined spatial projection including:
    - Type of projection i.e. State Plane, Geographic or UTM
    - Datum
    - Linear Unit
- Using the ArcCatalog metadata editor, the data is checked for complete FGDC compliant Metadata, including the following information:
    - All areas marked "Required" within the editor
    - Data Contact
    - Attribute Accuracy
    - Positional Accuracy
    - Data Source Information
    - Process Steps
    - Complete Definitions of all Attributes
    - Distributor
    - Metadata Contact
- If the metadata is non-existent or incomplete, the owner or steward of the data will be contacted so that the metadata can be completed to meet FGDC standards.
- Once the data is spatially validated, a crosswalk is used to associate the submitted data title with the ISO nomenclature adopted by the Archives for geospatial series.
- Run VB script to create a geo-PDF, a shapefile, and a file-based geodatabase for each dataset.
    - Each dataset will be labeled with the name and date of the dataset. This will be extracted from the data's metadata .xml.
    - Coordinates will be extracted from the data's metadata .xml.
    - The data will be placed in a folder labeled with the date of the data.
    - This folder will then be placed in another folder labeled with ISO sub-category theme and the series retention schedule number.

- This folder will then be placed in the last folder labeled with the basic ISO nomenclature adopted by the Archives. (See Attached Crosswalk List of Archive Categories and SGID Categories)
- These folders will be created before any data is transferred to the Archives FTP directory.

**See diagram below:**

ftp://ftp.agrc.state.ut.us/Archives Archive finding aid harvests metadata for information and coordinates.

- Biota
- Boundaries
  - MunicipalBoundaries26846-Series Schedule-26846 – contains retention information on media and format migration and distribution.
    - MunicipalBoundaries1998 Date of data
      - GeoPDF
      - Metadata
      - Geodatabase
      - Shapefiles
    - MunicipalBoundaries1999
    - MunicipalBoundaries2000
    - MunicipalBoundaries2001
    - Etc.
  - CountyBoundaries26845
    - CountyBoundaries2002
    - CountyBoundaries2003
    - CountyBoundariesPre2003
    - CountyBoundariesundated
- Climatology

**c. Imagery and Raster**

The SGID imagery will be extracted from the AGRC's FTP imagery directories and moved to the Archives' directory on the AGRC's FTP site. There the data will go through a short validation before it is ready to be transferred to Richfield. This spatial validation will include:

- Opening the imagery in ArcGIS to check for corruption
- Checking for complete FGDC metadata including flight and publishing date.

After the imagery has been spatially validated, a script will be run that will add the series schedule number to the name of the imagery file. The imagery file will then be zipped and placed in the correct folder in the Archives directory on the Salt Lake FTP site. The folder format for the imagery and raster area on the Archives' FTP site will be:

- Type of imagery (i.e. DOQ, NAIP, NED) Series schedule number - containing retention information on media and format migration and distribution.
  - Imagery type (compressed): the files will be in MrSID, TIFF, or JPEG formats. -
    - Date of imagery – included in zipped file title

In the future, once more storage space has been acquired the imagery will also be available in an uncompressed format.

Only imagery from the SGID has been looked at for archiving at this time. When there is sufficient funding to purchase a larger server, the storage of imagery from other agencies will be developed. The process of validation and ingesting imagery into the Archives' permanent storage and onto the FTP site in Richfield will be refined.

**d. Projects**

Projects containing many geo-spatial datasets, models for analysis, map documents (.mxd), documentation explaining the procedures and results of the project will have a project folder created to hold all of these "parts" of the project. All geospatial datasets will go through the same archive validation process previously outlined in Section 3b. If there is imagery or raster data used in the project, it too, would go through the imagery validation described in Section 3c. By following these procedures all geospatial data and imagery will be documented by FGDC metadata. If there are any models built for the analysis, the model builder documentation should be used and checked for clarity. Map documents (.mxd) should have Dublin Core metadata written about each one with a simple explanation of what will be found within the .mxd. Once all the parts of the project have been well documented a single Dublin Core metadata document should be prepared outside the project folder. This metadata should contain the general information on the project, its outcome and the extent coordinates of the project. The folder and the Dublin Core metadata should then be zipped together and ready to be placed in the projects folder in the Archives' FTP site.

