https://doi.org/10.2352/issn.2168-3204.2018.1.0.27
© 2018, Society for Imaging Science and Technology

Preservation Data Modeling for Systems Interoperability: the Single SIP Model in the Bayou City DAMS

Bethany Scott, Andrew Weidner; University of Houston Libraries; Houston, USA

Abstract

The University of Houston (UH) Libraries made an institutional commitment in late 2015 to migrate the data for its digitized cultural heritage collections to open source systems for preservation and access: Samvera, Archivematica, and ArchivesSpace. In order to ensure that preservation objects can be uniquely identified in Archivematica and referenced/accessed through the other systems, the UH Libraries implementation team has developed a "single SIP" data model in which a digital object's files and metadata are packaged individually prior to Archivematica ingest. The single SIP model provides flexibility in file management, avoids overloading Archivematica's processing capacity, and allows for direct persistent links from ArchivesSpace and Samvera to the preservation objects in Archivematica storage.

Introduction

The University of Houston (UH) Libraries made an institutional commitment in late 2015 to migrate the data for its digitized cultural heritage collections to open source systems for preservation and access: Samvera, Archivematica, and ArchivesSpace. As an initial step in this migration project, the implementation team produced workflows and tools to support preservation and access ingest activities, creating the Bayou City Digital Asset Management System (BCDAMS) ecosystem of modular components that work together to address all aspects of the digital curation lifecycle [1], including minting and resolving unique identifiers, managing controlled vocabulary terms, and assigning standardized metadata to access objects.

One challenge that the team faced was determining which files and metadata should be sent to preservation storage and how preservation packages should be structured in order to comply with UH Libraries' digital preservation policy [2]. We sought to balance requirements for preserving adequate contextual information about the original materials with limitations in Archivematica's processing and indexing capability for large transfers. To that end, we created a single SIP data model in which the preservation files and metadata for individual digital objects are packaged, and the resulting AIPs are aggregated in Archivematica storage through the use of the AIC functionality [3]. This model for preserving objects and metadata individually may be useful to other institutions seeking to scale up their use of Archivematica to preserve large collections of digitized archival content.

