MODULE 12

PRESERVING DIGITAL OBJECTS

Erin O'Meara and Kate Stratton

SOCIETY OF American Archivists

Case Study 2: University of North Carolina–Chapel Hill

By Jill Sexton, former Head of Digital Research Services; Meg Tuomala, former Electronic Records Archivist; and Gregory Jansen, former Lead Repository Developer

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 University of North Carolina at Chapel Hill is a public university supporting 78 bachelor's, 112 master's, 68 doctorate, and seven professional degree programs through 14 schools and the College of Arts and Sciences. More than 29,000 undergraduate, graduate, and professional students learn from a faculty of 3,600. UNC Chapel Hill Libraries is a large research library with a staff of around 300, and collections of 7.4 million volumes, 4.5 million microforms, and 25 million manuscripts.

At UNC Libraries, digital preservation efforts are coordinated by the Digital Preservation and Stewardship Committee, whose charge is to align information and activities related to all aspects of digital preservation, for example, the creation and review of policy documents, the management of perpetual access files for licensed content, and the creation of guidelines for donors and collectors of digital content. Furthermore, these efforts are distributed across several library departments including Library and Information Technology, University Archives and Records Management Services, the Wilson Special Collections Library, and Library Preservation.

Development of a preservation repository was an early priority for UNC Libraries. Starting in 2006, the Library collaborated with campus partners, especially faculty and graduate students from UNC Chapel Hill's School of Information and Library Science, to develop specifications for the new repository. Software development began in 2008, and in 2010 the Carolina Digital Repository (CDR) began accepting submissions. The repository accepts born-digital special collections, digital research data collections and other scholarly output, and digitized library materials.

Digital Repository Services is a sub-unit under Library and Information Technology, and is fortunate to have a unit of four

full-time staff dedicated to the development of the CDR.[69] Staff from Digital Repository Services in the Library and Information Technology department work closely with archivists from University Archives and Records Management Services and Special Collections Technical Services to determine requirements for system functionality, set enhancement priorities, and move materials into our preservation environment.

Following standards developed in the reference model for Open Archival Information Systems (OAIS), the CDR uses the Fedora Commons Repository as an object, model, and services provider and iRODS as a distributed storage and preservation system. We use a locally developed tool, the Curator's Workbench,[70] to facilitate the management, staging, description, and ingest of large batches of objects destined for the CDR.

The CDR also supports other means of ingest, such as automated ingest of content from aggregators via SWORD, and patron-initiated ingest of materials such as ETDs and research posters, via web forms. We offer a range of access controls based on campus LDAP[71] groups, which allow us to specify embargoes and access controls at the data stream level to any object in the repository.

Walk us through your ingest and AIP creation workflow.

Our high-level process is:

Submission preparation tool (Curator's Workbench and others)

Submission service (Admin web application)

- Validation (persistence module)
- Transformation to ingest batch (persistence module)
- Routine pre-processing of ingest batch (persistence module)
- Queues ingest batch with Fedora ingest service

Fedora ingest service

- First come, first served batch ingest
- Handling of diverse SOAP faults, protocol and service exceptions

69 https://cdr.lib.unc.edu/, captured at https://perma.cc/7V9K-KTBG.

70 https://github.com/UNC-Libraries/Curators-Workbench, captured at https://perma.cc/H79H-4YM4.

71 Lightweight Directory Access Protocol (LDAP), https://msdn.microsoft.com/en-us/library/aa366075(v=vs.85).aspx, captured at https://perma.cc/K95V-RDCQ.

- Verification of each ingested object, including fixity
- Container updates (persistence module in Services web application)
- Sending a JMS message when done
- Emailing submitter when done
- Fedora ingest sequence

Replication in iRODS

- Objects and datastreams copied to redundant storage systems

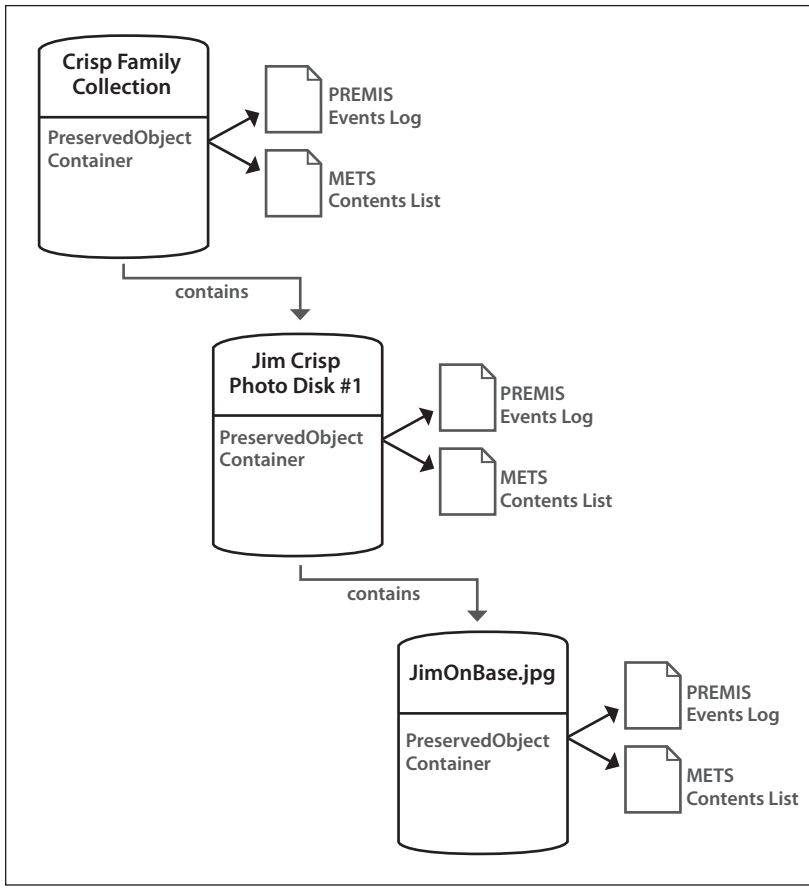
For a more detailed AIP creation workflow, visit the CDR website, http://cdr.web.unc.edu/ingest-overview/.

