OSSArcFlow

Digital Dossier

By Matthew Farrell and Noah Huffman
Duke University

OVERVIEW

Digital Curation at Duke University Libraries (DUL) impacts a number of staff members mainly across departments. In the David M. Rubenstein Rare Book & Manuscript Library (RL), a digital records archivist is tasked with digital curation tasks with 100% of his time. The archivist for metadata, systems, and digital records spends roughly 50% of his time on digital curation. Processing archivists spend around 5% of their time on digital curation activities on average (i.e., one processing archivist may spend significantly more than 5%, while others spend less). Likewise, staff members in the collection development department probably spend, on average, 5% of their time on collecting activities related to digital materials. The DUL Information Technology Services department is the other department where a number of staff are tasked with digital curation. The Digital Production Center (DPC) employs four full time staff (FTE) to digitize materials from DUL collections, of which RL collections is the largest source. The Digital Curation and Production Services (DCPS) unit employs four FTE as well. Finally, the Software Development and Integration Services (SDIS) unit employs seven FTE, though their portfolio supports activities beyond digital curation scoped for the purposes of this project. Note: beyond the scope of digital curation for OSSArcFlow is the Data Visualization Services unit, which offers curation and services around research data licensed, created, or otherwise held by units at Duke University.

EDUCOPIA INSTITUTE

DUL employs several discovery systems for library materials. A public catalog serves up catalog records for physical and digital materials; MARC records for this system are created using OCLC Connexion and Aleph cataloging software. Finding aids for manuscript and archival collections are created in ArchivesSpace, though completed finding aids are exported as EAD and hosted in a homegrown finding aid web platform. Component of the Digital Repository Program also serve as discovery systems. The Duke Digital Repository (DDR) is a Samvera repository with a Fedora 3 backend and holds a small amount of born-digital materials and a larger amount of digitized materials. It is viewed as the end goal storage for both types of digital content. DukeSpace, an implementation of DSpace, serves up scholarly communications, faculty publications, and electronic theses and dissertations. Finally, Tripod 2 is a legacy homegrown platform maintained for digitized materials created before the DDR existed. Finding aids and catalog records that describe digital materials point to DDR, DukeSpace, or—for digital materials that are either not accessible via the web for privacy reasons, or have not yet been ingested into DDR—URIs for staff-accessible storage.

These discovery and storage systems are supported by members of the ITS department. A single systems administrator supports ArchivesSpace, though this is only a piece of his portfolio. SDIS and DCPS staff develop and support DDR, DukeSpace, and the non-repository preservation storage. As for staff computing, a Desktop Support unit of three support most staff computers, while an additional two support staff members support Specialized Computing Environments (SCE). SCE machines include the digitization equipment used by the DPC, as well as two electronic records acquisition stations used for acquiring and processing born-digital materials. SCE support for these machines is limited to software updates and initial hardware configuration; day-to-day usage and support is self-managed by the respective users of those machines.

Documentation about the systems used for digital curation are scattered across a variety of platforms. Documentation related to the digital repository program is created in and communicated via a Confluence and JIRA instance. Cross-departmental projects use Basecamp to communicate and collaborate. Some documentation still exists in DUL's SharePoint instance, and intra-departmental documentation is often stored on a given department's share drive. In addition to Basecamp, Slack is used by various staff members and teams. Finally, every staff member in DUL uses email regularly.

Born-digital collections comprise 31 terabytes (TB) of data. Materials digitized by vendors or in- house by Rubenstein Library staff comprise 54 TB. Materials digitized by the DPC for the Digital Collections comprise 132 TB of data.

DIGITAL CURATION ACTIVITIES

The categories of digital content for OSSArcFlow purposes are: 1) born-digital materials acquired by RL in the collection of manuscript and institutional archives, 2) materials digitized in house or by vendor for patron requests or preservation, and 3) materials digitized in house or by vendor for the Duke Digital Collections program. A clear divide exists between digital curation activities for born-digital materials and those digitized from physical objects at Duke. Partially, this is due to how the two programs evolved: born-digital, in its earlier days, was the province of one RL staff member, while the digitization effort at Duke (Duke Digital Collections) has always involved multiple staff members from across departments.

Major phases in the curation lifecycle for born-digital materials include acquisition and basic reporting, arrangement and description, generation of derivatives for access or preservation, preservation storage, publication of descriptive record, access to materials via the DDR or in the reading room.

