One Stone, Two Birds: Data Assessment Framework for Digital Collection Audit and Preservation

Aaron Collie; Michigan State University; E. Lansing, MI
Lucas Mak; Michigan State University; E. Lansing, MI
Shawn Nicholson; Michigan State University; E. Lansing, MI

Abstract

The digital curation community is currently investigating scalable solutions for digital preservation and access. "As the size and complexity of digital collections increases how will curators classify, prioritize, capture, preserve, and present their historical holdings?" To meet these requirements the Digital Information Division of the Michigan State University Libraries has called for a comprehensive inventory of its preservation quality digital collections. A comprehensive inventory will produce an index of collections and items (inventory), provide descriptive information (classification), summarize the state of preservation (assessment), and inform workflow (policy). The Data Asset Framework, formerly the Data Audit Framework, a standard tool for providing a representative audit, is modified to accommodate the scale and density of digital collections acquired by libraries. Modifications to the DAF include 1) automated tools for metadata extraction, checksum generation, and format authentication 2) file storage summaries using disk analysis reporting and container format specifications, and 3) a modified interview schedule to gather metadata at the collection level.

One possible output of modifying the DAF to assess digital collections is the potential to map DAF data into PREMIS records for preservation purposes. This activity is unique on the collection level and enhanced reuse is the main benefit of such mapping. The stored XML data can be compared with new audit data for any updates and affords a novel means for tracking changes to digital objects at the collection level. This paper will detail the process used to 1) modify the DAF, 2) map DAF data elements into PREMIS and 3) build a business case for reuse for future audits. The composition and rationale for the tool set and the results attained form the concluding remarks.

Introduction

Background

Physical acquisitions come in physical packages: often through the mail, occasionally left on our doorstep, maybe through the loading dock, and even hand-to-hand. Common practices for the acquisition and management of physical assets have developed across a wide range of organizations from corporate, government, not-for-profit, and educational settings. Inventory management procedures have therefore adapted to accommodate dramatically disparate assets and the parameters of those assets.

Digital acquisitions attempt to break the mold. Digital assets sometimes come neatly packaged on physical media, but they can also easily be digitally synthesized, digitally duplicated, and digitally transferred. Just as with physical assets, adaptations have been made to many management models to account for these new parameters in inventory management.

Digital asset management is a flavor of inventory management which includes a management strategy for the ingesting, annotation, cataloging, storage, retrieval and dissemination of digital assets. Just like physical assets, organizations manage digital assets for any number of reasons. For example a business might manage internal documents and records, but also sell digital video, electronic books or three dimensional blueprints.

Cultural heritage institutions such as libraries, archives, museums and art galleries face a unique obstacle in that these institutions often are charged with identifying and acquiring assets which hold informational, artistic, or cultural value that is not dictated by market dynamics like supply and demand. This has long been the mission of such institutions and these organizations have therefore developed guiding policies and leveraged the knowledge and skill of professional librarians, archivists and curators to address this obstacle.

However, for the foreseeable future, these policies, workflows and tasks will be challenged under the "deluge" of digital content which has been shaken free of a physical form due to the aforementioned fundamental shift in asset management—unrestricted synthesis, immediate and exact duplication, and boundary-bending digital transfer. While many policies can and have been adapted in light of increasing acquisition of digital assets some will take time to become distinguished in the digital "cloud". An example of this challenge would be meeting specifications for content or metadata transformations in preparations for ingest into collaborative preservation environment such as those established by the HathiTrust http://www.hathitrust.org/ingest.

This flavor of digital asset management has been studied extensively by technologists and professionals alike, and these discussions manifest in our scholarly literature under the umbrella term digital curation. Many models have been employed and infrastructures built under the banner of digital curation, and these now face the same iterative optimization of the information management practices which existed before them.

It is clear that since we have begun managing digital assets the scale and complexity of digital collections has increased, and that like their physical manifestations there is not likely a one-sized-fits-all solution. Instead, organizations will be required to internally acclimate to managing assets with new characteristics such as non-rivalry, transmutability, and complex meronomy.

