Accumulating Metadata from Tributary Workflows for More Robust Archival Records

Aaron Collie; Michigan State University Libraries; East Lansing, Michigan
Devin Higgins; Michigan State University Libraries; East Lansing, Michigan
Lucas Mak; Michigan State University Libraries; East Lansing, Michigan
Shawn Nicholson; Michigan State University Libraries; East Lansing, Michigan

Abstract

Metadata generation oftentimes involves not just a single source but rather a tributary system of interoperating technical processes and workflows during capture. New methods of capture such as automated extraction of technical metadata creates a more robust archival record that will improve our ability to curate digital objects.

Recent development roadmaps from two well regarded open source digital preservation systems envision an integration of the Archivematica digital preservation system to prepare information packages for management and dissemination via the Fedora Commons middleware application. MSU Libraries recently piloted a proof-of-technology to transform the technical metadata output from the File Information Tool Set (FITS) utilized by Archivematica for ingesting into a Fedora Common installation. This is accomplished by transforming the metadata output of the open-source Archivematica digital preservation system into the Fedora Commons extension of METS.

This interactive paper will report on possible scenarios for the integration of these two preservation tools including the management of the resultant AIP and DIP; possible changes to metadata generation, indexing and searching; as well as provide observation on the applicability to similar workflows.

Metadata Traditions

Libraries have been involved in resource description for centuries. From clay tablets to filing cards, and from physical media to computerized and online systems, library metadata has been stored with different technology in different times. Due to both the emergence of new types of content and advancement of storage and retrieval technologies, libraries have been creating and storing metadata in different silos, including online catalogs, finding aids, databases, and even repositories for digital collections. Although many of these silos have been built or modified to allow improved communications, each is in some way limited by contemporaneous technologies and infrastructure. In more recent years the type, standards and sources of metadata have become more and more diverse and heterogeneous. Descriptive metadata is no longer the only desired type of metadata. Technical, administrative, preservation, and rights metadata are just a few of the many types of metadata currently in use to describe library content. Each of these broad genres of metadata oftentimes has a specific standard that allows for the nuances of particular content types or specialized languages—other types of metadata attempt to be encompassing and inclusive. Similarly, metadata is now captured through a variety of events in the lifecycle of a digital object: during file creation, in subsequent file edits, via software extraction, social tagging, and more traditional methods such as original cataloging, copy cataloging, and third party metadata creation.

Most libraries have experience working with more than one metadata standard, and oftentimes build systems that support multiple metadata standards. This is because in recent years, MARC (MAchine Readable Cataloging), a dominant encoding standard for bibliographic information, has been giving ground to a wellspring of new standards including Dublin Core, METS, MODS and others. These standards are not exactly replacements, but rather enhancements or alternative that are being implemented and managed in tandem with traditional standards.

In order to identify, capture and curate applicable metadata library systems must expand their capacity for handling complex objects including the plethora of standard and non-standard metadata formats. Authoritative archival records must forgo the traditional mantra “one record to rule them all” and instead leverage systems which can maintain a balance between standardization and innovation by supporting the tributary sources of metadata generation.

Mutable Systems

Fortunately, contemporary repository developers and managers benefit from a maturing landscape where numerous tools exist that address the challenges of metadata handling. While early digital collections management systems suffered from the same “silo” tendencies as traditional library systems, for example: format limitations; database-dependent metadata handling; strictly one-to-one metadata relations; and a plethora of scalability issues recent reports from the field indicate that the second wave of repository platforms have successfully built infrastructure to support the heterogeneity of digital collections—without limitation to organizational make-up.

Many of these systems are built on the micro-services model of digital object handling. This model acknowledges that turn-key repository platforms run the same risk of functional obsolescence as the digital objects they manage. The micro-service philosophy breaks down turn-key repository solutions into individual tasks that are optimized for performance and maintain broader community (rather than individual vendor) support.

It is clear that in just the short time that organization have been managing digital collections, modularity and flexibility are important characteristics of digital systems. Starting from scratch when building digital repository systems is oftentimes infeasible; cultural heritage organizations are typically involved in many pockets of digital initiatives and many have in-house operations that provide valuable functionality or benefit for the organization.