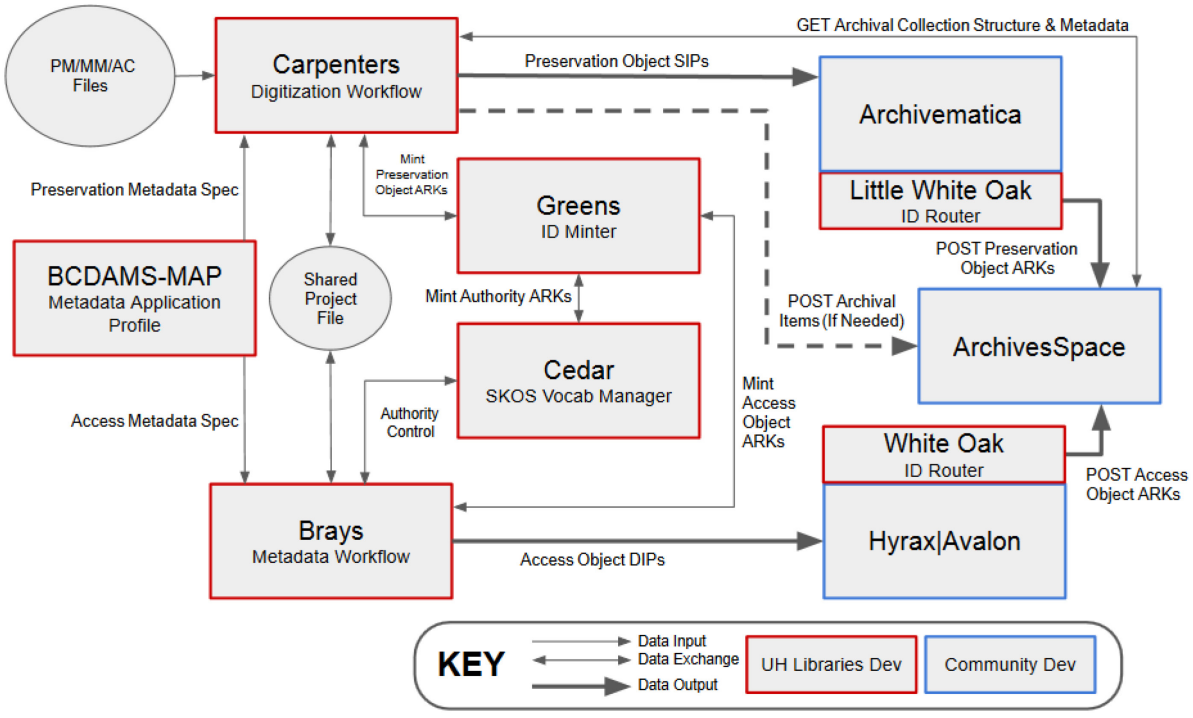

***Figure 1**. BCDAMS preservation and access workflow/architecture*

Digitization Workflow

The overall BCDAMS architecture and workflow for digitization projects is seen in Figure 1. Digitization projects begin in the Special Collections department, where physical materials to be digitized are selected by curators, described in an ArchivesSpace finding aid, and flagged for digitization. Through the use of the ArchivesSpace-integrated Carpenters file management app [4], developed as part of the larger BCDAMS ecosystem, the Special Collections project manager creates a shot list to be handed off to the Metadata & Digitization Services (MDS) department's Digitization Unit along with the materials, allowing digitization technicians to easily identify the archival objects that are included in the digitization project. As imaging is completed, digitization technicians assign preservation master files and access derivatives to those archival objects within the app, and references to the file locations, ArchivesSpace URIs, basic title metadata, and other data points are saved to a Carpenters (*.carp) shared data file for the project.

When digitization is complete and the project is ready to be sent for metadata creation, the MDS Metadata Unit is notified and the project is handed off. Metadata specialists load the shared data file using the metadata workflow app Brays [5], which automatically imports descriptive metadata for each archival object using its ArchivesSpace URI, so that this metadata can be quickly and efficiently reused. Metadata imported from ArchivesSpace includes titles, dates, collection information, and container/location identifiers. The Brays app displays required, recommended, and optional fields and validates the contents of certain fields based on the BCDAMS Metadata Application Profile [6]. Metadata specialists enhance the ArchivesSpace-derived core metadata for each object, adding information to fields such as creator/contributor, description, extent, and subject. Subject headings, including personal names, corporate names, places, topics, and time periods, are created and managed using the Cedar vocabulary manager [7]. As metadata specialists complete their work, they may flag items for rescanning, resequencing, or other filenaming/file management rework.

When descriptive metadata has been entered and saved and any digitization rework has been completed, the project manager reopens the shared data file in Carpenters to export SIPs. At that time, Archival Resource Key (ARK) persistent identifiers are minted for each preservation object and saved in the shared data file to be added to the access objects' metadata. The preservation files, any modified masters, and preservation metadata/submission documentation are exported from Carpenters and automatically packaged for Archivematica ingest according to the single SIP spec (described below). The Greens ID minter app [8] is key to managing related preservation and access objects, providing a unique ARK for each object and a persistent URL that references related objects in our access systems, Hyrax/Avalon and ArchivesSpace. After SIPs have been exported and preservation ARKs have been minted, the shared data file can then be reloaded once more in Brays, adding preservation ARKs to the descriptive metadata for the access objects, which are exported in the appropriate format for ingest into either the Hyrax or Avalon access repository.

Finally, when the access objects have been published in the access repository, single SIPs are ingested to Archivematica storage. Using Archivematica's automation tools, SIPs are queued up to begin the transfer/ingest process without the need for manual intervention or approval for each one. During Archivematica ingest, a new microservice developed by the BCDAMS implementation team manages activities relating to persistent identification of preservation objects. This microservice, code named Little White Oak [9], represents one of the last preservation actions that takes place before AIPs are prepared and stored in the Archivematica storage service, as shown in Figure 2. Little White Oak serves two functions: updating the erc:where metadata for each ARK so that the ARK URL resolves to the package's location in Archivematica storage, and posting the preservation ARK URL to the appropriate archival objects in ArchivesSpace as a new, unpublished digital object.

