**Digital Preservation Plan**

**Digital Preservation Working Group**

**University Archives and Preservation Unit**

**Library Services**

**Minnesota State University, Mankato**

**Created: March 4, 2016**

****

**Mankato State University student working at a computer. MSU-UA-309-14522. University Archives, Memorial Library, Minnesota State University, Mankato.**

<http://arch.lib.mnsu.edu/islandora/object/MSUrepository%3A6371>

**Statement of Purpose**

The Digital Preservation Working Group (a.k.a. DigiVengers) was formed in 2014 within the University Archives and Preservation Unit of Library Services at Minnesota State University, Mankato with the purpose of creating and implementing a digital preservation plan for our expanding locally held digital materials. The DigiVengers group consisted of Jamie Dalbey, Preservation and Government Documents Technician; Anne Stenzel, Archives Technician; and Heidi Southworth, Digital Initiatives Librarian. The following document is a working draft of our recommendations for effective storage, usability, maintenance and updates to our digital collections for the continued use and accessibility of students, faculty, staff, and the global online community.

**Table of Contents**

Brief Description of Digital Collections and Ingest Process

Metadata Standards Utilized

Locations of Digital Collections (Storage Locations)

Arrangement of Digital Collections in Storage Locations

Number of Copies of Items Retained

Backups Copies of Digital Collections

Checksums and other Preservation Tools

Long-Term Storage Strategies

**Brief Description of Digital Collections**

Digital Collections at Minnesota State University, Mankato consist of two repositories known as ARCH and Cornerstone. Together these repositories present the scholarly and historical record of the University and region. The following descriptions have been developed to help identify types of collections that might be located in either repository.

**ARCH: University Archives Digital Collections** – Preserves and provides access to historical information about Minnesota State University, Mankato and South Central Minnesota

**Cornerstone: A Collection of Scholarly and Creative Works** – Collects and provides access to the scholarly and creative output of Minnesota State University, Mankato.

The two repositories have different functionalities and the University Archives reserves the right to add collections into the most appropriate repository.

Examples of content by repository.

|  |  |
| --- | --- |
| ARCH | Cornerstone |
| Alumni Magazine | Conference posters/presentations |
| Bulletins | Creative works |
| Historical atlases\* | Datasets |
| Manuscript collections | Department journals |
| Newspapers | Department newsletters\* |
| Oral histories | Digital humanities projects\* |
| Photographs | ETDs and capstone projects |
| Yearbooks | Faculty publications (Articles, etc.) |
|  | Grants funded |
|  | Senior seminar projects |
|  | Student journals/creative magazines |
|  | URC/GRC |

\*example of a type of material that could go in either repository

**ARCH =** [**http://arch.lib.mnsu.edu/**](http://arch.lib.mnsu.edu/)

ARCH: University Archives Digital Collections collects and provides access to the scholarly and creative output of Minnesota State University, Mankato. Islandora is an open source web-based digital asset management system that stores, manages, showcases and preserves digital assets. Islandora leverages three core open source software components, all of which are well-supported and widely used. Fedora is the heart of the system and provides the digital asset management functions, Drupal provides for an extremely customizable user interface, and the fast searching and retrieval within Islandora is due to the use of Solr, which is also very scalable.

**Cornerstone =** [**http://cornerstone.lib.mnsu.edu**](http://cornerstone.lib.mnsu.edu)

Cornerstone: A Collection of Scholarly and Creative Works showcases the intellectual output of Minnesota State University, Mankato’s faculty, staff, and students by preserving their works digitally and presenting them to the world in an easy-to-find format. Cornerstone is an institutional repository for the University and runs on the Digital Commons platform from bepress (<http://digitalcommons.bepress.com>). Digital Commons is a suite of tools and services that enables institutions to manage, display, and publish scholarship to the web in a beautiful, highly visible showcase. Digital Commons offers the features of a traditional institutional repository software platform as well as professional-grade publishing software, management tools, and individual faculty and researcher pages to promote and disseminate scholarship and serve academia. Digital Commons is a subscription based service and has an annual fee paid by Library Services. The platform is proprietary and is a fully hosted and supported service with no servers to maintain or software to upgrade – it is all done by bepress. Our subscription includes unlimited space and a dedicated customer services representative. Cornerstone contains multiple collections (referred to as series). A complete list of series can be found here: <http://cornerstone.lib.mnsu.edu/communities.html>

**Brief Description of Ingest Process for Digital Collections**

ARCH (do not retain copies of zip files, thumbnails or JPEG2000) – Content to be added later

Cornerstone is managed by the Digital Initiatives Librarian. Items are added to Cornerstone by the Digital Initiatives Librarian and the Archives Technician using templates created by the bepress customer service representative. Information on how to ingest items can be found here: <http://digitalcommons.bepress.com/reference>. Textual documents are converted to PDF format by the Digital Initiatives Librarian or Archives Technician if needed. Only PDF format is used for textual documents. Other acceptable formats include: TIFF for images, WAVE for audio and MPEG-2 for video.