In this regard, libraries and other cultural heritage organizations are well served by mutable systems that grow with and not against contemporary technologies. This architecture will better model the actively changing human systems that interplay with digital collections software and infrastructure. A mutable digital workflow and infrastructure will more easily integrate with extant library operations as digital collections grow to scale.

Project Context

This is certainly true for Michigan State University Libraries. Over the past 15 years MSU-L has built a robust digital storage environment to support digital collections. This storage environment (expressed in Figure 1) offers extensive disk and bit level preservation but is challenged to provide functional preservation; other significant curation challenges such as web access and metadata creation remain manual processes or in-house operations. This strategy has resulted in the formation and support of divisional units with sufficient staffing and expertise to sustain operation. Like many other organizations a build-it-as-you-go methodology has established a strong foundation of human systems, enterprise storage, and operating procedures.

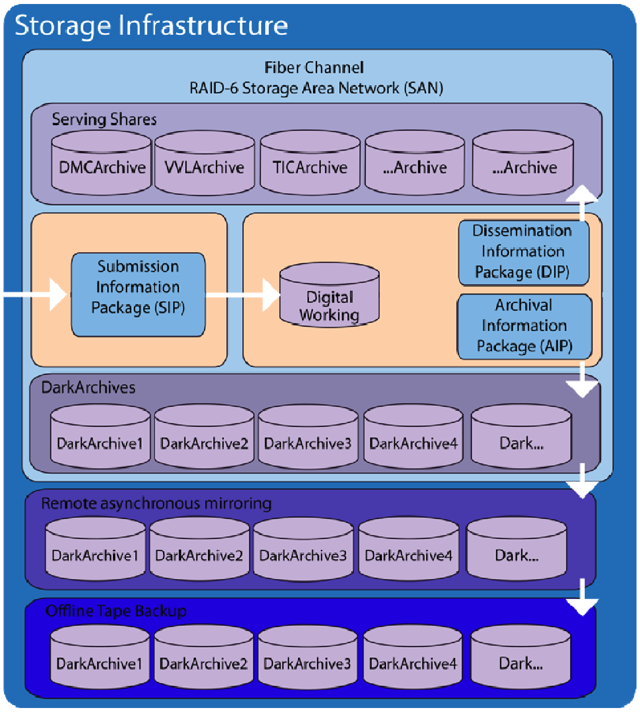


***Figure 1**. Three Tier Storage Infrastructure: Local SAN, Remote SAN, Offline Tape Backup*

A recent audit of digital collections and infrastructure provided a detailed snapshot of the stakeholders, current technology, and collections at Michigan State University Library. The assessment identified a number of 'preservation gaps' within the current system and a report was produced which detailed a series of milestones or functional requirements that were later used to measure software value and performance during an environmental scan of available options. It was clear from the report that two manual processes would benefit most from improvement: automation of ingest into archival storage, and better handling of complex collections.

From these baseline criteria, two software packages were identified as best meeting the needs of the organization: Archivematica as a digital preservation system to prepare information packages for management and dissemination and Fedora Commons as a robust middleware application to improve digital handling of multipart collections (Figure 2 is explained in detail within the Accumulating Metadata section, but is provided here as an overview of desired implementation).

Archivematica is "a free and open source digital preservation system that is designed to maintain standards-based, long term access to collections of digital objects" http://www.archivematica.org. Archivematica's micro-services approach provides an integrated (and customizable) suite of software tools in compliance with the ISO-OAIS model. Furthermore, Archivematica's stated goal is to afford ready access to an agile, lightweight and easy to manage digital preservation tool. In many ways, Archivematica is well suited to projects that enhance its interoperability with other storage and access systems. The micro-services approach to digital preservation allows developers to create workflows that expand or contract, or become more or less automated, in ways that suit the needs and preferences of their anticipated users. While the current release of Archivematica (0.9-beta) allows this level of customization only through back-end changes to the source code, the use of a branching workflow system that passes information from one job to the next along a chain allows new jobs to be dropped in (or removed) at specific points in the workflow without interrupting other jobs in the chain.