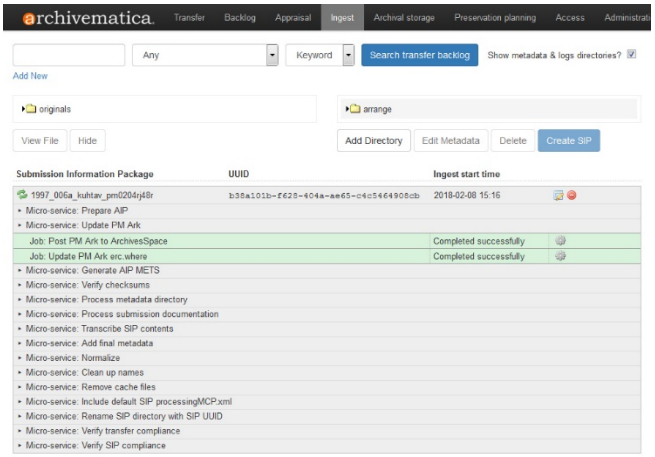

Figure 2. *Little White Oak microservice in the Archivematica interface*

Other Archivematica Implementations

Workflows and documentation on other institutions' implementations of Archivematica provided some guidance in designing the BCDAMS workflow and its system integrations. While Archivematica is a relatively recent option for digital preservation processing, several examples have been released, and for many institutions, development and implementation is ongoing.

At the "Using Open-Source Tools to Fulfill Digital Preservation Requirements" session held at the 13th International Conference on Digital Preservation (iPRES) in 2015, several talks highlighted the use of Archivematica and its integration with various open source archival software tools [10]. Andrew Berger of the Computer History Museum discussed its Archivematica implementation and "the use of other open source tools to prepare packages for submission" to preservation storage [10]. Ben Fino-Radin of the Museum of Modern Art (MoMA) cited long-term viability and sustainability as part of the rationale for selecting open source systems for digital preservation solutions – rationale which resonate with the goals and purpose of the BCDAMS project – and highlighted the use of a locally developed repository management tool, Binder [11], that allows Archivematica to be integrated "with existing proprietary systems deployed internally at MoMA" [10]. Bonnie Gordon described the Rockefeller Archive Center's integration between Archivematica and ArchivesSpace, through which rights information and technical metadata assigned or derived in Archivematica is passed to the appropriate ArchivesSpace resource records.

In 2016, the Bentley Historical Library completed its ArchivesSpace-Archivematica-DSpace workflow integration project [12]. The project goals included facilitating the creation and reuse of descriptive and administrative metadata across

systems and streamlining the process for depositing content into preservation storage. Although these goals mirror the desired outcomes of the BCDAMS workflow, the Bentley's workflow integration focuses on processing born-digital collections which make up a large portion of their holdings. To create efficiencies in processing large born-digital collections, the development work focused on adding an Appraisal and Arrangement tab in Archivematica's web dashboard, allowing users to "characterize distributions of file formats within acquisitions, identify sensitive data, and preview content" by visualizing the outputs of the individual tools that make up Archivematica's transfer microservices [12]. The Appraisal and Arrangement tab also allows for ArchivesSpace descriptive metadata to be created or edited within the Archivematica interface, and for digital objects in the Archivematica backlog to be associated with ArchivesSpace records. In the case of the BCDAMS workflow for digitized collections, the contents of the SIPs are generally less complex than those of born-digital collections, removing the need to conduct file format or sensitive data analyses. Additionally, for BCDAMS digitization projects ArchivesSpace objects must be identified and associated with their digital representations early in the process so that ArchivesSpace URIs can be passed to both Archivematica and the Samvera access repositories. While the integrations created at the Bentley are a mismatch for the BCDAMS digitization workflow described above, the tools in the Appraisal and Arrangement tab may be useful in processing UH Libraries' born-digital content in a future phase of the BCDAMS project.

