OSSArcFlow

Digital Dossier

By Josh Hogan and Christine Wiseman
Robert W. Woodruff Library, Atlanta University Center

OVERVIEW

Established in 1982, the Atlanta University Center Robert W. Woodruff Library is unique on a number of fronts. It is an independent, non-profit academic library and research center providing information services to the world's largest consortium of Historically Black Colleges and Universities (HBCUs): Clark Atlanta University, the Interdenominational Theological Center, Morehouse College and Spelman College. The Archives Research Center's (ARC) history dates back to the establishment of the Collection in 1925 under the auspices of Atlanta University's Trevor Arnett Library "Negro Collection." The AUC Woodruff Library has over 12 years of experience developing digital services, programs, and collections that expand access to hidden primary-resource collections.

Within the AUC Woodruff Library, the Digital Services Department (DSD) is responsible for managing and implementing digital curation and preservation activities including digital conversion projects, providing access to digital content, as well as library systems administration. DSD members work collaboratively with staff of the Archives Research Center since the bulk of the digital content originates from their collections, and with the Office of Computer and Information Technology (OCIT). In the DSD, the Department Head, Assistant Department Head, and Digital Initiatives Librarian all take active roles in ensuring that data curation tasks are carried out.

EDUCOPIA
INSTITUTE

OCIT staff provides expertise with helping set up and troubleshoot open source software platforms such as Omeka and BitCurator. ARC assists with setting policy and determining digitization and digital preservation priorities. OCIT is the Library's primary source for technology support. They are a contract IT organization, based in Virginia. Three of their staff members have offices on site, mainly for day-to-day desktop support. The Virginia staff members, who come to Atlanta monthly, provide more in-depth assistance both on-site and over the phone.   OCIT provides server space both on-site and in Virginia and have assisted us in setting up storage with Amazon AWS. DSD also works with a contract developer on a project basis when needed. At this stage it would be difficult to determine an exact percentage of time for each individual involved, but that may be a fruitful exercise going forward.

DIGITAL CURATION ACTIVITIES

Currently at AUC Woodruff Library, our digital collections are made publically available via several platforms. The chief platform is our institutional repository, an instance of Bepress's Digital Commons. The amount of digital collections hosted on that platform has grown rapidly in the past two years, and includes Electronic Theses and Dissertations, the open access scholarly production of our member institutions, student work, and, increasingly, digitized archival collections from the Archives Research Center. Our fastest growing set of materials stems from an NEH grant received in 2015 entitled, Spreading the Word. Other digitized archival collections, including historic catalogs, yearbooks, photographs, audio files, and video files are also on Digital Commons. At present, Digital Commons hosts over 8,800 digital items, which have been downloaded more than 859,000 times. The AUC Woodruff Library also hosts two instances of CONTENTdm for digital collections material.

One instance hosts the HBCU Library Alliance Digital Collection, a collection of primary resources from 23 HBCU libraries and archives that is comprised of over 16,000 images. The other CONTENTdm instance houses 60,000 digitized images from the Morehouse College Martin Luther King, Jr. Collection, that are available only in the reading room of the Archives Research Center.

Our digital exhibits also include digitized archival materials, and these are hosted on an instance of Omeka. Currently, we have four digital exhibits drawing from multiple archival collections. We are also in the process of adding an exhibit based on digitized materials from the Spreading the Word grant to be published in Spring 2018.

In addition to these systems, AUC Woodruff Library uses ArchivesSpace to create and maintain all finding aids for archival collections. ArchivesSpace serves as the backend; the front end is provided by XTF, and many digital objects are linked within the finding aids.   AUC Woodruff is currently exploring adopting the ArchivesSpace public interface as updates make that more feasible.