**Metadata Standards Utilized**

All items added to our digital collections include full-level metadata records.

For ARCH, all items are described utilizing the Minnesota Reflections Metadata Entry Guidelines from the Minnesota Digital Library (<http://mndigital.org/standards-best-practices/minnesota-reflections-metadata-entry-guidelines>). These guidelines contain descriptive, administrative, and technical metadata and are formatted by the Islandora platform using MODS metadata schema. Metadata records in ARCH can be exported out in CSV format using Dublin Core metadata schema.

Cornerstone at this time only contains descriptive level metadata based in Dublin Core metadata schema. Metadata in Cornerstone is customizable, so we work with our customer service representative to add fields. Any type of field may be added to any series within Cornerstone. Information on Digital Commons’ metadata schema can be found at: <http://digitalcommons.bepress.com/reference/87>.

For certain series in Cornerstone, administrative and technical metadata based on the Minnesota Reflections Metadata Entry Guidelines from the Minnesota Digital Library is generated and retained on separate Excel spreadsheets. At this time, this type of metadata is generated only for the 1959 and 1967 Greater Mankato Air Photos found in the Minnesota County Air Photos series. The spreadsheets are stored in the local Cornerstone Preservation (ml01) drive at:   
Cornerstone Preservation\Aerial Photograph Project\BDJ\_1967\Metadata\Metadata\_Master\_BDJ\_1967.xlsx   
and   
Cornerstone Preservation\Aerial Photograph Project\HXG\_1959\Metadata\Metadata\_Master\_HXG\_1959.xlsx

**Locations of Digital Collections**

In addition to digital items located in ARCH (Islandora) and Cornerstone (bepress Digital Commmons), we use the following locations for storing and working with digital copies.

1. Library Server (ml01) – Windows 2008 based server, Library owned/managed, 6 years old   
   1. **Archives Department Drive** = contains working files for digital projects for Archives staff and students; temporary/working drive; currently has 278 GB capacity; contains 259 GB content; free space is 19.5 GB.
   2. **ArchivesWorking Drive** = contains working files for digital projects for Archives staff and students; temporary/working drive; currently has 278 GB capacity; contains 237 GB content; free space is 41.8 GB.
   3. **REPORTERWorking Drive** = working drive for Archives students to store and work with copies for the Reporter (student newspaper) digitization project; temporary/working drive; currently has 244 GB capacity; contains 91.8 GB content; free space is 152 GB.
   4. **CornerstonePreservation Drive** = contains files related to Cornerstone project; preservation/permanent storage for Cornerstone; currently contains duplicate ARCH files that will be moved; currently has 527 GB; contains 307 GB content; free space is 219 GB.
2. **GOLD** – GOLD is a networked, attached storage unit, 24TB, from iXsystem called the FreeNAS mini. It is made up of 4 WD Red NAS Hard Drives that are 6.0 TB each. Operating system is called FreeNAS (open source, <http://www.freenas.org>). Contains 10.2 TB of usable storage space. Access or use of GOLD is restricted to Archives and Systems staff. Used solely for preservation copies for ARCH. Systems has 2 backup hard drives for the unit (6.0 TB SATA/64MB Cache WD60EFRX, NASware 3.0 WD Red NAS Hard Drives).
3. **Hard Drives** – currently have 4-5 hard drives of various sizes, formatted for PCs. Largest is 3 TB. These were used to store copies of digital items found in ARCH. Temporary storage units; now being decommissioned and replaced with GOLD.
   1. HD1: MSU Reporters, 40199-41466, SMHC items, 1.18 TB
   2. HD3: Student/Mankatonian, msu40000-40198 676 GB
   3. HD4: MSU Reporters, msu41465-42737, 1265 items 277 GB
4. **CDs** – Used to store copies of digital items found in ARCH - temporary storage. We have two copies of all CDs. One set of copies is kept off site at University Records Center, TC065 and the other copy is kept on site in the ARCHives upper storage cabinets, ML3097. No longer saving copies to CDs. Life expectancy of CDs = vary depending on storage conditions – 20-100 years, depending on errors. See CLIR report: <http://www.clir.org/pubs/reports/pub121/sec4.html>
5. **Other Storage Places**   
   1. **MavDisk** – individual networked drive, backed up by campus ITS; all staff currently have 32 GB space; individual drives cannot be seen by others unless shared (all or nothing share); good for working documents, not preservation.
   2. **OneDrive** – individual Office365 OneDrive accounts; limit of 20,000 items (files and folders); can share documents; good for working documents, not preservation.