At the University of Saskatchewan Library, the Archidora project integrating Archivematica and the Fedora access repository Islandora "enables the automated ingest into Archivematica of objects create in Islandora" [13]. In this workflow, access objects are first ingested to Islandora, which triggers preservation files to be uploaded to Archivematica. As a result, digitization staff are not required to interact with Archivematica. Furthermore, in this workflow Archivematica can take advantage of other standards built into Islandora, including metadata files encoded in MODS, METS, and PREMIS. While the use case outlined in a 2018 Code4Lib article, providing public access to digitized cultural heritage content in a Fedora-based repository, most closely matches the BCDAMS digitization workflow, we are not as concerned about removing digitization staff from the preparation of SIPs. Also, our projects are designed to take advantage of existing ArchivesSpace metadata rather than MODS or METS.

BCDAMS Preservation Data Model

While developing the BCDAMS digital projects workflow, we encountered several challenges that led to the creation of the single SIP model for preservation. The primary problem arose when we tested the Archivematica ingest process for a package representing an entire digitized collection. The full-resolution TIFF preservation master files for 167 objects included in the SIP are approximately 400 GB, and the files as originally packaged and transferred were arranged in the SIP in a hierarchical folder structure that mirrored the archival arrangement of the original materials in the provenance collection. The large file size of the transfer and the long file paths that resulted from multiple layers of nesting in the hierarchy caused Archivematica to stall or fail completely, frequently during the indexing microservice that allows stored AIPs to be searchable in the Archivematica storage service dashboard.

We debated adding a "post-processing" step in the workflow that would allow preservation SIPs to be restructured after export as needed to reduce their overall size and/or complexity of arrangement. However, a post-processing workflow raised challenges with minting preservation ARKs in such a way that they could be automatically (or at least efficiently) passed to the Metadata Unit for inclusion in the descriptive metadata for the appropriate access objects. Similarly, we struggled with trying to design a workflow in which multiple preservation objects are packaged together and a unique identifier is assigned to that package, since the resulting preservation identifier would not provide a 1:1 match with other uniquely identified objects (e.g., access objects in Samvera or archival objects in ArchivesSpace).

Because the overall BCDAMS ecosystem relies on a system-of-record style architecture in which different versions of files and metadata only exist in the appropriate repository – Archivematica for preservation files, Samvera for access files and descriptive metadata, and ArchivesSpace for archival hierarchies – the ability to automatically add ARKs that point to the associated objects in other systems is vital to ensuring the repositories are well integrated while reducing the need to manage duplicates in different repositories. In order to both assign unique identifiers for individual preservation objects (to be referenced in our access systems) and to avoid slowdowns or failures in Archivematica's transfer and ingest processes, we created new specifications for the structure, contents, and naming conventions for single-object SIPs, seen in Figure 3.

```
transfer directory: [project slug]_[pm ARK]
|___  objects
  |___  file.mxf (preservation master)
|___  service
  |___  file.dv (modified master)
|___  metadata
  |___  metadata.csv
  |___  submissionDocumentation
    |___  Carpenters project file and logs
    |___  PBCore.xml metadata file
    |___  readme.txt (if applicable)
```

***Figure 3**. Single SIP structure and contents for digitized film/video object*

The files that make up single SIPs are created early in the digital projects workflow, when archival collections are digitized and files are assigned to the appropriate archival objects. The novel aspect of this workflow compared to our previous tests is the way in which digitized collections are structured and packaged individually for preservation ingest. While imaging is taking place, the digital collection is managed in aggregate, as an overall project, using the Carpenters app. Upon export, Carpenters automatically creates SIPs representing each individual object in the project. Processes that take place at this time include: minting a preservation ARK for each digital object, moving files and submission documentation into the appropriate folder structure that Archivematica expects for ingest, and creating the metadata.csv with information imported from ArchivesSpace.