Currently, our digital collections material exceeds 40 TB, including both master and derivative files. Large, digitized AV files and grant funded content are backed up in the cloud via an Amazon Snowball into Amazon Glacier. Our main categories of digital content are digitized archival materials, born-digital archival materials, digitized scholarly communication from member institutions, born-digital scholarly content from member institutions, and born-digital institutional records and photographs. The AUC Woodruff Library has been involved in digital preservation activities since 2010 when it joined the MetaArchive Cooperative on behalf of the HBCU Library Alliance.

Since then, 17 collections from the HBCU LA Digital Collection have been ingested for long term preservation. In 2016 the library formed an interdepartmental Digital Preservation Working Group (DPWG) to address digital preservation in a more systematic manner.

The DPWG completed a digital preservation policy and three year plan, tested tools including BitCurator, established workflows for born digital content and set up cloud storage for master images with AWS.   In addition, digital objects are categorized into three tiers reflecting the priority in our Digital Preservation Policy. Content selected for ongoing preservation supports the teaching and scholarship of the Atlanta University Center and is evaluated based on funding, degradation and obsolescence risk factors, the collection development policy, and retention schedules. The Library was very pleased to be invited to participate in OSSArcFlow as it dovetailed nicely with our digital preservation plan.

GOALS FOR DIGITAL CURATION

- Create a Data Management Services plan and perform outreach to the member institutions about services we can offer.
- Consolidate content management systems. The AUC Woodruff Library is conducting an an evaluation of content management systems with the goal of migrating to a new platform in 2018-19.
- Begin a web archiving program. This program will initially harvest the websites of the Library, CAU and ITC with plans to expand to include Morehouse College and Spelman College as well as student groups and organization.

OSSArcFlow

Digital Dossier

By Paul Kelly
District of Columbia Public Library

OVERVIEW

DC Public Library employs two Digital Curation Librarian FTEs, and one Digital Projects Intern, whose hours vary but average at around 15 per week. 100% of the DPI's time is spent on digital curation activities. Of their 40 hours per week, DCLs spend roughly 80% of their time on digital curation, although, broadly speaking, the remaining 20% might be considered digital curation- related. Although other employees are somewhat in the orbit of these activities, they are considerably more ancillary to their core duties.

DC Public Library's Information Technology division has recently been phasing out local application support. Special Collections, therefore, relies mostly on a patchwork of hosted services to meet its digital curation goals. That said, IT still plays an important role in that they ensure that both DCLs possess admin-level user accounts on their desktop and laptop computers to allow testing of open source software that does not require a server component. IT of course also provides basic hardware support.

DIGITAL CURATION ACTIVITIES

DC Public Library utilizes three public discovery systems: Sirsi, ArchivesSpace, and CONTENTdm. Although we are currently working to ensure that metadata in one system references corresponding metadata in another (for example, a catalog record would point to a finding aid, and the finding aid would point to digital objects), that project is not yet complete.

EDUCOPIA
INSTITUTE

Internal documentation about collections and activities is generally shared via Google Drive, where we can comment, collaborate, and track changes, but is also stored locally on the library intranet.

DC Public Library's digital collections are comprised primarily of digital surrogates, although we are moving toward accessioning more born-digital material. Our non-public Preservica instance (which includes both Glacier and S3), contains 59 digital collections comprised of 357,810 objects, and takes up around 8.2 terabytes of storage. Formats included in preservation storage are TIFF, PCM, WAV, and MOV. These digital collections are mirrored, with some exceptions, in CONTENTdm, where access copies reside in PDF, MP3, and JPEG formats. To compare, current CONTENTdm storage usage sits as 65.05 gigabytes.

The major phases of the digital curation lifecycle at DC Public Library vary depending on whether material is born-digital, or digitized. Born-digital material undergoes pre-acquisition assessment before it is accessioned. Accessioning involves a member of staff, usually a Digital Curation Librarian or Archivist, entering donation data into a standard Google form, which then forms the basis of an ArchivesSpace accession record. For digitized surrogate materials, a new accession is generated from the record that pertains to the source collection. For digital material on physical media, that media is then imaged in the BitCurator environment, and reports are generated. Images and reports are stored in their entirety in Preservica, while selected documents are migrated to a standard format (usually PDF) and made available publicly through CONTENTdm. Digitized materials are almost always created by a vendor. When vendors return material to us, master and access copies are provided. The Digital Curation Librarians either create or approve collection metadata, and ingest masters to Preservica and access files to CONTENTdm.