Acquisition and basic reporting includes disk imaging or extracting logical files from media, calculating fixity information, running PII scans, running virus scans, extracting file and/or filesystem metadata, and arranging materials and metadata into a pre-storage SIP. Arrangement and description includes analyzing the contents of the digital materials for additional arrangement (if needed), and reusing the extracted metadata to create archival description in ArchivesSpace. Generation of derivatives can also happen at this step, though some derivatives are generated upon ingest to DDR. Preservation storage includes transferring materials to preservation storage servers and checking fixity. From these servers, materials may remain or ingested into DDR. After publishing the descriptive record(s), researchers may make requests for access to digital materials, which are then retrieved from storage and made available on a secure reading room computer.

The tools used in this workflow are Windows and BitCurator-based tools, and include those for disk imaging and logical file extraction, metadata extraction, description, discovery, access, and storage. One acquisition workstation dual boots between Windows and BitCurator, while the other workstation is a Windows machine that virtualizes BitCurator with Hyper-V. For disk imaging, RL staff usually use FTK Imager and less frequently rely on FC5025, Kryoflux's DTC application, Guymager, and CDRDAO. Logical file extraction tools include OSFMount or BitCurator's mounting scripts coupled with TeraCopy, rsync, or TSK Recover. Metadata extraction and reporting is handled almost exclusively in BitCurator, although fixity information using Hashdeep in whichever OS used to create the disk image or acquire the logical files. Tools used for metadata extraction and reporting in most cases are Siegfried and Brunnhilde, bulk_extractor, fiwalk, and ClamAV. For collections with large digital image or audiovisual components, MediaInfo or EXIFtool are used for additional file characterization.

ArchivesSpace is the system of record for archival description, particularly in aggregate. Staff create this description either through the ArchivesSpace web interface, via spreadsheet import facilitated by ArchivesSpace plugins, or via the ArchivesSpace API. Description is synthesized from the extracted metadata. Access to archival description comes from the public catalog and published EAD.

For digitized materials, the curation lifecycle includes a selection process, additional description of the physical collection, the creation of a digitization guide (spreadsheet), the linking of digitized objects in the repository to the collection's archival description, and publication of the digitized materials and derivatives to the web. Selection for digitization starts with a proposal, usually submitted by a curator. If approved by a Digital Collections committee, the physical collection is assessed to ensure that it has been described to a level conducive to item-level digitization.

Once the archival description is satisfactory, RL staff create a digitization guide including descriptive metadata, which DPC staff enhance with metadata about the digitization process. This digitization guide is used to load metadata into DDR once digitized objects are ingested. An automated, on-demand service creates links between digital objects in DDR and related archival description in ArchivesSpace. Digitized materials are made available through the DDR public interface as well as through the finding aid interface.

Preservation storage at DUL takes two forms: the Samvera-based DDR and networked storage provided by campus IT. DDR has web interfaces for administration and access, while networked storage provides staff with file system access. Both storage solutions have onsite backups, as well as synchronize to DuraCloud.

The roles in DUL that directly contribute to digital curation workflows are:

RUBENSTEIN LIBRARY
- Digital records archivist
- Archivist for metadata, systems, and digital records

DIGITAL PRODUCTION CENTER
- Digitization Specialist—Still Image Head
- Digitization Specialist—Audio
- Digitization Specialist—Video
- Digital Collections Intern

DIGITAL CURATION AND PRODUCTION SERVICES
- Metadata Architect
- Digital Repository Content Analyst (x2)
- Head

SOFTWARE DEVELOPMENT & INTEGRATION SERVICES
- Digital Projects Developer (x2)
- Digital Repository Developer (x2)
- Senior Applications Analyst
- Head

GOALS FOR DIGITAL CURATION

RL staff would like to see a clearer and more seamless method for repurposing technical and descriptive metadata between systems. This includes metadata extracted during born-digital processing or digitization to ArchivesSpace, from ArchivesSpace to DDR, and vice versa.

Second, we would like to reduce the amount of duplicative work. For example, staff should not have to manage the same metadata in three different systems. Similarly if metadata exists in one system, it does not necessarily need to exist in another system.

Finally, staff would like to not over-describe digital archival content as a prerequisite for ingest into DDR. Often the best information about a single digital object is already present in the object's archival metadata and it should not require staff to apply additional description. Altering descriptive conventions to facilitate digitization or ingest of born-digital materials is not scalable if we are to provide responsible preservation and access to all of our digital content.