Problem Statement

Digital acquisitions do not line up in single-file for ingest into a preservation archive. Not only do backlogs develop, but collections move through ingest workflows at differing rates—sometimes pooling about as they await decision making (policy) or problem solving (infrastructure). Even worse, workflow standards change and digital collections become estranged or even re-ingested. This is particularly evident as digital objects move through refresh cycles, including, not exclusive to, rescans to meet evolving capture rates.

Although somewhat relaxed after the second-wind rallying around micro-service models, digital librarians and archivists have expressed that ingesting digital content is viewed as "expensive." In many ways "ingest" has come to be understood as a formalized protocol. This is possibly due to the rigidity of standards-compliant software tools and the workflows which develop around those tools. This formalization as a "phase" or "stage" suggests ingest is viewed as a one-time, front-loaded cost.

Added to these hurdles, and as hinted at before, digital acquisitions arrive in variegated "packages" of inconsistent orders of structure. Some "packages" might include: download/upload (FTP), CD-ROM, DVD-ROM, CD-RW, DVD-RW, SSD (Flash drives or solid state hard drives), Zip Disks, 3.5" Disks, Spinning Disk Drives (Internal or External), and MiniDisc, to name a few. Other methods of acquisition may include born-digital or created content (e.g. digitization) or duplication/reformatting (e.g. backup) of content. It should not be that far off until acquisitions include entire disk arrays, virtual servers and storage, and entire content management systems (e.g. a merger).

It is suggested then that acquisition and ingest of digital assets is not quite black and white (e.g. "pre" and "post"), but rather a subset of the selection and appraisal tasks associated with traditional collections management. This suggestion is not novel, and is in fact supported by both the lifecycle model for digital curation and by the traditions of library science. As digital content accumulates, it will be helpful to develop tools, workflows and policies for assessing, prioritizing, and characterizing digital assets.

Auditing: Assessment, Maintenance, and Retroaction

Audits provide organizations with the means to identify describe and assess how they are managing their assets. The general condition of a collection, for example, cannot simply be deduced by walking the shelf or peering into a box of donations. Instead, in the digital environment, checksums must be validated against file manifests to conclude fixity of collections, individual files must pass quality control measures to conclude the fidelity of assets, directories must be mapped and recursed to measure entropy and density, and formats must be documented to inform migration policies. Already tools exist to perform some of these functions (e.g. JHOVE2, FITS, etc), and these tools will soon make their way into every digital curator's toolset. However, metadata extraction and format authentication are only one tool from a digital curators tool belt, and still to develop are more tools, better workflows, and more complete policies.

Another tool for assessing digital assets for preservation comes from the sister discipline of data curation. Data curation is a subsystem of digital curation which is specifically concerned with research data. Research data, primary source data, and cultural heritage data share many of the same parameters. The Data Asset Framework (DAF), formerly the Data Audit Framework allows organizations to "identify, locate, describe and assess how they are managing their research data assets" (citation). The DAF is a framework for sifting through unstructured data and semi-structured organizations to discover and assess valuable resources. It will be through the marriage of frameworks like DAF and tools like JHOVE2 (again reiterating: more tools, better workflows, and more complete policies) that a complete workbench for the audit and preservation of digital assets will emerge.

It is under this assumption that Michigan State University Libraries began investigating ways to adopt lightweight solutions for inventory management that can be iteratively developed and applied during ingest, maintenance, or any cycle of digital asset management. In fact one primary use case for conducting an audit and inventory of digital collections was to identify retroactive work—maintenance—for the digital collections.

The objectives of the audit are to 1) provide a flexible inventory of assets and aggregates of assets, 2) classify assets for institutional specific preservation purposes, 3) provide an assessment of assets to guide future preservation strategies for the institution, and 4) inform the development of policy and workflows.

