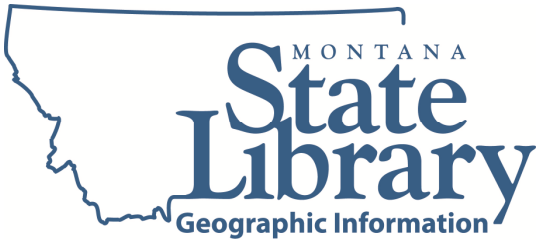
Montana State Library
Spatial Data Transfer Design



Prepared for GeoMAPP, December 17, 2011
by Diane Papineau, Gerry Daumiller, Evan Hammer, Jennie Stapp, and Grant Austin

Introduction

The Montana State Library (MSL) was integrated into the Geospatial Multistate Archive and Preservation Partnership (GeoMAPP) as a full partner in February 2011 after a little over a year as a GeoMAPP informational partner. MSL has served as the state GIS clearinghouse for almost two decades. In this capacity, MSL manages a large spatial data collection and makes GIS data available via web applications, web map services, and as downloadable data. For years MSL has maintained a GIS data list, and in 2008 MSL launched the Montana GIS Portal[1] based on the ESRI GeoPortal Toolkit. Though MSL does accept spatial data into its collection from state agencies and other data creators, Montana's most frequently-updated spatial datasets are compiled and created inside the State Library. MSL is the theme steward for nine of the fourteen framework layers that comprise the Montana Spatial Data Infrastructure (MSDI).

MSL has long been recognized as the "archives" for GIS data, though informally executed. Prior to joining GeoMAPP, MSL's process to archive GIS data was to simply not throw any data away. As MSL joined the GeoMAPP project, MSL chose to take a "library collection development policy" approach to managing a GIS data archive rather than a "records management" approach which makes use of records retention schedules often used for state government document management. Guidance offered via a library collection development policy defines what MSL should collect based on state statute and user needs and provides in policy the support to weed or permanently remove data as appropriate.

MSL manages its own internal data center that supports, in part, the GIS clearinghouse and the informal GIS archive. Storage is comprised of a file system and database management system (DBMS) for managing active datasets and a file structure created on a Storage Area Network (SAN) for Dark Archive storage. All data storage is backed up to State of Montana standards.

[1] Montana GIS Portal: http://gisportal.msl.mt.gov

Note: This data transfer demonstration effort focuses only on archiving spatial data stored and served by MSL. It does not cover archiving MSL's electronic maps, paper maps, map project files, text documents, or web pages.

Planning to Incorporate Preservation Practices

MSL staff members spent a significant portion of 2011 reviewing GeoMAPP documentation and envisioning how to apply archiving practices to Montana's geospatial clearinghouse workflows. The GeoMAPP data transfer best practice and design documentation taught MSL about practices and challenges applicable to archiving spatial data. It also spawned an objective critique of MSL's existing data storage, data access tools, and management processes in light of what is needed to consistently and professionally archive spatial data.

One early task involved clarifying six important terms and how they would define and support a revised dataset management process that includes archiving as MSL moves forward:

> *Data Collection*—data MSL stores and preserves that fits the draft Collection Development Policy. The collection may include multiple copies of data in different forms stored in the Dark Archive as well as the Accessible Archive and the Active Store.
>
> *Dark Archive*—a physical location on MSL's SAN, storing copies of datasets MSL accepts into its data collection. The Dark Archive stores the preservation copies of electronic maps and GIS map projects. The Dark Archive may not permanently store spatial data that is archived by other parties such as the National Agricultural Imagery Program (NAIP) imagery and wetlands. For preservation purposes, MSL may initially store preservation copies of these data to ensure stable data discovery.
>
> *Security Dark Archive*—a physical location in the State of Montana's Data Center (offsite from MSL) storing an exact copy of the Dark Archive. The integrity of the data in the Security Dark Archive will be checked using the same processes applied to the Dark Archive at MSL. The Security Dark Archive and the Dark Archive at MSL will each serve as the restore source for each other.
>
> *Accessible Archive*—a physical storage location offering superseded data made easily accessible to patrons via the Internet. This is older data that is

valuable to many, justifying online, self-service access (i.e. superseded Cadastral, superseded Land Cover, etc.). The storage and service vendor for this Accessible Archive is yet to be determined.