**5. Data Receipt by Archives**

The data has been spatially validated, extracted into a geo-PDF, a shapefile, and geodatabase and are now in the correct folder in the Archives directory on Salt Lake FTP site. Before the data is "pushed" down to Richfield, a second script will be run to create a separate metadata .xml file containing the name of the file, a descriptive title (abstract), creation dates, file size, whether it is a shapefile or geodatabase, scale or resolution, projection, datum, extent coordinates, and keywords for a search tool. After the folders have been pushed down to the Richfield FTP site, the script will extract the URL of each geo-PDF, the zipped shapefile and geodatabase moved and added to the second metadata .xml. This second metadata .xml will be sent to Archives where it can be used to record and store all the information about every submitted dataset in a finding aid.

**a. CheckSums:**

The open source product called RSYNC is used to transfer the files between the two AGRC servers has a process that takes place over a Secure Shell (SSH) connection. SSH encrypts the file on the sending end and de-encrypts it on the receiving end, thus checking the integrity of the file. RSYNC does have a CheckSum flag that can be

turned on. At this time, though, AGRC has deemed this to be a redundant action in light of the check that SSH does on the file being transferred.

However, this process and decision not to use it, does not take into account interstate transfers. AGRC will be testing BAGit from the Library of Congress for the interstate transfer.

**b. Frequency of Ingest**

With each development of a series retention schedule, the owner or steward of the data will work with the analyst to set up the frequency of ingestion of data. The SGID will be on an annual ingestion schedule except for the few datasets that change at a high rate. These datasets would include parcels, administrative boundaries, tax entities, zoning districts, and others. It will be up to the analyst and the GIS coordinator to sit down together and look at the needs of the agency and the public before they develop an ingestion schedule.

**6. Demonstration Validation**

**a. Intrastate data transfer:**

1. After the geospatial vector data has been transferred to the Archives' FTP site, a sample set of the data will be opened and checked, using same validation checks (Section 3b) applied to the data when it was first submitted to the Archives. If this sample set of data passes these checks, the transfer will be considered accurate and complete. When the imagery is submitted and transferred to the Archives' Richfield FTP site, a sampling of the imagery will be downloaded, opened and examined to check for metadata and imagery corruption. If the sample imagery passes this validation, the transfer will be considered complete. The Archives will set up schedules to do checksums to confirm the data is unchanged over time.

2. To show the efficiency of the finding aids created by the Archives, researchers with a geospatial background and without a geospatial background will be asked to research geospatial data using these aids. A survey will be taken asking if the finding aids helped in locating the data on the Archives' FTP site and if after finding, downloading and opening the geospatial data, there was sufficient metadata explaining the data.

3. A second test of the intrastate transfer will involve the project data file. After the transfer, again researchers with geospatial knowledge and those not acquainted with the data or the methodology will be asked to use the finding aids to find and download the project files. After the researchers have opened and examined the data they will be surveyed to give their opinion of the documentation found in the project file.

**b. Interstate Transfers:**

1. During the interstate transfer of the Utah's pre-selected geospatial data, the Library of Congress' new open source BAGit transfer application will be used and tested to insure the data is transferred successfully. In using

the BAGit application, all the parts of the spatial data, imagery or project files are recorded on the BAGit contents list found in each BAGit folder. This list is made when the data is loaded into the BAGit file before the transfer. A declaration text is the third file found in the BAGit transfer folder. The BAGit declaration text is created by the sending computer when the data is loaded into the transfer file, recording information the receiving computer should find in the folder. When the file is transferred, the receiving computer will read the declaration text file; checksum the BAGit list, confirming that all the parts of the data have been received and that none were corrupted during the transfer. The BAGit interstate transfers will be from a Linux file system to a Windows file system. Although it is believed that checksum processes done between these two OS are independent, any issues that may arise will be documented.

2. After the interstate transfer and the checksums have been confirmed and completed, the state receiving Utah's data should open and examine the documents, data and imagery. This should be done by both researchers with geospatial experience and those that do not have a geospatial background. After examining the transferred data, another survey will be given asking for information about the transferred data. The results of this survey will be to document information on whether the data was geospatially correct after the transfer, did the Dublin Core and FGDC Metadata explain the data well enough that it could be used to by a different state, and after opening the project folder, was the metadata and documentation adequate for another state to use the information.