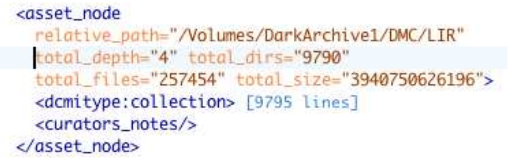
The audit has been designed under the guiding framework of the Data Asset Framework with a series of modifications meant to address the scale and density of library collections. Primary modifications include automated metadata extraction, simplified forms and templates, and assumed an increased access to internal documentation and shared file space. A discussion of these modifications and possibilities for future work follows.

Methodology

Michigan State University (MSU) is a public research university in East Lansing, Michigan. It was founded in 1855 as the pioneer land-grant institution and has served as a model for land-grant colleges in the United States. MSU Libraries hold over 5,000,000 print and electronic monographic and serial resources, 200,000 maps, and 40,000 hours of spoken word recordings. In addition to these public collections, the Libraries hold over 2,000,000 preservation quality digital assets which make up roughly 26 Terabytes (TB) of information.

The three shares which constitute what is termed the "dark archive" or "preservation archive" ("DarkArchive1", "DarkArchive2", "DarkArchive3") are 14TB, 13TB, and 15TB respectively for a total capacity of 42TB. Digital files are stored in a shared disk file system on enterprise spinning disk drives in a RAID 5 disk array and are made accessible using a dedicated storage network (SAN). Everyday interaction with the "dark archive" is through the use of the application-layer network protocol Server Message Block (SMB), also known as Common Internet File System (CIFS). This setup is considered a common enterprise storage architecture that is compatible with the variety of use cases and operating environments of large organizations.

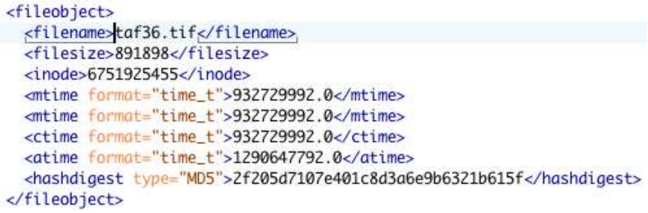
The DAF Form 2 "Inventory of data assets" was expressed as a custom XML schema which captures a collection as an "asset node" with attributes which describe the directory path, total number of subdirectories, total number of files, total number of bytes, and deepest directory depth from the relative path. This data is used to represent the entropy and density of a collection, and will inform low hanging fruit and high risk collections.



```
<asset_node
  relative_path="/Volumes/DarkArchive1/DMC/LIR"
  total_depth="4" total_dirs="9790"
  total_files="257454" total_size="3940750626196">
  <dcmitype:collection> [9795 lines]
  <curators_notes/>
</asset_node>
```

***Figure 1**. "Inventory of data assets"*

The DAF Form 3A "Data asset management (core element set)" is also expressed as XML using the Forensic XML schema which is a simplified and flexible schema that can capture the metadata stored natively in a filesystem using the tuple returned by the unix system call stat() or the python os.stat function. This example was generated using the dfxml_tool.py from AFFLIB.org.



```
<fileobject>
  <filename>taf36.tif</filename>
  <filesize>891898</filesize>
  <inode>6751925455</inode>
  <mtime format="time_t">932729992.0</mtime>
  <mtime format="time_t">932729992.0</mtime>
  <ctime format="time_t">932729992.0</ctime>
  <atime format="time_t">1290647792.0</atime>
  <hashdigest type="MD5">2f205d7107e401c8d3a6e9b6321b615f</hashdigest>
</fileobject>
```

***Figure 2**. "Data asset management (core element set)"*

Along with the methods described above, a variety of tools were used or investigated for use while conducting the data audit. The following tables present a summary of the tools used during this project. This list is not comprehensive, and many of the primary functions can be completed with simple command line system calls, however these tools generally provide a level of customization or ease-of-use that befits the intent of the data audit. More tools can be found on the NDIIPP Partner Tools and Services Inventory [1]. Also of note are the Archivematica (0.7-alpha) [2], Archivist's Toolkit (2.0) [3] and Curator's Workbench [4] toolsets, which pre-package many of these tools or tools with similar functions.