Active Store—a group of MSL SAN storage locations holding the most active data (primarily the most current data) made available in different forms primarily for online, self-service access. Different storage devices and locations will serve Active Store data: for internal use in an SDE database, on portable hard drives for large dataset manual delivery, on servers for online MSL mapping applications, as downloadable files, and as web mapping services for patrons. Data in Active Store locations is data that will be used most frequently by library patrons (internal and external).

Clearinghouse—a group of data discovery tools and resources, including the Montana GIS Portal, web mapping applications, web map services, webpages offering data download, as well as staff time for manual packaging of large datasets for patrons.

The next effort undertaken as MSL moved toward the data transfer demonstration involved envisioning a new, streamlined dataset management process that incorporates formal archiving practices and concepts. To start this process, MSL documented existing data repositories and file flow processes. This enabled the modeling of natural spatial data clearinghouse workflows that could be mimicked in a new management system. With this draft GIS dataset management workflow in hand (Appendix A), MSL wrote initial technical requirements for a new GIS Dataset Management System to automate as much of this dataset management work as possible. This tool will assist with spatial data and metadata ingest, metadata updates, metadata publishing, dataset storage and management, dataset integrity auditing.

Considering the abbreviated period of time involved with GeoMAPP (less than one year), it was not possible to create, test, and deploy a new management system prior to completing the GeoMAPP data transfer demonstration. Instead, MSL created a prototype environment that would mimic the new management system proposed. MSL created a Dark Archive file structure on the SAN that reflects the pattern by which data is accessioned into MSL's collection. (Figure 1, right). MSL also tested and chose to use the Library of Congress's tool "Bagger"[2] for transferring data to the Dark Archive and validating data transfer. This GUI-driven tool (Figure 1, left) automates many archival

[2] Bagger available via http://sourceforge.net/projects/loc-xferutils/files/loc-bagger

processes, including data packaging, creating a manifest of the contents, creating a checksum, and validating file integrity.

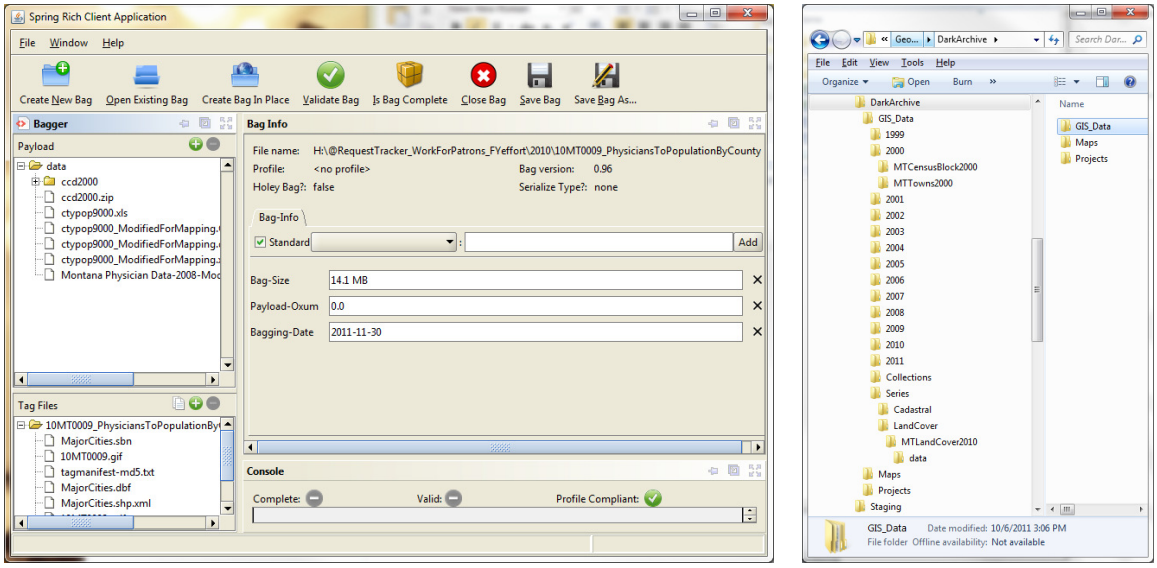


Figure 1: Bagger User Interface (left) and MSL GIS Dark Archive File Structure (right)