One exception for born-digital material is web archive data, which is harvested via the Archive-It service, stored by Internet Archive, and described at both Accession and Resource levels in ArchivesSpace, as well as at the website level in Archive-It. Broadly speaking, almost all spreadsheet metadata is edited in both Excel and OpenRefine, and XML is manipulated with a variety of Python scripts. DC Public Library defines its different categories of digital content as audio, video, multipage document, single page document, web archive, and other (which pertains to types that have been processed but for which no standard internal workflow is established, such as disk images or email archives).

GOALS FOR DIGITAL CURATION

DC Public Library's immediate digital curation goals are, one, to increase our internal capacity to digitize or migrate multiple media types, two, to implement a new digital collection management system that handles both access and preservation, and three, to actively seek out more born- digital collections for acquisition.

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.

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.

OSSArcFlow

Digital Dossier

By Dorothy Waugh
The Stuart A. Rose Manuscript, Archives, and Rare Books Library
Emory University

OVERVIEW

Almost everyone at the Rose Library interacts with our digital assets in some capacity, whether they process born-digital materials, process digitization requests for patrons or exhibits, or assist researchers with accessing digital materials in the reading room. In particular, our processing and accessioning archivists work with born-digital collections as and when they are acquired and processed.

The following positions have a percentage of time allocated to digital curation work:

- Digital archivist: Manages our born-digital collections (100%)
- Digital archivist and metadata specialist: Manages metadata and the systems that provide access to digitized objects (100%)
- Head of Collection Services: Administrative role in planning for digital collections and systems (currently 20% because we are in the process of planning a new digital repository for Emory Libraries, usually this percentage would be slightly smaller)
- Processing archivist: Co-ordinating the digitization of AV collections (20%)
- University Archivist: Administrative role in planning for digital collections and systems (15%)
- Associate Director: Administrative role in planning for digital programs and projects (currently 20% because we are in the process of planning a new digital repository for Emory Libraries, usually this percentage would be slightly smaller)

- Records Manager: Managing electronic records created by the university (5%)

Information technology support for digital curation activities at the Rose Library varies by program. The born-digital program, which manages born-digital collections coming to the library as part of our manuscript collections, is supported by one of Emory Libraries' Desktop Consultants. His position typically focuses on providing desktop support to library employees, but is always willing to help us as we require support for tools and processes specific to our born- digital workflows. The challenges that we encounter often force our Desktop Consultant outside of his comfort zone somewhat and we often end up pitching in to solve problems together!

The development and maintenance of our current repository is supported by one software engineer who has time allocated in his position for this purpose. We also receive limited support from 1-2 system administrators as and when problems occur. Much of the digitization of Rose Library still images and AV materials is conducted in-house by Emory Libraries' Digitazation and Digital Curation team. In this capacity, staff from this team might also offer information technology support as and when it is required.

The Rose Library uses several discovery systems:

- Our finding aids are stored and managed within a homegrown system, Emory Finding Aids. This system is managed by the Rose Library, but stores finding aids for all of Emory's special collection libraries.
- Copies of our digitized books are accessible through Primo. Our current digital repository, which we call the Keep, provides access to our processed digitized and born-digital AV materials, although researcher access to this content is limited to the reading room. Preservation copies of our AV materials and born-digital materials are also stored in the Keep. Access to preservation copies is restricted to specific library staff members.

- We provide access to digitized still images through Luna. For the most part, these materials are not openly available online. Instead, access levels have been applied using IP authentication
- ArchiveIt is used to harvest Web content.