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**[72]

72 Lavoie and Gartner, "Preservation Metadata," 14.

Figure 3. AIP Model



Source: https://cdr.web.unc.edu/aip-description/, captured at https://perma.cc/8KK7-MPXJ.

The model shows a sample archival information package (AIP) with a simple hierarchy. It is composed of the top-level object, the mid-level object, the smaller discrete digital object, and their associated metadata and datastreams.

The top-level object allows multiple digital objects to be grouped together intellectually; within the top-level object are smaller digital objects, often mapping to real-world constraints. It is important to note that both Crisp Family Collection and Jim Crisp Photos Disk #1 are digital objects themselves and containers of subsequent digital objects. As such, both have associated METS contents lists detailing

the structure of the digital objects contained within. The lowest level of an AIP always contains individual files. All three digital objects making up the AIP make use of PREMIS events logs to record preservation events that occur to the digital objects at various levels.

While the entire structure is considered a single AIP, it can be made up of multiple submission information packages (SIPs). For example, if a user were to create a SIP called Jim Crisp Photos Disk #2 from a second disk image, it could be ingested and added to the Crisp Family Collection AIP, and then both Jim Crisp Photos Disk #1 and Jim Crisp Photos Disk #2 would be considered parts of the same AIP. That is to say, an AIP can be drawn at different levels. On one hand, individual objects can be preserved as individual AIPs; generally, though, AIPs hold an aggregated collection of objects. Aggregated AIPs are preferred because then not only are the individual objects preserved, but the relationships between objects and their underlying structure are also preserved. This is useful from at least two points of view: from a digital preservation perspective, the concepts of authenticity and trustworthiness are enhanced if an entire structure is preserved. Along the same lines, from an archival perspective, having structure and relationships preserved conveys a sense of context for the objects.

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

The CDR uses iRODS (integrated Rule Oriented Data System)[73] to implement its preservation storage environment. A key feature of iRODS is its ability to automate file operations. Digital objects and associated metadata files written to the CDR's storage infrastructure are automatically replicated across our storage grid and onto archival tape. One copy of every file is stored on a server administered by the Library, one copy is stored on a server administered by the campus's Information Technology Services, and the archival tape copy is stored offsite in a geographically separate location. We also automate file characterization, checksum generation, virus scans, thumbnail creation, and access copy generation. Replication and fixity checks are performed on a quarterly basis for every file in the repository, and

73 https://www.irods.org, homepage captured at https://perma.cc/BDU5-LELC.

results of these checks are logged in the repository. Preservation events for each object are recorded in PREMIS.

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?

One of the challenges we face in our efforts to preserve digital objects is prioritization, that is simply deciding which files to preserve first. Like many institutions, the amount of digital content we aim to preserve exceeds our storage and hosting capacity. Though storage has become more affordable over the past several years, hundreds of terabytes of storage is still not cheap. We believe defining preservation levels for certain digital objects, file types, and collections, and prioritizing these based on risk and long-term value, would allow us to more effectively and efficiently preserve a larger quantity of digital content over the long term.

Another challenge is hardware migration and system updates, which always loom in the not-so-distant future. Unlike a bricks and mortar building that persists for decades with regular maintenance and updating, digital preservation systems essentially have to be torn down and rebuilt every five years. In order to maintain a strong and trustworthy digital preservation system, purchasing new servers, renegotiating contracts for storage, and evaluating new software happens regularly. In a digital preservation system there is near-constant evaluation of new tools, and pieces of the infrastructure need to be upgraded on a regular schedule.

In the future, we'd like to add more normalization activities into open formats as well as virtualization as an alternative for preserving application environments.

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?

When evaluating tools we look for open-source tools that have grown robust user communities around them. We chose Fedora and iRODS as core components of our system not only because of the functionality they support, but because of the community around the tools. Our colleagues are our most valuable resource.

When we first started building our digital preservation program in 2008, there weren't a lot of tools available to support the workflows we identified in our environment. As a result, in 2010 we began development of the Curator's Workbench, an open-source tool that supports the preparation of digital collections. It was specifically designed with archivists in mind, incorporating features that support archival practice, from accession through arrangement, description, and SIP preparation.

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

Consult with a wide range of constituents at your institution to make sure all of their needs are going to be met—including archivists, librarians, preservationists, IT staff, and potential depositors. Start small. Consider the staff and expertise you have on hand and don't plan to implement processes that you can't manage. It's possible to incrementally scale up your preservation efforts as you build capacity and expertise in your organization.

Remember that digital preservation is not just a single event, it's an ongoing process. Plan for the long-term sustainability of your program. If you plan to host your system on local hardware, be sure to budget for hardware migration every three to five years.

Don't go it alone. Try to get involved in some form of collaborative group. Some cooperative initiatives to examine include the following organizations:

- Internet Archive[74]
- HathiTrust[75]
- APTrust[76]

74 https://archive.org/.
75 http://www.hathitrust.org/.
76 http://aptrust.org/.