Fields in the metadata.csv include:

- dcterms.title (ArchivesSpace object title)
- dc.date (ArchivesSpace date)
- uhlib.aSpaceUri (ArchivesSpace URI)
- dcterms.identifier (preservation ARK)

- dcterms.isPartOf (Cedar collection term ARK)
- uhlib.note (human-readable collection title from ArchivesSpace)
- partOfAIC (AIC batch identifier entered in Carpenters interface)

In order to aggregate the hundreds or thousands of single SIPs that may be contained in one digitization project, we take advantage of Archivematica's AIC functionality, through which "multiple AIPs can be intellectually combined into one AIC, or Archival Information Collection" [3]. In this workflow, the AIC represents all the preservation masters created in one digital project or batch. The addition of the "partOfAIC" field in the metadata.csv allows preservation administrators to enter an AIC identifier (based on the archival collection/accession number and the batch number) upon exporting the SIPs from Carpenters. When the packages have been ingested and stored in the Archivematica storage service, the AIC identifier is indexed, and packages associated with that batch are returned in a keyword or phrase search of the stored AIPs. Including this AIC or batch identifier accounts for archival collections in which accruals of digitized content from a single archival collection are made over time. In the future, we may need the ability to download the packages or re-run microservices for just one batch from the larger provenance collection, and the AIC functionality allows for that grouping of AIPs in the Archivematica storage service.

For archival collections, we also create one SIP which contains the ArchivesSpace-exported EAD finding aid for the collection, so that the intellectual arrangement and hierarchy of the digital objects is preserved along with the objects themselves. The finding aid will be exported from ArchivesSpace after preservation objects and their ARK URLs have been posted to ArchivesSpace by the Little White Oak microservice, so that the EAD file contains an ordered/nested list of references to the preservation identifiers included in the AIP name for each stored package. Once the EAD SIP has been stored in Archivematica, it can be added to the AIC for the batch of digital objects associated with the appropriate archival collection.

Conclusions and Future Work

Two discrete avenues for further development arise from the single SIP model. First, we plan to explore potential interface improvements for searching and browsing AIPs in the Archivematica storage service's web dashboard. Features we would like to add to the dashboard include: the ability to customize the number of results per page (currently, only 10 stored AIPs are displayed per page), a checkbox or other method of selecting multiple stored AIPs, and the ability to conduct batch processing (such as creating AICs or re-running microservices through the Archivematica re-ingest feature). The availability of a simplified AIC creation process and the ability to conduct preservation actions on batches of stored AIPs are crucial to the long-term viability of the single SIP model, since the model will not scale without these features as thousands or tens of thousands of individual/single-object AIPs proliferate in preservation storage.

Second, we plan to investigate further improvements to the user interface of the Carpenters app with the goal of creating a more streamlined process for assigning files to the appropriate digital and archival objects and allowing for further time-saving automation such as automatic file renaming to UH Libraries' naming conventions. Currently, files may be assigned either by dragging and dropping them from the filesystem to Carpenters, or by moving files on the filesystem into a folder structure representing the project's archival objects selected in Carpenters. Especially for preservation packages, a more streamlined way of assigning submission documentation files, which are often repeated for every object in the project, could improve efficiency in preparing a project for SIP export.

Finally, a separate but related goal is developing a workflow to accession and preserve born-digital archival materials using the Carpenters app to package them for Archivematica. Several differences between digitized and born-digital content present themselves. One difference in the overall way that born-digital materials are handled is in the processing configuration required in Archivematica. By contrast to digitized collections, in which items are imaged to FADGI-based local specifications in which file formats are deliberately selected because they are preferred as long-term preservation formats, born-digital content received from donors is often somewhat disorganized and may include a large range of different file formats, not all of them preservation-worthy, necessitating normalization during Archivematica processing. These differences in processing requirements are accounted for through the installation of a second Archivematica pipeline which is specifically configured to handle born-digital materials. Furthermore, born-digital content may not be practical to add to the access repositories, since the data model of those repositories doesn't allow for items to be described in aggregate, and collections of born-digital content may include thousands or tens of thousands of files with little existing descriptive metadata, which may best be viewed, searched, and browsed in a hierarchical arrangement. With this in mind, a single SIP for born-digital content could include a copy of the original files, any files normalized to preservation formats, and metadata representing a single digital storage item, such as a floppy disk, CD/DVD, or hard drive.

