The evolution of digital preservation at the Getty Research Institute: How workflows have evolved in the past five years to address our varied needs

Laura Schroffel
Digital Archivist,
Special Collections
Getty Research Institute
Lschroffel@getty.edu

Teresa Soleau
Head, Library Systems & Digital Services
Getty Research Institute
Tsoleau@getty.edu

Lorain Wang
Digital Archivist,
Institutional Archives
Getty Research Institute
Lowang@getty.edu

CONFERENCE THEME ADDRESSED

• Our paper addresses the topic of mapping out sustainable digital preservation approaches and communities by describing how we have handled tiered digital preservation requirements; ie. specificity needed for some items versus large-scale ingest and processing.

DESCRIPTION

• This short paper will discuss the challenges and practical decisions that have driven the Getty Research Institute to develop new workflows and adapt existing ones for born-digital materials.

TYPE OF SUBMISSION

• Short paper

ABSTRACT

Five years ago, at the Getty Research Institute, we implemented the Ex Libris product Rosetta to preserve and provide access to our digital collections, mainly digitized resources from the Institutional Archives and Special Collections, reusing workflows from our former DAM to create METS records and populate the system. Over time it became clear that new processes were required for born-digital material and, furthermore, we would need to accommodate the contrasting needs of Special Collections and Institutional Archives to preserve their born-digital content. For these reasons we have adjusted those original workflows, creating new tools and developing additional deposit processes along the way, with a very small team.

Working within the limitations of our digital preservation system and available resources has been challenging. While we have tried to standardize the workflows involved in digitization and digital preservation, we sometimes make adjustments to work with our varied and diverse collections. For example, due to differences in volume and access needs, the processing of Special Collections content tends to involve extensive, hand-edited metadata work while Institutional Archives files are processed in the aggregate. We try to make decisions that follow best practices for digital preservation, but ultimately we're committed to getting the material into the system quickly, often prior to full processing, so it can be identified, validated, stored redundantly, and made accessible if appropriate, as backlogs continue to grow. In our efforts to preserve as much as possible now, we are likely creating more work for ourselves in the future, a possibility we confront regularly when weighing the importance of preservation needs versus access needs.

In this paper, we will discuss these challenges and focus on the practical decisions we have had to make when developing new deposit workflows, or adapting existing ones, on content that is in some way different than what came before. Our discussion will describe our different processes for digitized material and born-digital material both from Institutional Archives and from our Special Collections. We hope to provide guidance for those who are moving forward with "good enough" approaches instead of waiting for some magical day when we have all the time, staffing, and expertise to give our full attention and care to each filestream.

1 INTRODUCTION

At the Getty Research Institute in Los Angeles we have been using Rosetta, the digital preservation solution from Ex Libris, since 2012. The system is based on OAIS (Open Archival Information System) principles and uses many of the standard community-developed digital preservation tools and metadata formats such as Jhove, DROID, and METS. The Rosetta data model is based on PREMIS semantic units. For Objects, these units are Intellectual Entity (IE), Representation, File, and Bitstream. Although Rosetta is a vendor solution, and so in some ways proprietary, the vendor works closely with Rosetta customers to continually enhance the product following best practices for digital preservation.

Prior to implementing Rosetta we used another Ex Libris product, a digital asset management system called DigiTool. When migrating from DigiTool to Rosetta in 2012, we configured our new system to work with the digital material we already had. Since the majority of that material was digitized and our main focus had been providing access to the material through our public discovery system, our configuration choices and deposit workflows were based on assumptions inherent in that kind of content, over which we have control of format and file names. This is not the case for born-digital content, which is often far more complex, not only in terms of the aforementioned areas, but also in regards to level of description and access needs. We, therefore, had to create new processes and reassess our system configuration to accommodate the needs of these new materials.