**Arrangement/Organization of Digital Collections in Storage Locations**

1. **File/Folder Organization**
   1. **ARCH on GOLD** = organized into 6 base categories and then specific collections within:
      1. Student Publications
         1. Student Magazines
            1. The Student 1889-1891
            2. The Mankatonian 1891-1913
         2. Student Newspapers
            1. Among Ourselves 1926-1927
            2. School Spirit 1929-1933
            3. College Spirit 1933-1935
            4. College Reporter 1935-1968
            5. The Reporter 1968 - ongoing
      2. University Publications
      3. Athletics
      4. SMHC
      5. The Campus Experience
         1. University Photograph Collection. Photographs, 1868-Ongoing. MSU Archives Collection 309.
         2. Minnesota State University, Mankato Homecoming. Oral History Interviews, 2015-. SMHC Manuscript Collection 1808.
         3. Vikings Training Camp Stories, 2015
      6. University Offices/Departments
   2. **Cornerstone on CornerstonePreservation Drive** = organized by series (collection) name and then by year as appropriate. See <http://cornerstone.lib.mnsu.edu/communities.html> for a complete list of series.
2. **Consistent File Naming Structure**
   1. **ARCH** = File names for digital items in ARCH have the same pattern:

MSU Identifier Number\_SMHC or UA\_Collection Number\_Box Number\_Folder Number\_Item Digitized Number.File extension abbreviation

msu00072\_SMHC\_103\_07\_022\_0006.tif

* + 1. **MSU Identifier Number** – for the MSU Identifier Number, use the following chart.

|  |  |
| --- | --- |
| MSU Identifier | Collection |
| msu00001-msu00072 | Yearbooks, University Student Publications. Collection, 1888-Ongoing. MSU Archives Collection 130. |
| msu00072-msu10000 | Manuscript Collections (SMHC and UA) |
| msu10001-msu24999 | University Photograph Collection, 1868-ongoing. MSU Archives 309. |
| msu25000-msu39499 | NEGATIVES University Photograph Collection, 1868-ongoing. MSU Archives 309. |
| msu39500-msu39999 |  |
| msu40000-msu50000 | University Student Publications. Collection, 1888-Ongoing. MSU Archives Collection 130. |
| msu50001-msu60000 |  |
| msu60001-msu99999 |  |

* + 1. **SMHC or UA** = use the appropriate abbreviation for the archival collection. Use either SMHC for the Southern Minnesota Historical Collection or UA for the University Archives.
    2. **Collection Number** = use the number of the collection that the digitized item belongs to.
    3. **Box Number** = use the number of the box that the digitized item is stored in. Use two digits for 1-9.
    4. **Folder Number** = use the number of the folder that the digitized item is stored in. Use three digits.
    5. **Item Digitized Number** = use a four digit, unique number
    6. **File extension abbreviation** = include the three character file extension preceded with a period (Example: .pdf or .tif).

* 1. **Cornerstone** = File names for digital items in Cornerstone have the same pattern:   
       
     c\_Series Name\_Manuscript Number\_Original File Name. File extension abbreviation  
       
     c\_biol\_fac\_pubs\_1000\_Land\_Allison\_Oncotarget\_2015\_10\_31\_Article.pdf
     1. **c** = precede the file name with a lower case “c” for Cornerstone
     2. **Series Name** = use the abbreviated series name found in Cornerstone
     3. **Manuscript Number** = use the manuscript number assigned to the item upon ingest.
     4. **Original File Name**  = use the file name as submitted
     5. **File extension abbreviation** = include the three character file extension preceded with a period (Example: .pdf or .tif).
  2. **Logs** – Logs are Excel spreadsheets that contain a comprehensive list of all digital assets.
     1. **ARCH log** = Currently saved in a folder at: N:\Library\_Services\Archives\Digitization\Logs
     2. **Cornerstone log** = Currently saved in a folder at: N:\Library\_Services\Archives\Digitization\Logs

**Number of Copies of Digital Items Retained**

While the LOCKSS principles call for seven different copies in seven different locations, we recommend the following:

1. If an item is **Born Digital**, three digital copies are retained. The Master copy is retained on a networked drive (TBD) for ARCH or on the CornerstonePreservation drive for Cornerstone. Two additional copies are kept in ARCH or Cornerstone and one on an external Hard Drive or on GOLD.
2. If an item is a **physical item that has been digitized** and the Archives retains the original copy, two digital copies are retained. One copy is kept in ARCH or Cornerstone. The Master copy is retained on a networked drive (TBD) for ARCH or on the CornerstonePreservation drive for Cornerstone.
3. These number of copies do not include backup copies (see next section).