Disk Analytic Reporting

JDiskReport http://www.jgoodies.com/freeware/jdiskreport/
Karen's Directory Printer http://www.karenware.com/powertools/ptdirprn.asp
Xinorbis http://www.freshney.org/xinorbis/index.htm
WinDirStat http://windirstat.info/

Technical Metadata Extraction

JHOVE2 http://www.jhove2.org
FITS http://code.google.com/p/fits/
fiwalk / dfxml_tool http://afflib.org/software/fiwalk

File Format Identification / Authentication

DROID http://droid.sourceforge.net/
TRiD http://mark0.net/soft-trid-e.html
libmagic http://sourceforge.net/projects/libmagic

Bulk File Renaming

Bulk Rename Utility http://www.bulkrenameutility.co.uk/Main_Intro.php
Rename Master http://www.joejoesoft.com/cms/showpage.php?cid=108
Filewrangler http://development.christopherdrum.com/software/

Results

The audit revealed that the most prevalent assets were image files (TIFF – 86.6% by KB, 88.9% by files) and audio files (WAV – 7.3% by KB, 1.2% by files) which together make up more than 90% of the collection. The entire collection contains more than 100 file types indicating preservation masters (e.g. TIFF), compressed files, production files (e.g. a project file), and unidentified files.

File Formats

	File Sizes (KB)	% of Total	Files	% of Files
tif	23,468,057,362	86.6%	1,263,077	88.9%
wav	1,974,985,084	7.3%	16,424	1.2%
avi	332,344,303	1.2%	1,117	0.1%
bz2	239,090,764	0.9%	75,084	5.3%
vob	140,557,804	0.5%	280	0.0%
pdf	124,908,752	0.5%	16,425	1.2%
m2v	56,943,309	0.2%	60	0.0%
zip	35,964,100	0.1%	225	0.0%

