MODULE 12

PRESERVING DIGITAL OBJECTS

Erin O'Meara and Kate Stratton

SOCIETY OF
American
Archivists

Appendix B: Case Studies

Case Study 1: Rockefeller Archive Center

By Sibyl Schaefer, former Assistant Director, Head of Digital Programs

Please provide a brief description of your organization (or library or archives within the organization) and a brief description of the digital preservation program/repository (team composition, systems, tools, and infrastructure).

The Rockefeller Archive Center (RAC) is an independent repository and research center dedicated to the study of philanthropy and the diverse domains shaped by philanthropy. It was established in 1974 to assemble, preserve, and make accessible the records of the Rockefeller family and their wide-ranging philanthropic endeavors (including the Rockefeller Foundation, the Rockefeller Brothers Fund, and the Rockefeller University). Today, the Center's growing holdings include materials from numerous non-Rockefeller foundations and nonprofit organizations, making it a premier center for research on philanthropy and civil society. It is also a major repository for the personal papers of leaders of the philanthropic community, Nobel Prize laureates, and world-renowned investigators in science and medicine.

The archives side of the RAC consists of five main areas: donor relations and collection development, reference, processing, archival services, and digital programs. The digital programs team consists of four full-time employees (one assistant director, one archivist, and two assistant archivists) and is responsible for maintaining and developing online access tools, digitization, archival technology support, and digital preservation. Although responsibilities vary from team member to team member, in general there are about 1.5 full-time employees dedicated to digital preservation tasks.

We use a variety of systems and tools to aid our digital preservation work. We have purchased a Forensic Recovery of Evidence Device (FRED)[55] and use that, as well as a Device Side Data FC5025[56] to image

[55] http://www.digitalintelligence.com/products/fred/, captured at https://perma.cc/49CX-TXZC.

[56] http://www.deviceside.com/fc5025.html, captured at https://perma.cc/KU4N-AU42.

digital media. We are in the process of purchasing a Kryoflux[57] for imaging as well.

We use Archivematica to ingest and package our AIPs. Our main AIPstore is file system based (not a repository like DSpace or Fedora). It is backed up to tape regularly, and we are planning on storing tapes offsite in a newly built storage facility. We check the fixity of the AIPs at regular intervals using the Ace Audit Manager.[58] We are also currently in the process of joining the MetaArchive[59] network to store selected AIPs in a LOCKSS[60] system. We initially made the decision not to adopt a repository system because we couldn't afford the technical expertise required. Since then, our in-house technical expertise has been expanded and the expertise required has been lowered by systems like Hydra[61] and Islandora,[62] so we may revisit that decision in the future.

Walk us through your ingest and AIP creation workflow.

In general, after we receive a data transfer or image a disk, we run virus checks, extract the directory structure of the disk image or data transfer, generate checksums, and then move the data to a staging server. A staging server is a separate area where we store data while preparing it for archival storage. Some of the data we receive from donors is transferred in bags,[63] in which case we leave them as bags for ingest. Prior to ingest, we assess any access or use restrictions that pertain to the materials. We try to do this at the largest aggregation possible, and if necessary, divide aggregations up by their restrictions.

Our ingest method follows the Archivematica ingest workflow, which processes materials through a series of microservices: file format identification, assignment of unique identifiers, metadata extraction, etc. We review rights and restrictions for the materials so we can enter PREMIS information during ingest. During this process, Archivematica packages the ingest into a bag (if it's not already bagged)

[57] http://www.kryoflux.com/, captured at https://perma.cc/V9MN-2MDN.

[58] https://wiki.umiacs.umd.edu/adapt/index.php/Ace:Main, captured at https://perma.cc/J3RT-XKBK.

[59] http://www.metaarchive.org/, captured at https://perma.cc/36Y7-HQSB.

[60] http://www.lockss.org/, captured at https://perma.cc/RE25-XXFN.

[61] http://projecthydra.org/, captured at https://perma.cc/P9ZH-D3F7.

[62] http://islandora.ca/, captured at https://perma.cc/T488-4WXH.