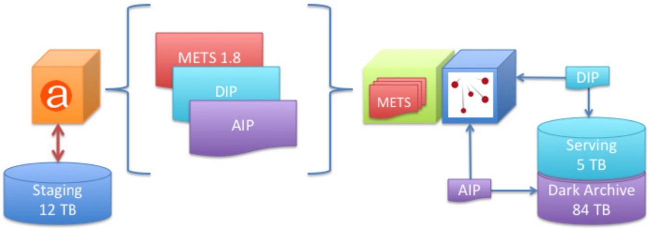


***Figure 2**. Staging, Ingest, Serving and Archiving*

Similarly, Fedora Commons—a well-established open source middleware application offered by DuraSpace—provides a flexible framework suitable for custom repository creation and plethora metadata formats. It was designed explicitly to offer streamlined handling, management, and discovery of digital content. The latest Fedora instance attractively provides supporting services and applications including search, OAI-PMH and RDF support, plus basic preservation workflow. Fedora Commons is well suited to customization and integration with extant workflows by utilizing the digital object model to simplify handling of multipart collections within the complex of library infrastructure.

Tributary Workflows

Michigan State University Libraries (MSUL), like other libraries, has created and stored various types of metadata in different systems and acquired metadata from different sources. Not surprisingly, the bulk of these metadata are stored in the online library catalog as MARC records. Most of the legacy MARC records originated from in-house catalogers or downloaded as copy from bibliographical utilities like OCLC. More recently, the majority of metadata capture is the result of batch-loading vendor and publisher metadata. Other than for-a-fee sources, MSU-L is also harvesting MARC records from the public domain HartiTrust digital collections catalog using the Z39.50 protocol. Additionally, staff and student workers also create descriptive metadata for locally digitized collections. Some of these records can be generated through repurposing existing MARC data. Though some locally digitized items have been cataloged and included into the library online catalog as well, significant amount of these ephemeral materials are accessible only through dedicated websites built on relational databases. Some of the technical and other non-descriptive metadata of digital collections are also stored in relational databases. Nonetheless, a great deal of non-descriptive metadata is embedded in individual digital files. These technical metadata can potentially be extracted using standalone software like JHove, DROID, and File Utility among others. The tributary nature of the current metadata landscape has posed both challenges and opportunities to curation of digital objects.

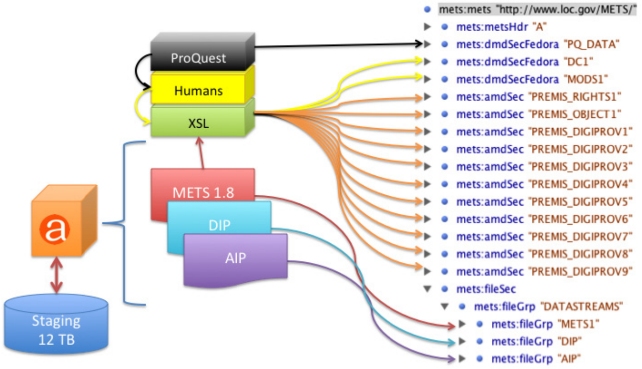
Accumulating Metadata

Michigan State University Libraries desired a system that would streamline the handling of metadata and digital content, remain flexible to support inevitable changes in digital content processing, and disseminate objects in a meaningful manner. To meet these desires, development has focused on:

1. Providing a means to accumulate and manage multiple types and standards of metadata from human systems, vendor systems, as well as technical and software systems
2. Integrating general and specific workflows with extant library operations including ongoing infrastructural build-out
3. Maintain access and present meaningful display of collections while remaining flexible under-the-hood to accommodate changes in collections, infrastructure, or software tools

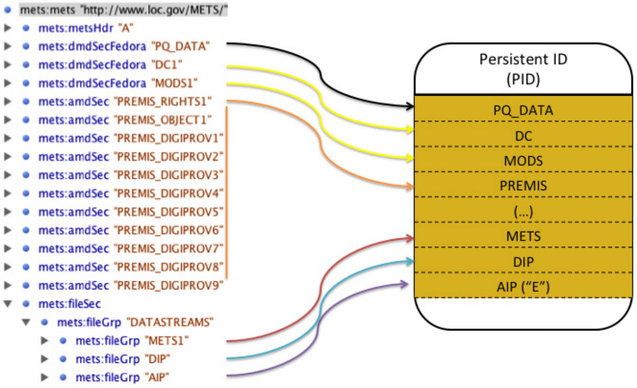
Figure 2 shows a high level model of the prototyped system. While integration of Archivematica and Fedora Commons is not provided out-of-the-box, both advantageously utilize standards and best practices that provide developers with a number of possibilities for interoperability. One particular standard utilized by both systems is the METS metadata standard; a standard specifically suited to describing the transfer of objects within complex systems. Archivematica creates a METS (v. 1.8) record that at a high level describes Archivematica's pipeline style processing as PREMIS events, but can be customized to produce detailed technical metadata due to the FITS microservice. Fedora Commons is able to use a specific extension of METS (Fedora Extension 1.1) in order to ingest items into the digital object model; at a high level, this record is actually describing the ingest process itself. However, despite standardization a number of differences between the two implementations require (in the least) restructuring.