**Backup Copies of Digital Collections**

1. ARCH backups by PALS - PALS currently uses disk-to-disk backup systems for Islandora, using an ORACLE SunFire Storage Server. Current backup procedures are as follows:
   1. Weekly (Saturday mornings) PALS takes full data dumps of all databases and data files, production and development.
   2. The data dumps are copied to two separate Virtual Tape Libraries.
   3. PALS has archive logging turned on for all production databases
   4. The files are copied to multiple physical locations as they are generated.
2. Cornerstone backups by bepress – bepress maintains the following backup system:
   1. All production servers are maintained at a high availability colocation facility with multiple backbone connections and back-up generators. The facility is secure and requires physical tokens (badges) for access.
   2. Failover web, database, and storage servers are maintained to continue to serve content in case of failures.
   3. All databases have real-time redundancy that runs continuously, and nightly back-ups of our entire database are done. The nightly back-ups are stored away from bepress’ colocation facility in a separate physical location.
   4. All uploaded files are stored in triplicate in a redundant storage cluster, as well as being backed up offsite to a third-party cloud service that specializes in data archiving and back-up.
   5. bepress sends monthly back-ups of all other data to a third-party archival service, an industry leader in data protection and recovery services. The archival service maintains back-up tapes for one year.

**Checksum Tools and other Preservation Tools**

1. Checksums
   1. Checksums on all digital objects are currently ran every three months.
   2. Checksum format used is SHA-1. SHA-256 or higher will take longer to run and are recommended for use by NIST for generating digital signatures, time stamps and for other applications that require collision resistance (see <http://csrc.nist.gov/groups/ST/hash/policy.html>). At this point, there is no consensus for which checksum format to use for ensuring data integrity, but the literature states that SHA-1 or MD5 are okay to use for just generating checksums. PALS recommended SHA-1 to us to use.
   3. Tools available to run checksums are:
      1. **Exact File** = <http://www.exactfile.com>
      2. **DROID** = <http://www.nationalarchives.gov.uk/information-management/manage-information/preserving-digital-records/droid/>
      3. **HashMyFiles** = <http://www.nirsoft.net/utils/hash_my_files.html>
      4. **Fixity** = <https://www.avpreserve.com/tools/fixity>
   4. Recommendation is to use Fixity in conjunction with Bagger/BagIt, but others can be run.
   5. Checksum files are to be stored in Bagger/BagIt with the digital copies and metadata.
2. Other Preservation Tools
   1. Metadata Extraction Tools
      1. **ExifTool** - <http://www.sno.phy.queensu.ca/~phil/exiftool/> - use to extract metadata from multiple files on a hard drive or other device via a command line interface. Potential uses: for working with hard drives from Theatre and Athletics.
      2. **NARA File Analyzer and Metadata Harvester** – <https://github.com/usnationalarchives/File-Analyzer> - use to examine a large series of files (such as on a hard drive) and scan to see what is on it. Potential uses: use to scan large hard drives from Theatre and Athletics to get a sense of what types of files we are to be working with. This program will also generate SHA-1, MD5 and SHA-256 checksums.
      3. **MediaInfo** - <http://mediaarea.net/en/MediaInfo> - use for metadata extraction for audio and video files. Potential uses: for working with our oral history files and other AV materials.
   2. Metadata Checking – use **MDQC** (<https://www.avpreserve.com/tools/mdqc>) to help check bulk metadata. Potential uses: use to double check metadata entry from projects sent out to vendor for large scale scanning.
   3. File Structure/Containment
      1. **ReNamer Lite** = <http://www.den4b.com/?x=downloads&product=renamer> - use to rename files. Potential uses: use to clean up and rename files according to our Consistent File Naming Structure.
      2. **Bulk Rename Utility** - <http://www.bulkrenameutility.co.uk/Main_Intro.php> – use to rename files. Potential uses: use to clean up and rename files according to our Consistent File Naming Structure
      3. **Bagger and BagIt** - <https://sourceforge.net/projects/loc-xferutils/files/loc-bagger/2.1.3> - use to group/bag together files and associated metadata and checksums. Use in conjunction with Zip files.

**Long-Term Storage and Backup Strategies**

Need to explore this topic further. Possibilities for cloud storage for our backups and copies include:

1. Azure - <https://azure.microsoft.com> – cloud based storage from Microsoft, could potentially partner with ITS for service.
2. Potential other partners for Cloud based storage
   1. PALS
   2. Digital Commons Private LOCKSS Network
   3. Minnesota Private LOCKSS Network