[63] https://github.com/LibraryOfCongress/bagit-java, captured at https://perma.cc/9QPD-BKHW.

and generates an AIP that includes the standard bag manifest files, as well as log files generated during the virus scan and file format identification processes, any metadata that was submitted with the package, any additional submission documentation, the actual objects themselves (and normalized versions, if those were created), and a METS file containing the structure of the objects and PREMIS events, agents, and rights information. See the Archivematica wiki for a more detailed description of workflow elements.[64]

At the end of ingest, the AIP is stored on our Isilon storage server[65] and a DIP is generated. We have connected Archivematica with our Archivists' Toolkit (AT) database so that we can match digital objects with their corresponding AT components, and automatically generate AT digital object records with the correct rights and restrictions information.

As mentioned earlier, selected AIPs will be copied on to a different server in order for the MetaArchive crawler to ingest them into the MetaArchive LOCKSS system. Costs for this service are not insignificant, so we have decided to submit only materials in which the digital version is considered the preservation copy. For example, we have terabytes of digitized microfilm. Because we consider the microfilm to be the preservation copy, we will not ingest that data into MetaArchive.

Currently we are only packaging digital objects that have been processed through Archivematica. We're planning on using the backlog function in Archivematica for unprocessed materials so we can manage them in the same system and identify any necessary preservation issues.

Provide a generic or specific example of an AIP (in a diagram or schema). How does your AIP represent or account for the three main elements described in Lavoie and Gartner's diagram?

- **content information**
- **preservation description information**
- **packaging information**[66]

[64] https://www.archivematica.org/wiki/AIP_structure, captured at https://perma.cc/A98G-T2WB.

[65] Isilon is a networked attached storage server sold by EMC.

[66] Brian Lavoie and Richard Gartner, "Preservation Metadata (2nd edition)," DPC Technology Watch Report 13-03 (Digital Preservation Coalition, May 2013), 14. doi: http://dx.doi.org/10.7207/twr13-03.

The Archivematica wiki shows how our AIPs are structured:

Content information

- *Content data object*: Original objects stored in the data/objects directory
- *Representation information*: File format identifiers are captured using FIDO[67] during ingest and recorded in the METS/PREMIS file. The appropriate PRONOM IDs are also stored.

Preservation Description Information

- *Reference:* Archivematica assigns UUIDs to each object in the ingest, as well as the ingest itself. These are all listed/organized in the METS <structmap>.
- *Context*: Relationships are also listed in the METS <structmap>. We have also added some complexity to this information by altering our data ingest structure when several digitized images comprise one PDF, so that the master TIFFs, adjusted JPEGs, and access PDFs are all related correctly. We are looking forward to the Yale-sponsored Archivematica development allowing ingest of disk images, in the hope that it will extract filesystem information automatically.
- *Provenance information*: This is probably a weak spot for us, because the finding aid in AT has the authoritative version of this information. I have debated including a finding aid in the submission documentation, but our ingests are often broken up into groupings smaller than the collection level, so we would have to ingest the same finding aid numerous times. This is not a big deal for small finding aids, but some of our finding aids are quite large. We are titling the ingests by the unique finding aid ID so there is some link between the two.
- *Fixity information*: This information is auto-generated by Archivematica and stored in the METS file as well as the BagIt manifests.

[67] https://github.com/openplanets/fido, captured at https://perma.cc/JM6C-WUXN.

Packaging Information

The Archivematica files we generate cover the major packaging information reflected in the Lavoie and Gartner publication, the weak spot being the descriptive metadata, which is captured in the AT. We are looking to integrate our descriptive information more. We have done the first step—connecting our archival data system (the Archivists' Toolkit) with Archivematica to ensure that object UUIDs are linked and recorded in the AT. The next step is to do the reverse: record descriptive information currently in the AT in the AIP. Archivematica does provide for descriptive metadata ingest, but it is based on Dublin Core, and all our descriptive data is in EAD, so there are hierarchical data mapping issues to resolve. We've held off delving into this because we're planning on migrating to ArchivesSpace and potentially adopting systems like Islandora or Hydra, both of which are being integrated with Archivematica and may alleviate this gap.