MSL chose to archive data in its original formats rather than define an archival format because each data format offers unique functionality. Though MSL realizes that shapefiles are usable in many GIS software programs, the team did not want to lose functionality inherent in original file types by converting all data to shapefiles as an archive standard type. The datasets selected for transfer included:

- Series Data—Landcover file geodatabase (raster data, 100MB)
- Stand-Alone Data—Montana Towns shapefile (.08MB)
- Stand-Alone Data—Montana Census Blocks shapefile (57MB)

MSL established a packaging and file naming convention in the form of <*extent*><*theme*><*time period*> (i.e. MTcensusblocks2010) and created a GIS dataset management spreadsheet to mimic the proposed system's database. This spreadsheet had the following metadata fields defined:

FGDC metadata—Title, Time Period, Originator, Publisher, Other Online Location, Larger Work Citation

New Archival metadata—Date Archived, Zipped Megabytes, Checksum, Last Archive Review Date, Accessible Archive Online Location

New Administrative metadata—Clearinghouse Online Location, File Location, Data Format, Data Format Version, SIP Metadata URL/network location, DIP Metadata, Source Data (if derived), Derivatives, Compression, Status

Note: User metadata records, including SIP metadata and any additional archive metadata, will be exposed to patrons for data discovery through a linkage between the management database and the Montana GIS Portal database. The complete metadata recorded in the GIS Dataset Management System database, including administrative metadata, aids in dataset management and archiving, not patron data discovery.

Data Transfer Demonstration

With these initial planning tasks completed, MSL proceeded with the actual data transfer demonstration to test these ideas and record any problems encountered. For the demonstration, MSL worked through the following procedure and modified it with each use. The last dataset that was transferred used the following process, representing MSL's draft data transfer process to date:

1. Select data for transfer that meets the requirements of the draft Collection Development Policy and place it in the Dark Archive staging folder.

2. Run a virus check on the data using ESET NOD32. Even though NOD32 is always checking for viruses on network storage devices, it is best to execute this virus check specifically on the file before placing the data in the Dark Archive.

3. Preserve a copy of the original metadata adding the word "original" to the file name. Edit the metadata as necessary to describe the data set's existence in the archive and create an archival metadata record that passes the Montana GIS Portal's validity requirements.

4. Pre-populate some metadata fields in the dataset management spreadsheet, pulling information from the data's FGDC metadata record (Title, Time Period, Originator, Publisher, Other Online Location, Larger Work Citation).

5. Using Windows Explorer and Bagger, package the data for transfer to the Dark Archive:

    a. Place the dataset, the original metadata file, and the archival metadata file in a zip file using the file naming convention.

b. Open Bagger and create a new bag. Set parameters as version 9.6 with no profile.

c. Using Bagger's green plus icon (Figure 1), navigate to and select the zip file for bagging then click Open.

d. Save the bag: File>Save Bag As.

e. In the Save In text box, use the browse button to save the bag in the archives. In this dialog box, give the bag the same name that was used for the zip file.

f. Accept defaults in the Save In dialog box. The bag will include: the dataset and metadata in a Data folder, plus other Bagger-generated files such as a checksum and manifest as shown in Figure 2.

g. Validate the bag transfer to Staging using Bagger's Validate Bag button. Check that the bag is complete using the Is Bag Complete button (Figure 1).