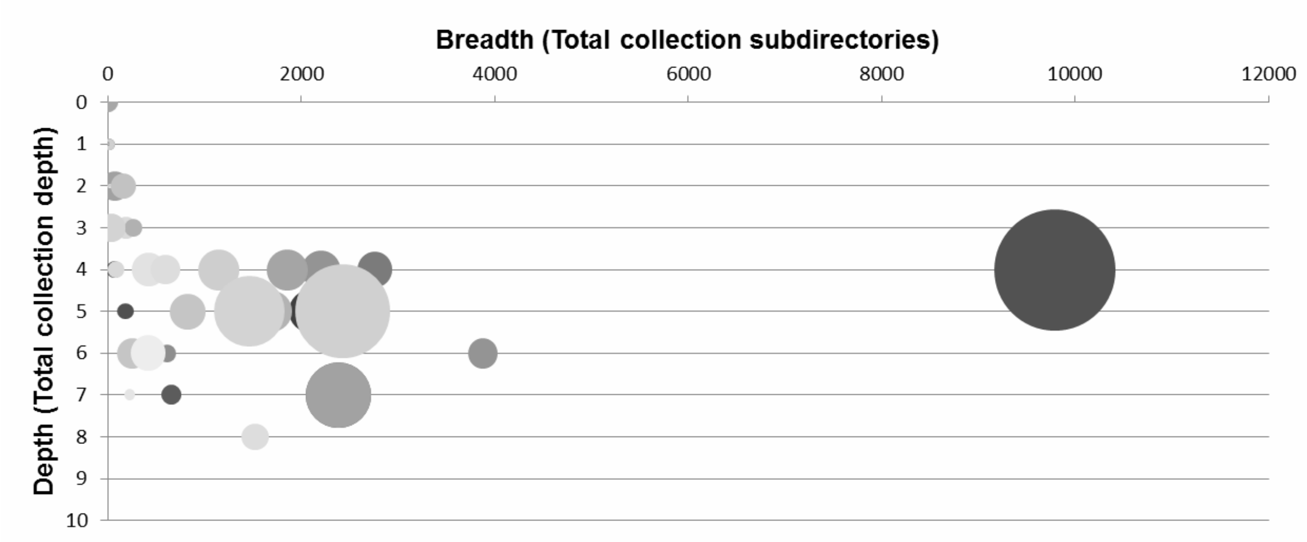


Figure 3 *Density of collections (size of dot represented by number of files)*

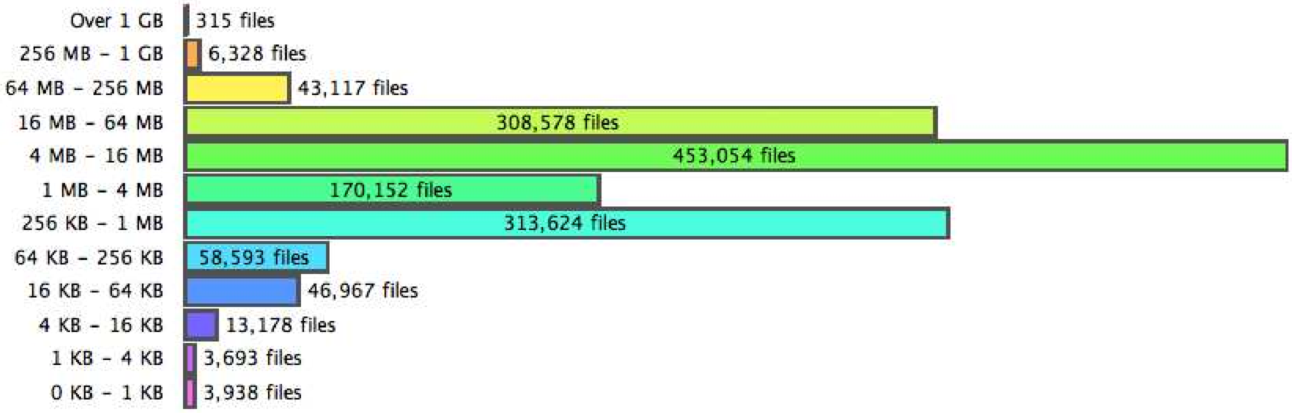


Figure 4 *File densities (output of JDiskReport)*

The largest collection was over 3.6 TB (257,454 total files) spread out over 9790 directories and had a max depth of 4 directories. The smallest collections were empty directories, and indicate the need for maintenance action plans. Based on a total of 76 aggregations, an average collection might be 327 GB (13,178 files), and would contain 520 directories with a maximum directory depth of 3. However, these numbers should not be extrapolated to compare to other dark archives, but attempt to characterize one example archive.

Distribution of files size clusters in the 256k-64 MB (87.6%) range, with the majority in the 4-64MB range (53.5%). Files with a 0 – 1 KB file size could represent blank files (e.g. blank OCR) and files over 1 GB could indicate a preservation risk; both sets will be queued for item-level investigation. Outlier collections appear to have few files at a great depth or few files at a great depth. One collection was found to contain no files, despite numerous directories. Another collection was found to have a max depth of 7 but a disproportionally small number of files. These collections could represent high entropy collections, and may represent items which will be of interest for further assessment. Results which will require further investigation include a large collection of compressed files (74,000+ files), project or temporary files, and files in non-standard format.

Discussion

Because archivists and librarians have greater access (than an external auditor) to internal documentation and permissions for shared disk space they are better positioned to utilize the snowballing method of data collection as alluded to in the many published DAF case studies. In fact, a greater familiarity with the organization under audit and a larger resource base were found to reduce the time needed to complete many of the DAF activities. This time is well accounted for, however, as library collections have the potential to have already been "selected" or deemed somehow valuable for preservation. This means that library collections may be considerably denser than disparate collections of research data which only share the umbrella organization's namesake. Library collections also tend to cluster around resource centers (hard money) rather than individuals (researchers, grants, etc) and may exhibit a level of uniformity inherited from the resource centers mission and policy. Further study and comparative analysis of these data types will shed more light on these questions.