A proof-of-concept was developed which used manual processes (Shell and XSL) to complete a basic transfer of the outputs of Archivematica to Fedora Commons. Archivematica complies with the OAIS model and therefore utilizes the SIP, AIP and DIP terminology. SIPs (submission information packages) are the unprocessed items which initiate the workflow, AIPs are the archival copy of the SIP (stored in the Bag hierarchical structure), and DIP files are derivative access copies of the SIP. At the MSU-L our security and preservation infrastructure requires that DIP content would then be managed internally in our serving environment by an access system (in this case we are using Islandora), while the AIP would be externally read from our DarkArchive preservation environment. By default, Archivematica produces a METS XML file for each information package, which can then be transformed by XSLT to fit the required Fedora Extension of the METS standard. Figure 3 shows the mapping of the Archivematica output into the Fedora Extension of METS.



mets:mets "http://www.loc.gov/METS/"
mets:metsHdr "A"
mets:dmdSecFedora "PQ_DATA"
mets:dmdSecFedora "DC1"
mets:dmdSecFedora "MODS1"
mets:amdSec "PREMIS_RIGHTS1"
mets:amdSec "PREMIS_OBJECT1"
mets:amdSec "PREMIS_DIGIPROV1"
mets:amdSec "PREMIS_DIGIPROV2"
mets:amdSec "PREMIS_DIGIPROV3"
mets:amdSec "PREMIS_DIGIPROV4"
mets:amdSec "PREMIS_DIGIPROV5"
mets:amdSec "PREMIS_DIGIPROV6"
mets:amdSec "PREMIS_DIGIPROV7"
mets:amdSec "PREMIS_DIGIPROV8"
mets:amdSec "PREMIS_DIGIPROV9"
mets:fileSec
mets:fileGrp "DATASTREAMS"
mets:fileGrp "METS1"
mets:fileGrp "DIP"
mets:fileGrp "AIP"
ProQuest
Humans
XSL
METS 1.8
DIP
AIP
a
Staging
12 TB

***Figure 3**. Archivematica Output mapped to the Fedora Extension of METS (v.1)*



mets:mets "http://www.loc.gov/METS/"
mets:metsHdr "A"
mets:dmdSecFedora "PQ_DATA"
mets:dmdSecFedora "DC1"
mets:dmdSecFedora "MODS1"
mets:amdSec "PREMIS_RIGHTS1"
mets:amdSec "PREMIS_OBJECT1"
mets:amdSec "PREMIS_DIGIPROV1"
mets:amdSec "PREMIS_DIGIPROV2"
mets:amdSec "PREMIS_DIGIPROV3"
mets:amdSec "PREMIS_DIGIPROV4"
mets:amdSec "PREMIS_DIGIPROV5"
mets:amdSec "PREMIS_DIGIPROV6"
mets:amdSec "PREMIS_DIGIPROV7"
mets:amdSec "PREMIS_DIGIPROV8"
mets:amdSec "PREMIS_DIGIPROV9"
mets:fileSec
mets:fileGrp "DATASTREAMS"
mets:fileGrp "METS1"
mets:fileGrp "DIP"
mets:fileGrp "AIP"
Persistent ID
(PID)
PQ_DATA
DC
MODS
PREMIS
(...)
METS
DIP
AIP ("E")

***Figure 4**. Fedora Extension of METS mapped to Digital Object Model*

Figure 4 shows the mapping of metadata and digital content into the Fedora Commons digital object model. By default, the Archivematica METS output does not contain descriptive metadata, though it is possible to supply some metadata through the Archivematica dashboard or by supplying a supplemental CSV file. In our test case, we utilized Electronic Theses and Dissertations as our content type, and thus desired to integrate metadata capture with other library operations including original cataloging. For this particular use case, an additional XSL transformation was designed to match dissertation records with library catalog metadata. The result of that process was additional descriptive Dublin Core and MODS records.