Staffing levels have also shaped our processes. During the early years of implementing a digital asset management system we had a software developer to focus on writing code to get our content into the system but we no longer have that resource. We also do not have a dedicated staff member for digital preservation and, as our digitization program and acquisition of born-digital materials have both grown over the years, we find smaller slivers of our time can be dedicated to formal preservation planning and policy drafting.

In this paper, we will describe some of the issues we encounter in trying to preserve our resources and how those issues manifest differently for digitized content and born-digital content, from both Institutional Archives and Special Collections.

2 DEPOSIT METHOD

A brief description of the baseline deposit methods will help expose some of the issues with the extensibility of our processes. For digitized content we create what we call a “shot list,” which is a simple Excel file that contains a list of the files included in the IE, labels for each of the files to be used in the structure map (these are the labels that appear in the viewer for users), a hierarchical representation (parent-child relationships), and file formats for each of the representations. We create a preservation master file, an access file and, in most cases, a modified master file (also known as the mezzanine-level). Each of the different representations of the file have the same file name with a different suffix added to the end.

To create a Rosetta-compliant METS XML file[1], we use a custom Perl script and configuration file to combine the “shot list” with a Dublin Core XML file. The METS file and filestreams are then deposited into Rosetta where the system enriches the METS with additional data and stores it in the permanent repository. All further changes to IEs in the repository are tracked in these IE-level METS files.

Born-digital Special Collections materials are also deposited into Rosetta via a METS deposit method, but we had to create a custom XSL stylesheet to make the METS because the Perl script made assumptions about file naming, file formats, and types of representations that do not hold true for our born-digital files.

When trying to deposit born-digital content for the first time using our existing scripts we realized that our use of the file name string for multiple purposes would not work. The file name string is used as a file ID in the METS record and so needs to comply with METS guidelines for IDs. When we create the file names during digitization, we can follow those rules, but that is not the case with born-digital content and so our deposit scripts either break or create non-compliant METS that need to be hand-edited. For example, file IDs in METS must begin with a letter or underscore (not a digit), and can only contain letters, digits, periods, hyphens and underscores.[2] Therefore, they cannot lead with numbers or have spaces, periods, diacritics or special characters of any kind - all common occurrences in file names of born-digital files. For this reason the Special Collections born-digital METS creation stylesheet uses arbitrary unique IDs as file IDs rather than deriving them from file names.

Another complication is that, for METS deposits in Rosetta, files within a SIP are stored flat in a single folder so files must have unique file names. This requires us to assign a prefix containing the file’s related item number and then a unique numeral if further disambiguation is required (e.g. CM1_filename.txt). We keep a record of any differences between original file names and preservation versions of file names by using the original names as labels in the METS structure map.

One other significant difference between born-digital deposits and digitized deposits is that for born-digital material a mezzanine-level file is generally not produced. The IEs contain a preservation master representation and sometimes a set of access derivatives. For Special Collections, in cases where original order can be improved or manipulated to better reflect the already extant structure of a finding aid, only the structure in the access file representation would be changed to make the intellectual structure more legible. Preservation file structure always reflects original order.

The METS deposit method proved to be impractical for use with Institutional Archives since their born-digital files number in the tens of thousands or larger and manual creation of “shot lists” is

[1] http://www.loc.gov/standards/mets/profiles/00000042.xml

[2] <METS> Metadata Encoding and Transmission Standard: Primer and Reference Manual, page 86 https://www.loc.gov/standards/mets/METSPrimer.pdf

not feasible. To handle this volume, we require procedures that retain directory structures and are less time-intensive than hand-editing METS files. Institutional Archives files are, therefore, deposited using Rosetta's CSV method, where metadata is submitted as a CSV (produced using a script from Archives New Zealand to convert metadata generated by DROID) and the filestreams are packaged as a zip. Rosetta unpacks the zip, and the system generates a METS file that reconstructs the original directory structure of the deposited files using the file paths contained in the CSV as a logical structure map.

Even with our depositing procedures established, we continued to encounter additional problems due to the "messiness" of born-digital files that required us to make additional adjustments.