It was often discussed during project planning that it is difficult to determine what would constitute a barebones inventory and classification measure for preservation purposes. Early in the project initiation phase it was noted that batch metadata creation and extraction tools such as FITS and JHOVE2 would produce redundant information (and duplicate precious CPU cycles when working at the TB scale) if more than the bare minimum of information was collected in the audit phase. As using these tools is likely a follow on activity in what will be an iterative auditing process JHOVE and FITS can potentially provide detailed technical metadata for digital preservation purposes. One execution of an iterative audit could store technical metadata in PREMIS (Preservation Metadata: Implementation Strategies) format and be included in a METS (Metadata Encoding and Transmission Standard) package. For example, the MIX (Metadata for Images in XML Standard) metadata output for digital image files (e.g. TIFF) from JHOVE could be extracted and put into <objectCharacteristicsExtension> under PREMIS and be wrapped under the <techMD> element within the METS <amdSec>. Since METS and PREMIS have been developed in different time with different focuses, there are some redundancies between the two schemas [2]. Decision points would include whether to repeat certain metadata in different sections in PREMIS and METS and how to include PREMIS metadata into the METS package. It would be a local implementation decision whether to put the whole PREMIS XML file into the <amdSec> under METS or to break the PREMIS file apart and put individual elements into corresponding sub-elements (e.g. techMD, digiProvMD etc.) within the <amdSec>.

It should come as no surprise that DAF Form 3A could be automated and mapped directly to Dublin Core XML or RDF, and form 3B could be automated and mapped to PREMIS or METS (or other applicable standard) by using the output from JHOVE2 or FITS. At the time of project initiation, these tools were in alpha and beta status.

Conclusion

The MSU audit emphasized reusing and enhancing existing tools and protocols. By detailing the processes employed to modify the DAF, it is hoped that others can experiment with auditing digital collections at the Library scale. Going forward MSU will look for ways to build from this initial audit. Some of these follow up activities may include capacity planning, informing requirements for new systems, conducting gap analysis for multimedia services or migration policies, indexing the metadata (e.g. solr), de-duplication reporting (e.g. checksum comparison), additional metadata creation (e.g. FITS, JHOVE2), developing a preservation action plan, or developing a monitoring action (e.g. BagIt, text manifests). The most critical activities going forward are to: 1) standardize file hierarchy and naming conventions in order to facilitate programmatic access to the archive (a retroactive objective); 2) standardize creation, ingest, and description of new content (a current objective); and 3) develop measures to actively monitor the general condition of the collection (a future objective).

These objectives will build upon the fundamental design of the audit. Auditing digital assets is an effective inventory management and maintenance procedure which can apply not to one "phase" or "stage" but to the entire lifecycle of digital collections. Viewing the output of tools, workflows and policies which emerge as a result of an audit as a flexible and lightweight foundation will enable the type of cumulative curation which is scalable and iteratively improved.

References

[1] NDIIPP Partner Tools and Services Inventory (2010). Available at: http://www.digitalpreservation.gov/partners/resources/tools/index.html

[2] "Archivmatica" Available at: http://archivematica.org/

[3] "Archivist's Toolkit" Available at: http://www.archiviststoolkit.org/

[4] "Curator's Workbench" Available at: https://github.com/UNC-Libraries/Curators-Workbench

[4] Library of Congress (2010). Using PREMIS with METS. Available at http://www.loc.gov/standards/premis/premis-mets.html

Author Biography

Aaron Collie received his M.S in Library and Information Science (2010) with a specialization in the Data Curation Education Program from the University of Illinois. His recent work has focused on developing a value proposal for modifying ETD workflows to acquire and preserve doctoral research data. He is currently the Digital Curation Librarian at Michigan State University.

Lucas Mak serves as the Metadata and Catalog Librarian at the Michigan State University Libraries. He earned his M.S. in Library and Information Science from the University of Illinois at Urbana-Champaign. His recent works have been focusing on metadata workflow automation.

Shawn W. Nicholson has over a decade experience in libraries and has written and lectured on use and reuse of numeric data. His current scholarly activities center on long-term curation for research data. He earned a M.S. in Political Science and an M.S. in Library and Information Science from the University of Illinois Urbana Champaign. He presently holds administrative responsibility for the Digital Information Division of the Michigan State University Libraries.