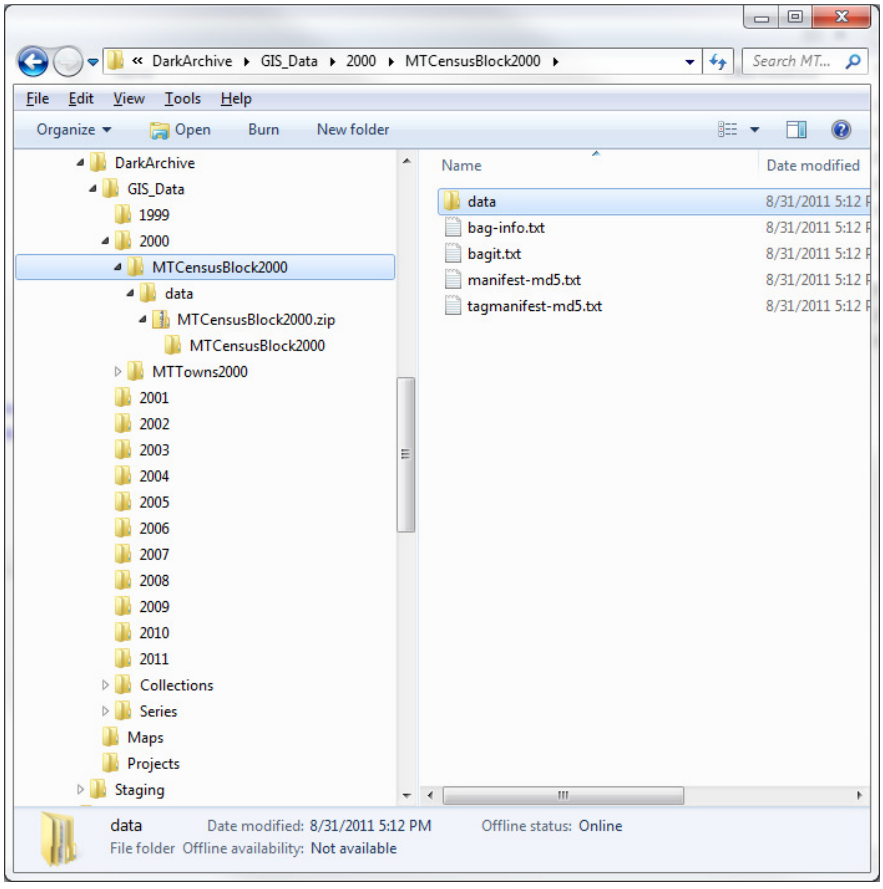


Figure 2: Bagger Generated Files with Data Transfer

h. Close the bag.

6. Enter archival and administrative metadata in the management system spreadsheet, including recording the zip file's checksum.

7. Delete dataset from the staging area.

Data Transfer Evaluation

Overall, the data transfer demonstration task met and even exceeded MSL expectations. The data transfer rates did not vary from what MSL has encountered previously as the state's GIS Clearinghouse; therefore it did not alter MSL's standard approach of transferring large datasets after work hours to avoid slowing down the network. The data transfer rates for the demonstration data were as follows:

- Landcover geodatabase zip file (100MB): 30 seconds
- Montana Census Blocks shapefile zip file (57MB): 7 seconds
- Montana Towns 2000 shapefile zip file (84KB): 1 second

MSL took a different approach to defining the Dark Archive data storage structure from that recommended in the GeoMAPP Data Transfer Best Practice. GeoMAPP suggests that data in an archive be categorized into one of nineteen ISO 19115 Topic Categories and stored in folders named for these categories. MSL considers ISO topic categories critical to how data is discovered by patrons. The patron data discovery system is the place where ISO categories will be leveraged. However, using ISO categories to guide staff with data storage placement presents difficulties because data often can be classified as meeting more than one ISO topic category. Selection of one ISO category for a dataset is then subject to staff member preferences, which may vary within the team. MSL chose instead to store the data by the time period of content unless the data is part of a series (i.e. cadastral, transportation) or if it was generated as part of a discrete project and is considered a collection. Parent-level collection and series metadata records will report the association between datasets.

MSL chose to zip the data before storing in the Dark Archive to save space in this MSL repository and in the Security Dark Archive stored offsite. Also, bagging one zip file instead of individual files produces one checksum, which may streamline dataset management and dataset integrity checking in the workflow. MSL chose not to use the Bagger zip functionality because the resulting data package made available to patrons had an excessively deep file structure, burying the data in multiple levels of folders.

The nature of how MSL approached the data transfer demonstration task was to learn from each transfer and strengthen the process. MSL revised the procedure between

each run, continuing to test the procedure each time. The process recorded in this document represents the draft data transfer and archiving procedure to date, though there are several areas that need further testing and consideration.

For many existing datasets in the collection, MSL chose to use its existing file names, which are based on an older, MSL file naming convention. The significant effort required to change these in a large collection (and associated discovery tools) is too great with limited resources. However, existing datasets that are en route to the Dark Archive will use the new naming convention. Newly-accessioned datasets will use the file and folder naming convention wherever naming is required.

The draft dataset management process (Appendix A) suggests that Submission Information Packages (SIPs) are defined as datasets that may have been modified by MSL staff before archiving and distributing to make the data more usable to library patrons. In those cases where substantial data modification is required for data to be made useful, there is some question whether the exact original data submission will also be archived. From a resource management standpoint, it makes little sense to archive data that is not useful to and end-user. This decision will be made on a case-by-case basis.