Because the proof-of-concept met our expectations for processing digital content, managing multiple content and metadata formats, and improving handling from acquisition to access it was determined that the design should move into the next stage of development by creating an automated workflow as proof-of-technology. This meant modifying the Archivematica software.

To make these workflow alterations, it was first necessary to understand Archivematica's internal data architecture. At its core, the Archivematica workflow is channeled and regulated by the MCP (master control program), a central program written in Python that draws on configuration and processing instructions stored in an MCP database. Each job performed within the Archivematica workflow must be defined within this database in ways that notify the MCP of its location within the chain of jobs, as well as which specific tasks it will perform. The MCP "reads" this information and processes accordingly, following the chain until it reaches an end value of NULL. The processing chain is not strictly linear--user choices made within the Archivematica dashboard allow the chain to branch off in a number of pre-programmed directions. Therefore, any functionality introduced into Archivematica to increase compatibility with Fedora Commons must not apply in all instances, and can easily be bypassed or automated by making adjustments to an XML configuration file.

Once a new job chain link is integrated into the larger job chain, the MCP follows the instructions included in the database to perform command line operations or scripts to move or transform files or data. Any number of processing interpolations are possible, though care must be taken not to disrupt tasks that occur further down the chain.

Thus Archivematica can be customized to complete operations that produce output suitable for ingest to Fedora Commons, by default or by user choice. The level of interoperability possible between Archivematica and Fedora Commons is improved by the flexibility of the Fedora Commons content model. Multiple digital items can be compounded into one object, and data streams can be customized to accept any type of information.

MSU Libraries is eager to share both our progress (code) and our use case (described here) in an effort to advance a fuller integration of Archivematica and Fedora Commons. The primary scripts are currently being refined for contribution to the Archivematica codebase, and the team is interested to learn how they may better collaborate with others involved with similar integrations.

Discussion

Recent discussion on relevant lists and in the wake of several digital preservation conferences indicate that there is continued interest in an integration of Archivematica and Fedora Commons. Current roadmaps from these two well-regarded teams reflect this vision though there has been little (public) articulation of the possible advantages, synergies and/or challenges. It is the hopes of the authors that this paper and use case can provide insight into some of the more intimate issues that might guide efforts towards strategic interoperability:

1. Prescription of Archivematica output to content models and/or solution packs (direct mapping of DIP to content models for greater standardization)
2. Continuation and articulation of Archivmatica's OAIS-based AIP & DIP handling throughout Fedora Commons
3. Interoperability (and GUI management) of Fedora Commons Service Definition & Deployment with Archivematica workflow chains
4. Integration of AIP monitoring, JMS Messaging, and respective (or combined) database rebuilds from Filesystem

The ultimate goal of this project was to implement a process that would handle document and metadata files, utilize the functionality of Archivematica to render information packages conforming to digital preservation best-practices, and adapt its workflow to create packages ready to be ingested into Fedora Commons.

References

[1] J Van Garderen, P. (2010). ARCHIVEMATICA: Using Micro-Services and Open-Source Software to Deliver a Comprehensive Digital Curation Solution. *IPRES2010*. http://www.ifs.tuwien.ac.at/dp/ipres2010/papers/vanGarderen28.pdf

Author Biography

Aaron Collie is the Digital Curation Librarian at Michigan State University Libraries and is responsible for the build-out of a second-generation digital collections preservation and access system. He received his M.S in Library and Information Science ('10) with a specialization in the Data Curation Education Program from the University of Illinois.

Devin Higgins graduated from the Graduate School of Library and Information Science at the University of Illinois in 2012. He is currently Digital Library Programmer at Michigan State University, where his interests include building digital collections that promote user exploration, designing automated data flows, and the digital humanities.

Lucas Mak is the metadata and catalog librarian at Michigan State University Libraries, East Lansing, Michigan. He holds a MS in library and information science from the University of Illinois at Urbana-Champaign. He is interested in authority control, metadata cross-walking, and cataloging workflow efficiency improvement through batch processing and 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.