The Rose Library has roughly 10,000 audio and 3,000 video objects stored in the Keep (each object includes both a preservation copy and an access copy of the file). The Keep also currently holds preservation copies of approximately 500 disk images and 6,500 individual born-digital files. In total, we have between 50-60 terabytes of data stored in the Keep. Our digital assets management system, in which we keep both preservation and access copies of digitized still images, contains about 157,000 files. Files are not accessed via this system. In addition to these materials, the Rose Library also has a backlog of born-digital and audiovisual materials stored on a variety of media and awaiting further processing.

DIGITAL CURATION ACTIVITIES

The Rose Library uses the following tools and environments to support various digital curation workflows:

- Born-digital workflow: The BitCurator suite of tools, Forensic Toolkit (FTK), FTK Imager, KryoFlux, FC5025, Isobuster, Quick View Plus, BagIt
- Audiovisual workflow: Final Cut Pro for capture and codec creator, MPEG Streamclip, BagIt
- Digitized still images: Silverfast
- Digitized books, serials, newspapers, etc.: Kirtas
- Some digitization work is outsourced, including all film

GOALS FOR DIGITAL CURATION

- Continued development and implementation of a new digital repository. We are currently in the early planning stages of development for a new Samvera-based repository that will act as a common repository for all Emory Libraries content, including Rose Library content.
- Establishing shared policies and framework to guide repository development and subsequent workflows.
- Improving access and discovery tools. Establishing strategies and methods that are more robust, more sustainable, and better support research using our digital assets.

OSSArcFlow

Digital Dossier

By Megan Rohleder
Kansas State Historical Society

OVERVIEW

The Kansas Historical Society (KSHS), established in 1875, is an Executive Branch state agency. Designated as the trustee of the state in 1879 and as the repository of official government records in 1905, the Historical Society collects and preserves the story of Kansas history. In the State Archives division, there are 8 full time positions that are responsible for our agency's digital curation activities. These positions and percentage of time typically spent on digital curation are:

- State Archivist 10%
- Digital Initiatives Coordinator 90%
- Digital Archivist 90%
- Digital Specialist 30%
- Digital Assistant 70-80%
- Digital Initiatives Photographer 100%
- Electronic Records Archivist 60%
- Government Records Archivist 40%

KSHS also has an in-house, dedicated IT team available for digital curation activities. The team is comprised of a Database Administrator, a Systems Administrator, and an Application Developer. We also receive IT support from our vendors including newspapers.com, NextScan, Preservica, Archive-It, Image Science Associates and HubTack.

DIGITAL CURATION ACTIVITIES

The State Archives houses both digitized and born-digital material in the digital collection. The scope of these items reaches across our manuscripts and public records departments and includes photos, documents, maps, audio- visual material, and a large collection of digitized and born digital newspapers. The State Archives currently stores about 100 TB worth of items in a variety of storage environments. KSHS utilizes multiple systems of discovery for our digital collections.

We have five “home-grown” systems that include our archives and museum catalogs, KansasMemory.org, the newspaper database, and Dart (our internal content management system). Other points of access for our material include Ancestry.com, newspapers.com, Newsbank, Territorial Kansas Online, Civil War on the Western Border, Chronicling America, Family Search, ATLAS (Associated Topeka Libraries Automated System), and Archive-It.

Information about our digital collections is shared internally through multiple systems. Staff can update Dart and the museum catalog with notes about accessions, updates to the collection, and any other relevant information. Information about our newspaper collection is shared through the newspaper database. This database houses inventory information about both our analog and digital newspaper collections. There is a significant portion of our collection stored on servers and information about those items is kept in inventory spreadsheets.