One area we would like to improve is the recording of preservation actions taken prior to ingest. So, if we imaged a disk, we want to document what software and hardware we used to create the disk image, or if we extracted files from an image, etc. We currently record this type of information in a separate database, but ideally it would be stored in the METS/PREMIS record.

How do you perform preservation management activities on these objects? Specifically, how do you ensure the integrity of the objects in your care?

If needed, files are normalized during ingest. We store one copy of the AIP in our AIPstore and run regular fixity checks on them. In the very near future, some of our AIPs will also be ingested into MetaArchive for geographic redundancy as well as regular fixity verification. We still have a lot of work to do as far as disk imaging, description, etc., but once we have those priorities worked out, the plan is to create file format audits to identify at-risk formats. Because our ingests have been only of digitized items, I am comfortable holding off on this. The priority now is to get data off of obsolete media and migrate those files as needed (or desired).

What are some of the challenges you face preserving digital objects? What are some next steps and new features you want to add to your digital preservation program?

Our biggest challenge is simply being a small team with big goals. Prioritizing those goals is essential, but it also means we have a list of tasks we know we have to address but cannot currently commit to. Our main focus now is surveying collections for digital media and then imaging that media. At the same time, we are setting up workflows for the transfer and ingest of electronic records received from donors, and also processing and describing electronic records. We also are trying to integrate the management of digital records throughout the archives, so that those currently accessioning paper records will be accessioning digital, those processing paper will process digital, etc. Incorporating digital workflows will require training across other units in the archive, especially with the addition of new tools and software for staff to use.

My next steps are to continue to automate tasks, provide different levels of access within the same system (for donors, for archivists, for researchers), manage descriptive data more effectively, and develop workflows for disk image ingest. As mentioned earlier, we plan on reviewing repository systems like Islandora and Hydra, probably within the next two years.

What do you look for when evaluating tools and developing workflows for preserving digital objects? What type of requirements are essential to achieve your goals?

We are very much an open source shop and try to implement open source tools and solutions whenever possible. I feel the transparency of open source programs is essential in long-term digital preservation. I also look for the widespread adoption of a program or tool, sponsorship by a notable organization, and fairly recent development work done on it. BagIt is a good example of both of those—widely used (enough that we can request bags from vendors) and sponsored by the Library of Congress. Documentation of the tool—how it works as well as how to use it—is also important. Tools that archivists outside of the digital team will be using have to be user-friendly. This is a big reason that I decided to use FTK for processing rather than other open source options.

As far as developing workflows, I aim for automation of repetitive tasks, simplicity, and ease of use.

Do you have any advice for repositories just starting out in digital preservation? What are some of the first steps that someone could take?

I have to agree with the OCLC reports on getting started with born-digital materials:[68] the first step should be to identify and separate digital media from your collections (creating separation documentation, of course), and then get that data (preferably as disk images) onto a backed-up server. Next steps include virus checks, the creation of fixity information, and file format identification. Reviewing the results of those three activities should give you an idea of the potential value of the material as well as potential risks and/or preservation issues.

One thing I strongly advise against is simply adopting some sort of preservation system right off the bat. Preservation is a core archival function, and I really worry that institutions that contract with these types of services are simply handing it off instead of building long-term institutional knowledge and policies on how to manage digital materials. I am constantly reviewing our tech support's decisions to see if they make sense from an archival standpoint, and I find things that need modification all the time. For example, the Linux arguments used when mounting a disk for virus checking—I had numerous arguments that I added in order to ensure the integrity of the data was protected. I have the ability to do this because I can access these commands in my workflow or in my systems. How does that work when the system is closed? And what happens when no one is responsible for building the knowledge within the institution to ask that type of question?

[68] Demystifying Born Digital series, http://oclc.org/research/activities/borndigital.html, captured at https://perma.cc/K27D-9FCN.