Next Steps

Further work needs to be completed to more formally define the MSL Spatial Data Collection Development Policy. Emphasis is placed on Montana GIS Clearinghouse data which has a statewide focus and MSDI data which incorporates both local and state data. Additionally, MSL staff will work with all MSDI theme stewards to develop archiving plans for each theme.

Montana is currently exploring different funding models to fund GIS data development activities including on-going development of the MSDI. An initial funding proposal includes an annual budget for a GIS archives program. This funding need will continue to be pursued as part of the larger funding discussion to be taken to the 2013 Montana Legislature.

MSL has draft system requirements for a GIS Dataset Management System (Appendix B) that will either be developed in house or by a solution vendor. This system will be used to continue the on-going process to inventory and archive the extensive MSL GIS data collection. In this GeoMAPP deliverable document, the system requirements represent the progress on developing requirements for a data management and archiving tool at

the time that this paper was written. Once MSL works through some of the issues identified in the data transfer evaluation, these system requirements will be updated.

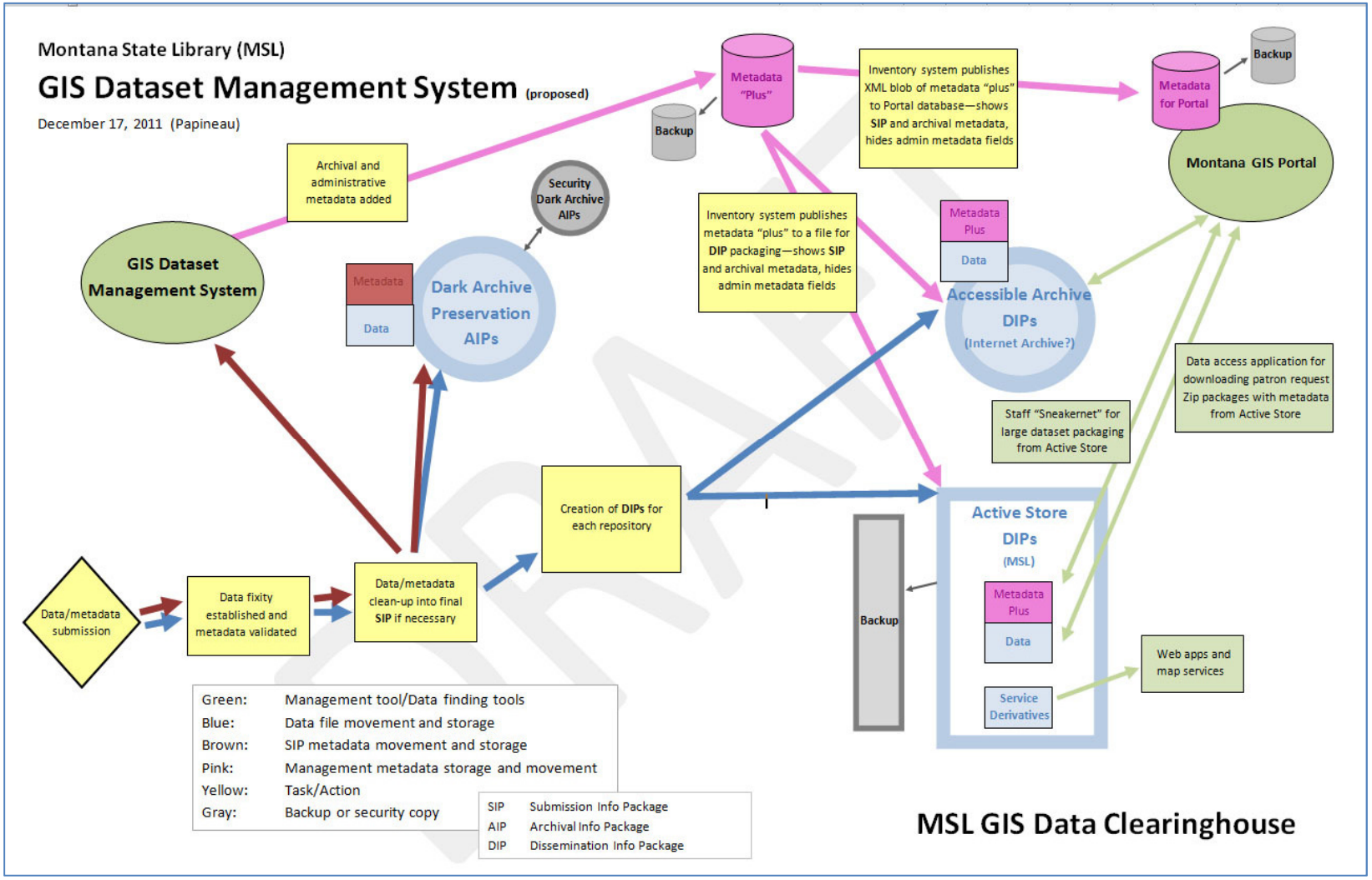
Draft requirements also need to be written for how the Security Dark Archive is integrity checked at its offsite location; as well as how that repository will be kept in sync with the Dark Archive at MSL.