The digital collection at the State Archives includes workflows for both born-digital and digitized material. These categories include:

DIGITIZED

- Non-Newspaper, Digitized on site
- Non-Newspaper, digitized by a partner
- Non-Newspaper, digitized by a vendor
- Newspaper, Digitized on site
- Newspaper, Digitized by partner
- Newspaper, Digitized by vendor

BORN DIGITAL

- Newspapers
- Government Records
- Manuscripts/Photos
- Archive-IT

These workflows vary in scope and complexity, but some major phases of each workflow as they currently exist include creation or acquisition, staging, description, and storage. The curation lifecycle typically ends at this point, as few long-term preservation activities are being completed on our digital collection. There have been periodic attempts to generate and check fixity information on our newspaper collection. We also record initial fixity information as part of the digitization workflow on our non-newspaper, digitized in-house items, but we do not currently complete ongoing fixity checks against those checksums.

The Historical Society uses many tools and environments to complete our digital curation workflows. These include:

- Preservica
- Dart (our own system)
- Network Drives for storage
- DIT (custom software that's part of Dart and Museum Catalog)

- FADGI compliant software- Golden Thread
- Newspaper database
- Rsync (used to acquire born-digital newspapers)
- Archive-It
- BagIt (used as part of the Chronicles in Preservation Workflow)

GOALS FOR DIGITAL CURATION

The State Archives has identified three digital curation goals that we would like to achieve.

1. We would like to develop structure to the digital curation process. This includes developing a realistic and scalable digital preservation plan. Ideally, we would have written documentation available across the agency.
2. We want to focus on a tiered approach to digital preservation, which will allow us to prioritize content and create appropriate preservation environments for each prioritized category. We would then like to actively seek out that digital content.
3. Consistent implementation of the preservation plan.

OSSArcFlow

Digital Dossier

By Kari Smith and Joe Carrano

Massachusetts Institute of Technology Archives and Special Collections

OVERVIEW

The MIT Institute Archives and Special Collections (IASC) serves as the "memory" of MIT, collecting and preserving records that document MIT's history and the people who have been a part of that history. The collections include both published and unpublished materials in various formats. This information does not reflect digital curation practices outside of IASC.

Can you indicate a percentage of time that each library, archives and curatorial staff member spends working on digital curation?

- Digital Archivist: 95%
- Institute Archivist and Program Head for Digital Archives: 50%
- Processing Archivist: 5%
- Director for Digital Preservation: 100%
- Digital Operations Coordinator: 50%

Describe information technology support available for digital curation activities:

- Most of technology support is done by the digital archivist and program head.
- As of 2017, the archives is now receiving some technology support for developing new tools as well as ongoing maintenance and tweaking of core systems (1 engineer (developer)).
- Enterprise systems support is still being determined for digital curation systems.

EDUCOPIA INSTITUTE

- We host ArchivesSpace through LYRASIS and we have an SLA with Artefactual staff who provide our IT support for our locally installed Archivematica instances.

Characterize your organization's digital infrastructure. Do you have one discovery system or several?

- We are well along with our comprehensive digital preservation storage program. Curated content is currently stored on institutionally managed network drives.
- We currently do not have an active discovery system for most archival material, but the intention is to use the ArchivesSpace public user interface.
- Some our finding aids are available online as html/pdf's on the archives subdomain of the libraries' website. A subset of these finding aids and archival publications have catalog records searchable through BARTON, our OPAC.
- Some of our digitized archival material is discoverable through DOME, the Libraries' DSpace instance.
- Most discovery is mediated through individual reference requests via email to the archives.

How does staff share documentation about collections and activities?

- Collections documentation is mainly shared through archival control files (physical and digital) and activities and processes are kept on local shared network drive folders. Some documentation also lives in places like MIT Dropbox, Google Drive, and the MIT Confluence wiki space.

What is the size/scope of your digital collections?

- 10.1 TB currently manage in local network storage, 2 PB on offline drives across 70 collections.

DIGITAL CURATION ACTIVITIES

Define digital curation at your organization? What are the major phases of your digital curation lifecycle?