Although these represent significant differences between the overall workflow for digitized content and born-digital collections, using Carpenters for packaging born-digital materials could provide similar benefits to the digitization workflow in that it allows for the automatic reuse of previously entered metadata. In order to take advantage of Carpenters' integration with ArchivesSpace, individual digital storage media could be added to ArchivesSpace as archival objects during processing, allowing processors to complete data entry work such as entering label text or dates found on individual disks or other items to be transferred. As these storage media are imaged or securely transferred to working space on a network drive, they could then be assigned in a Carpenters project, where label/title data and dates of creation entered in the finding aid would be automatically added to the Carpenters-generated Archivematica metadata.csv in the exported SIPs. Once SIPs are transferred to the Archivematica backlog, preservation administrators may wish to take advantage of the reporting features for born-digital content available in the Appraisal and Arrangement tab, especially if the collection has not yet been fully processed. Reports on file formats and/or sensitive personal data present in the born-digital files may be used to inform decisions such as archival arrangement and access restrictions in the finalized archival collection.

References

[1] Weidner A, Watkins S, Scott B, Krewer D, Washington A, Richardson M. Outside The Box: Building a Digital Asset Management Ecosystem for Preservation and Access. Code4Lib

Journal. [Cited 2018 March 5]; Issue 36. Available from: http://journal.code4lib.org/articles/12342

[2] UH Libraries Digital Preservation Policy. [Cited 2018 March 5]. Available from: http://libraries.uh.edu/wp-content/uploads/DigitalCollectionDevelopmentPolicy.pdf

[3] AIC, Archivematica 1.6 documentation. [Cited 2018 March 5]. Available from: https://www.archivematica.org/en/docs/archivematica-1.6/user-manual/archival-storage/aic/

[4] Carpenters Github repository. [Cited 2018 March 5]. Available from: https://github.com/uhlibraries-digital/carpenters

[5] Brays Github repository. [Cited 2018 March 5]. Available from: https://github.com/uhlibraries-digital/brays

[6] BCDAMS Metadata Application Profile. [Cited 2018 March 5]. Available from: https://vocab.lib.uh.edu/bcdams-map

[7] Cedar vocabulary manager. [Cited 2018 March 5]. Available from: https://vocab.lib.uh.edu

[8] Greens Github repository. [Cited 2018 March 5]. Available from: https://github.com/uhlibraries-digital/greens

[9] Little White Oak Github repository. [Cited 2018 March 5]. Available from: https://github.com/uhlibraries-digital/little-white-oak

[10] Using Open-Source Tools to Fulfill Digital Preservation Requirements, contributed talks. [Cited 2018 March 5]. Available from: http://oss4pres2015.web.unc.edu/contributed-talks/

[11] Binder documentation. [Cited 2018 March 5]. Available from: http://binder.readthedocs.io/en/latest/

[12] Eckard M, Pillen D, Shallcross M. Bridging Technologies to Efficiently Arrange and Describe Digital Archives: the Bentley Historical Library's ArchivesSpace-Archivematica-DSpace Workflow Integration Project. Code4Lib Journal. [Cited 2018 March 5]; Issue 35. Available from: http://journal.code4lib.org/articles/12105

[13] Hutchinson T. Archidora: Integrating Archivematica and Islandora. Code4Lib Journal. [Cited 2018 March 5]; Issue 39. Available from: http://journal.code4lib.org/articles/13150

Author Biography

Bethany Scott is Digital Projects Coordinator at the University of Houston Libraries where she currently works on projects to implement systems such as ArchivesSpace, Archivematica, BitCurator, and Omeka. As a representative of UH Special Collections, she contributes knowledge on digital preservation, born-digital archives, and archival description to the DAMS Task Force. She received an MS in Information Studies from the University of Texas at Austin.

Andrew Weidner is Metadata Services Coordinator at the University of Houston Libraries where he oversees the Metadata Unit and is the project manager for Bayou City DAMS implementation. Prior to joining UH, Andrew worked as the New Mexico state project coordinator for the National Digital Newspaper Program at the University of North Texas (UNT). He holds master's degrees in Library Science from UNT and History from the University of Texas of the Permian Basin.