More research will also be conducted on ways to improve discovery and delivery of archived GIS data. An initial review of the Internet Archive[3] site shows that a handful of shapefiles are available for download. MSL plans to test distribution of GIS data via this site in conjunction with the MSL State Publications program.

MSL will work with external partners that may have spatial data of archival value to develop standardized data transfer procedures which include virus scans and integrity checks on data being submitted to MSL. The goal for this step would be to ensure that any incoming data has a verifiable checksum generated prior to being transferred to the library so that we can verify the data that arrives here is consistent with what our partners submitted.

[3] Internet Archive: http://archive.org

Appendix A: *Draft* GIS Dataset Management System



Appendix B: *Draft* Requirements for GIS Dataset Management System

The system shall be able to:

1. Ingest a set of certain metadata fields into the system and validate against requirements.
2. Push SIP metadata from bagged file into a Blob field (SIP Blob).
3. Ingest data packages to the Dark Archive.
4. Generate series and collection IDs.
5. Generate unique IDs for each dataset that is recorded in its metadata.
6. Move data packages from one location to another.
7. Generate a checksum upon data ingest and record that checksum in a database field.
8. Validate a checksum against the original checksum at ingest.
9. Auto-populate the location of the data destination upon transfer.
10. Auto-populate automatically discernible administrative metadata fields, such as date uploaded.
11. Permit certain automatically-populated fields to be human edited after populating and then not overridden by the system.
12. Distinguish between original metadata uploaded and metadata fields edited and used for administration and archiving.
13. Be extensible for future growth (adding metadata fields and when possible asking the system to populate those fields by surveying the data and metadata holdings; modifications to any system components such as a database, the GUI, report formats, etc.).
14. Permit migration of the contents of the system to any new management system as technology changes.
15. Permit removing of test metadata records or mistakes within the system. Button “Remove” but it marks that record as removed (not deleting actual record)
16. Perform regular audits/validations of dataset integrity.
17. Send notification if an audit/validation check error occurs.

18. Restore a corrupt dataset inside either the Dark Archive or Security Dark Archive (use each other as the recovery source).

19. Send notification when a dataset/collection/series is due for reappraisal.

20. Record actions in an audit trail.

21. Configure access and permissions.

22. Recognize that a newly-ingested dataset is part of an existing collection or series leveraging a series or collection ID.

23. Identify parent/child relationships among data and among metadata records

24. Automatically push selected XML metadata out into a blob field in a separate database for use with the Montana GIS Portal.

25. Create Administrative metadata fields in the system that are populated manually, including a Notes or Comments field and an Other field, and the ability to attach documents.

26. Identify groups of data by format to facilitate batch format migration.

27. Harvest and upload datasets from locations outside of the Montana State Library's network, including performing proper packaging and fixity generation.

28. Generate reports based on predefined and ad hoc queries.

29. View/edit/move records and data in bulk via metadata queries (see next entry).

30. Respond to queries that answer the following questions about data in the MSL spatial data collection:

    - What Dark Archive datasets are due review?
    - What datasets are used in <this> web application?
    - What datasets are used in web map services?
    - What datasets cover <this> part of Montana?
    - What datasets are downloadable?
    - Which datasets are framework data?
    - What datasets are stored in the Accessible Archive?
    - What datasets are stored in the Active Store?
    - What datasets take up a lot of space?
        - What datasets were created by <this> agency?
        - What datasets are stored in <this> format?