- Decide to acquire digital material
- Determine Digital Preservation Requirements
- Receive Digital Content
- Ingest/Process Digital Content
- Manage Preservation Object Storage
- Make Digital Content Available

Provide an overview of the tools/environments used throughout your digital curation workflow.

- For known content we use Exactly to bag and transfer materials from donors or from in- house/vendor digitization.
- Web archiving is done with Archive-It and Webrecorder.
- Email is appraised and transferred using ePADD.
- Finding aids and accession records are in ArchivesSpace.
- BitCurator is used for unknown materials or those that require examination before accessioning. This entails additional tasks like virus checking, possible disk imaging, and identification of PII; Kryoflux if need to image floppy disks.
- During the Ingest stage we use Archivematica to identify, normalize, and package digital objects for preservation and dissemination.
- Fixity tool is used to monitor file attendance and integrity of all archival files on network drives.

Who [which roles] contributes to digital curation in your organization?

- Digital Archivist
- Institute Archivist and Program Head for Digital Archives
- Director of Digital Preservation

- Digital Content and Reformatting Team
- Scanning and reformatting QC associate
- Metadata entry assistant
- System Administrators
- Digital Library Systems Engineer

What do you consider to be your categories of digital content?

- “Categories” can correspond to format types, but in some cases categories correspond more easily to the way in which the material was accessioned.
- “Categories” here refers to content that shares a common workflow.
- Known born-digital content
- Unknown born-digital content
- Unstructured mix of known and unknown born-digital content
- Digitized content
- Content by category:
  - Video
  - Image—Digitized
  - Image—Born Digital
  - Web
  - Audio
  - Text
  - Geospatial
  - Database
  - Structured Database
  - Text—Digitized
  - Web with Video
  - Disk Image

How variable are your workflows across different categories of digital content?

- Workflows are fairly consistent for known born-digital content and digitized content. Unknown digitized content requires more work to analyze and identify information about the files for appraisal, processing, and preservation planning, such as running additional tools and reports.

GOALS FOR DIGITAL CURATION

- Successfully implement web-based transfer tool.
- Expand on the Content Types characteristics to include level of effort and capabilities for working with it and knowing the tools that will assist with arrangement and description in the next year or two.
- Implement the integration of ArchivesSpace and Archivematica.
- Implement appropriate and sufficient digital preservation storage and services.

OSSArcFlow

Digital Dossier

By Shaun Trujillo

Mount Holyoke College

OVERVIEW

Can you indicate a percentage of time that each library, archives and curatorial staff member spends working on digital curation?

- Head of Archives and Special Collections 1-5%
- Special Collections Archivist ~5-10%
- Metadata Librarian >40%
- Digital Projects Lead >20%
- Digital Library Applications Manager >40%
- Associate Director of Discovery and Access 1-5%

Describe information technology support available for digital curation activities.

MHC's Library, Information, & Technology Services (LITS) is a merged organization (Library + IT), which has proven to be a very productive environment for quickly prototyping and testing of technology services. We have direct access to a systems and networking associate, and can use roughly 5 hours of their time weekly. The output from our digitization program has grown at a steady rate through the years and IT support has consistently met storage and backup needs.

Characterize your organization's digital infrastructure. Do you have one discovery system or several?

Several - Islandora, the OPAC (Aleph, Ebsco Discovery), 5 college consortial finding aid site

How does staff share documentation about collections and activities?
Electronic finding aids, Islandora digital repository, library website

What is the size/scope of your digital collections?
Roughly 20,000 locally digitized objects. ~500 outsourced digitized audio/video assets. ~200 electronic records accessions. ~10 media accessions

DIGITAL CURATION ACTIVITIES

Define digital curation at your organization. What are the major phases of your digital curation lifecycle?
Primarily creation/digitization/capture, accession (for born digital records and media), description, storage, derivative creation (automated), publication

Provide an overview of the tools/environments used throughout your digital curation workflow.