**7. Discovery and Research**

a. The first geospatial data stored by the Archives will be available through the Archives' Richfield FTP site. The data will be found in the cascading folder model described in Section 3b this paper. A client will use the Archive's indexing site and their finding aids to determine what geospatial data is available and where it is on the Richfield FTP site.

b. In the near future, the Archives hopes to work with the AGRC in the development of a map application that will allow for searching not only the geospatial data on the Archives' FTP site, but also the other scanned data and hard copy records which Archives has tagged with the coordinate extent of a certain geographic area.

1. This application will have several historically dated Utah maps associated with it. These maps will range from the early territorial boundaries of Utah to a present day state boundary map. A researcher will be able to be selected any of these maps to draw and select an area of interest. The drawn extent of the area of interest would send coordinates back to the Archives' finding aids allowing the researcher to search and return all digital data, including the geospatial data, and information on the archived hard copy data pertaining to that extent. This application is still in draft form, but it is hoped that it will be available in 2010.

**Appendix 1**

***Crosswalk from Archive ISO schedules to SGID Nomenclature***

| General retention schedules | General Series Number | SGID Classification |
|---|---|---|
| Biota | 26838 | Bioscience |
| Boundaries | 26813 | Boundaries |
| Climatology, Meteorology & atmosphere | 26814 | Energy |
| Elevation | 26821 | Elevation |
| Environment | 26835 | Environment |
| Geo-scientific | 26793 | Energy |
| Geo-scientific | 26793 | Geoscience |
| Health | 26822 | Health |
| Imagery, Base Maps, & Earth | 26823 | Indices |
| Inland Water | 26820 | Water |
| Location | 26826 | Location |
| Planning and Cadastre | 26815 | Cadastre |
| Planning and Cadastre | 26815 | Planning |
| Society | 26816 | Demographic |
| Society | 26816 | Economy |
| Society | 26816 | Political |
| Society | 26816 | Recreation |
| Transportation | 26828 | History |
| Transportation | 26828 | Transportation |
| Utilities and Communication | 26829 | Utilities |

Data Transfer Demonstration Recap:

**Issues Encountered with Data Transfer**

1. Metadata inclusion challenges. This process of checking the data after the transfer resulted in discovery of missing metadata in the source files.
2. Challenges were encountered with consistency in zipping files, ESRI format file geodatabases were not zipped up, but shapefiles were.

3. The GeoPDF created for each feature class did not display “friendly” attributes, it was as though it was showing the ESRI file format internal Ids. The GeoPDF creation process needs to be analyzed to understand where the issues lie.

4. RSYNC was used to transfer the data from SLC to Richfield. It uses SSH which checks for file integrity. It has a checksum flag that can be turned on, but currently we are not utilizing it. Though by turning it on would allow us to test the file against the checksum over time. This option needs to be explored more.

5. A project file “EnhancedPenalty26969” was transferred, however this file was very very lacking in documentation to describe all the myriad of files associated with a project.

6. There is currently no Virus scan on files uploaded to the Salt Lake FTP site or when the file is transferred to the Richfield FTP. This issue needs further research to find the best time from a technology perspective to scan for viruses.

7. Naming conventions are a bit of an issue, with the State Geographic Information Database names recently changed. There needs to be more discussion when data is archived, which name does it contain, the old filename or new filename etc.

8. For archiving purposes, it's better not to store compressed (zipped) files, although they are mighty handy when capturing shape files. North Carolina is getting both a zipped and non-zipped version--one for public downloading, and one for storing.

9. There needs to be a better line of communication as to what is needed on an ftp site for the customers. Is it possible to automate any of the steps in the process? I think it would eliminate some of the inconsistencies, which are operator error.

10. A ReadMe file needs to be created whenever a project file is placed in the Archive. This file should include information on what files, etc. are included in the project.

11. A GeoPDF usually takes about 10 minutes to create, depending on the size of the dataset.
ArcMap is opened and the dataset is added. The dataset is then exported out to a GeoPDF. Once this process is complete, the GeoPDF is opened to verify all of the data was exported.

12. It would be great if the steps to create a GeoPDF, geodatabase, and shapefile were automated. This would lessen the time needed and the risk of error.