For example, we retain the original file names of born-digital content whenever possible, even if they do not conform to recommended conventions. Changing a name could break links in databases, website files, and other documents. When dealing with thousands of files, it is often difficult to determine the implications of minor name changes. Over time, however, we learned that certain characters, such as periods followed by a space, can cause system problems during the deposit and long-term storage processes. We consequently developed a list of characters that we strip or replace in names prior to deposit and track these changes and the original names in the METS. Diacritics also cause system problems specific to our CSV deposit method, but at this time Institutional Archives has chosen not to remove them from file names. We have set the files aside until the vendor resolves the issue. This decision will need to be reevaluated at a certain point if the problem still hasn't been fixed and the need to preserve the files outweighs the desire to maintain the integrity of file names.

The less standardized nature of born-digital files also makes it difficult to fully take advantage of beneficial features of the system. Rosetta has a validation stack that uses DROID to identify the file format and a suite of tools to extract technical metadata from some of the files, validating them against the standards for those formats. This function is useful in the case of digitization. Since we are creating the files, we want to know if they are corrupt or non-compliant in some way so that we can change our processes to create valid, well-formed files.

In contrast, this feature impeded our ability to deposit born-digital files into the system. Early in the process of depositing Institutional Archives' born-digital files, we knew that we would encounter a significant number of file format validation errors. We originally decided that we would address these issues prior to deposit and spent a great deal of time testing tools that identified files that were not well-formed. This led to the inevitable discussion (which we perhaps should have had earlier) of how we would handle the malformed files, which, in turn, led to our conclusion that we did not have the capacity or level of expertise to "fix" hundreds or thousands of problem files. In the end we chose to ignore the errors, letting the files pass into the permanent repository without alerting us, but to have the details of the errors recorded in the METS.

This approach has worked well for Institutional Archives content, which lives in its own repository in Rosetta. In contrast, Special Collections born-digital material exists in the same repository as the digitized content. Until we have time to develop procedures to accommodate the conflicting needs of digitized and born-digital content in the same repository, we need to coordinate the deposit of Special Collections material (turning error handling rules on or off) to ensure that file format or validation errors do not prevent born-digital content from going into the repository.

3 ACCESS

As mentioned at the beginning of the paper, our original use case for digitized content assumed that public access of some sort (either fully open or restricted to on-site use) was a priority. In order to deliver files to end users through our public discovery system, we include access derivatives that can be rendered by standard web browsers. These access files are easily created as part of our standard digitization process.

To provide public access to born-digital content, however, additional steps are required. Files need to be evaluated and cleaned of personal identification information. Special Collections converts files to access formats, requiring files to be assessed at the file level in a time consuming process. In some situations, due to lack of software or technological access, we cannot convert files at all and, in those cases, we preserve the files in Rosetta but do not provide access to them. As there is always a concern that data is irrevocably changed, Special Collections is transparent with users and reports which formats were converted and includes a warning to users that access representations do not always exactly mirror originals.

For Institutional Archives, which tends to deal with larger volumes of files that are often for internal use only, conversions are done on a case by case basis. We most commonly convert sets of files composed entirely of video or audio content that can be transformed using batch processes. For other content, we make files accessible in the original format, regardless of web browser compatibility.

Another system limitation we have encountered in making files accessible is that an access set must be deposited along with the preservation set in order for the access set to reflect the original directory structure. Adding a logically structured representation at a later point is not possible in the system. Since born-digital

content is typically highly structured, we sometimes need to decide to put in the extra time now to deposit a suppressed access set along with the preservation set, even when the files must remain restricted for the time being. This requires us to make decisions about access before we are ready to do so.

While a large portion of Institutional Archives' born-digital materials is restricted from public view, the ability to provide staff access to files is a major concern. Such requests have been rare so far, but we know that the Rosetta viewer is not ideal for displaying our content and we do not have the resources to convert all our files to formats that can render in the viewer. Thus, when staff transfer files to us, we suggest that they keep a locked set of the files on their network drive for access or we will export a zipped set of the files from Rosetta upon request. This is not a sustainable practice and we do not recommend it to others.