- ArchivesSpace (Archives' collections management system - previously Archivists' Toolkit - full migration to AS in Winter 2016-17)
- ResourceSpace (Digital asset management syste - home of master files from digitization workflows)
- Islandora (Digital collections frontend - previously CONTENTdm)
- Omeka (Archives led digital exhibitions and crowdsource transcription site)
- DSpace (Institutional archive for student and faculty generated data, as well as college wide programs and initiatives)
- Archive-It (Web archiving service for snapshotting the college websites. Strategic capture of sites based on Archives' collecting policy)
- Network Attached Storage (Acting repository of record for most of the original work produced by the Library and Archives. Currently houses ~20TB of data)

Who [which roles] contributes to digital curation in your organization?
Head of Archives and Special Collections, Special Collections Archivist, Metadata Librarian, Digital Projects Lead, Digital Library Applications Manager, Associate Director of Discovery and Access

What do you consider to be your categories of digital content?
ON-DEMAND DIGITIZATION:

- This is the most fully realized workflow for the creation, publishing, and longterm storage of digital objects. Workflow is centered around the ResourceSpace system's request and delivery features. Most of the products of this workflow are described and managed in ArchivesSpace.

PROGRAMMATIC DIGITIZATION:

- Large-scale scanning/digitization of archival collections.
- Challenges around level of description - Best practices of aggregate description vs. item level description. Input/output of description for ArchivesSpace objects has been a pain point, but we’ve gleaned some ideas from Duke’s python scripts on Github.

BORN DIGITAL E-RECORDS:

- Current workflow was born from a 2009 National Historical Publications and Records Commission (NHPRC) grant and relies on a locally developed Digital Records Transfer system (DRXFER) and the Data Accessioner to produce SIPs and basic DIPs.

BORN DIGITAL A/V EVENT RECORDINGS:

- A steady growth of new A/V content each semester provides storage and throughput challenges, as well as questions regarding access and permissions.

BORN DIGITAL STORAGE MEDIA ACCESSIONS:

- Rare donations of storage media from staff and community. These assets provide similar issues as our born digital A/V materials. We rely on FTK Imager to produce E01 and raw disk images and Bitcurator for file system manifests and reporting.

EMAIL ACCESSIONING:

- We collect the MBox files of outgoing officers. We're interested in incorporating the epadd tool for help in processing email archives.

How variable are your workflows across different categories of digital content?

The workflows across and even within categories are fairly variable and in some instances ad hoc. Process can change depending on the context of the project. Overall, it is fair to say that each of the categories listed above has an independent/unique workflow. The two categories that are closest in terms of workflow are ‘on-demand digitization’ and ‘programmatic digitization’.

GOALS FOR DIGITAL CURATION

- Providing better access points for digital content while maintaining archival context, i.e. better linkages between finding aids and collection guides and digital library frontend systems.
- Streamline A/V, media, and other born digital accessioning. Automation, automation, automation! Low cost, low staff processing of digital materials.
- Standardization and clarification of digital curation practice, e.g. standard and consistent definition of an AIP for long-term retention, formalize handoffs and syncing of data across systems, providing audit trails and improving availability in collections system of record.

OSSArcFlow

Digital Dossier

By Susan Malsbury and Nick Krabbenhoeft
New York Public Library

OVERVIEW

The New York Public Library includes three research libraries that collect archival material: the Humanities and Social Science Research Divisions (within the Stephen A. Schwarzman building), the Library for Performing Arts at Lincoln Center, and the Schomburg Center for Research in Black Culture. NYPL has a centralized processing department called Special Collections and Preservation Services which provides technical and support services for all three research libraries. This dossier will give background on digital curation at NYPL but primarily discuss the work performed by the Digital Archives and Digital Preservation Programs which are situated within Special Collections and Preservation Services.