4 CONCLUSION

Digital preservation is an iterative and active process. While depositing the content and providing access to it are important steps, our work is far from over. There is still a great deal that needs to be addressed.

Modifying existing IEs in Rosetta is cumbersome and we need to decide how to handle changes to a set of files in Rosetta where material is further processed or additional content is added at a later date. This issue is especially complicated for Institutional Archives as we have made a decision to deposit files in Rosetta prior to removing unnecessary files or rearrangement due to backlog.

There are also differences in the way the departments view iterations of born-digital content. While Special Collections has decided to deposit processed files as an access set retaining the unprocessed files as a preservation set, Institutional Archives views the processed files as a new version of the preservation set. Whether we replace the existing preservation set or deposit the new set as an entirely new IE is a question we have not yet tackled.

On the access front, we are trying to decide whether to describe the preservation or access set in the IE-level Dublin Core record that is used in our discovery system. This is a larger issue when the access set represents only a portion of the full set of files. Should we describe the files we are providing access to or the files we are preserving? This dilemma is a direct result of our decision to use Rosetta as both a preservation and asset management system where the Dublin Core record serves dual purposes.

Looking forward, we anticipate that we will need to migrate content from one file format to another. Rosetta has tools to assist in that effort, though we have not yet tested them. These tools rely on the files having a specific format ID in the system, which is not true in cases where we have prioritized getting large amounts of content into the system over addressing format issues at the file level. We know that we will need to run the format identification process on some sets of files in the future to facilitate this work.

The variety of born-digital formats and our decision to ignore format issues during deposit presents additional challenges related to our status as founding members of IIIF (International Image Interoperability Framework). Our goal is to provide access to our materials through that specification, instead of through Rosetta, in the next year as we are eager to separate preservation from access for our publicly accessible content. For digitized content the transition will be fairly straightforward and largely automated because the digital objects are consistently structured and the file formats are known. However, when it comes to providing access to born-digital content through IIIF we expect to encounter more bumps in the road and realize that it may not be possible for us to transition without significant manual effort.

The challenges and decision points outlined in this paper are not unique to our institution. While established policies and access to tools and resources vary for each institution, comfort with imperfection is always necessary to make progress in digital preservation. As illustrated, our processes for depositing content into our preservation system have grown organically over the years. Flexibility and the willingness to adjust workflows as needed is essential. Even with forethought, we have found ourselves addressing issues as they arise and then developing policy and documentation along the way. This agility is only possible if you learn to recognize when enough energy has been spent analyzing or evaluating an issue. It is important to consider best practices when making decisions but they should not stand in the way of taking action. Ideally, system limitations should not drive preservation policy, but sometimes it is impossible to work around these constraints, even as systems evolve. While it is easy to be distracted by the minutiae, one should not lose sight of the overarching goal.

In a world where technology changes so rapidly and everyone is looking for the next flashy app or visualization it's hard to advocate for a process that has the sole aim of keeping things exactly the same. Still, we are engaged in the work and determined to get our materials into a managed system where they can be identified, monitored over time using checksums and provenance events, and stored redundantly on geographically dispersed and technologically diverse media.

Table: Workflow comparisons

	Digitization	**Special Collections born-digital**	**Institutional Archives**
Deposit	METS	METS with unique File IDs	CSV
Representation types	Preservation (always), mezzanine (typically), access (typically)	Preservation (always), access (typically)	Preservation (always), access (rarely)
File naming	Created by institution using local guidelines	Created by institution for access. Retains originals with invalid characters removed for preservation	Retains originals with invalid characters removed for preservation. Sometimes adds file extension.
Access file name example	gri_980065_b01_f01_01.jpg	gri_980065_CM01_01.jpg	filename-acc.mp4
Access via Rosetta viewer	Typically full access	Limited access	Extremely limited access