DIGITAL CURATION ACTIVITIES

Historically, NYPL has not operated from a comprehensive digital curation strategy but has primarily focused on adding digital capabilities and integrating these processes within the general workflows of the library with a focus on making the files accessible to researchers. For example, in the past decade, the program responsible for transferring analog sound and video, the Preservation of Audio and Moving Image (PAMI) program, has transitioned from tape-to-tape to tape-to-file workflows.

EDUCOPIA
INSTITUTE

The digital outputs from these programs are cataloged by staff in the Special Formats Processing Program (SFP) and made available through a digital platform in the reading rooms with assistance from collection librarians, as they were prior to digitization of the workflow. By this broad definition, the majority of Research Libraries staff are involved in digital curation activities.

NYPL currently manages the following streams of digital materials:

- Born-Digital Archives—archival materials collected on physical digital media and hosted cloud storage
- Born-Digital Original Documentation—performance recordings commissioned by the library
- Digital Imaging—still images of books, photographs, and other research material
- Microfilm Mass Digitization—still images of microfilmed materials
- Audio and Moving Image Digitization—archival materials collected on audio and video media formats

In general, the workflow for these streams has the following phases:

1. **Acquisition**: Materials is either acquired in a native digital format (born-digital archives, born-digital original documentation) or a digital surrogate is created from physical materials (digital imaging, microfilm mass digitization, audio and moving image digitization). In the second case, this is accomplished both by vendors and in-house staff.
2. **Ingest**: Acquired material is processed by in-house staff who perform quality assurance, create packages for long-term preservation, and describe the materials (if a description does not already exist).
3. **Storage**: Material is placed in a managed storage environment where it is monitored for preservation risks and available for access reuqests.
4. **Data Management**: Descriptive and other metadata is loaded into a system to enable management of stored materials and provide metadata for discovery systems.

5. **Access**: Library users access descriptions of digital content through catalog records or finding aids. Service copies of still image, audio, and moving image content are made available through a digital content platform.

Each stream has grown organically and may share systems with other streams depending on the phase. The following discusses the systems for the born-digital archives stream and notes where other streams intersect with born-digital archives on a system.

Digital Curation of Born-Digital Archives

Two programs take an active role in managing born-digital archives. The Digital Archives Program was established in 2011 and is part of the Archives Unit and consists of a digital archivist, digital archives assistant, and library technical assistant, all who are FTE who spend 100% of their time on digital curation activities. The Digital Preservation Program was established in 2015 and consists of 1 FTE who spends 100% of his time on digital curation activities.

Acquisition

During the collection development stage, the Digital Archivist uses site visits to collect information for pre-acquisition scoping decisions. Information includes the size of the collection, file types and hardware in the collection, arrangement and file naming conventions, and the digital environments used to create files. The collected information is used to anticipate resources needed to ingest and process the collection, and for informing the arrangement and description. In terms of transferring the materials, Digital Archives has developed several strategies depending on the size and complexity of the materials. Once in the custody of New York Public Library, these materials, alongside the collected metadata and collection documentation, comprise the SIP.

Ingest

At the beginning of Ingest, administrative and physical control are established through the accessioning process to ensure all the Library has received the agreed upon material. All media is given a unique identifier; inventoried in a custom FileMaker Pro database, also used for managing physical archival material; and all physical media is photographed. A determination is made as to whether the physical media is archival or transfer media. Media is considered archival when the media object contains working files generated or edited by the creator during their professional and personal activities. This includes computers and physical media objects (floppy disks, CD/DVDs, zip disks, external hard drives) that contain evidential information regarding the creation and activity surrounding the files. Media is considered transfer media when files have been added to the media purely to provide a method of transport or storage by the creator/donor. Archival media is put through the disk imaging workflow where disk images are created on one of two forensic workstations and transfer media goes through the file transfer workflow. For email archives, staff convert the materials to the mbox format for processing, while retaining the original email archive.

Once the Digital Archives program has staged the materials for arrangement and description, processing archivists appraise and arrange the material using Forensic Toolkit (FTK) on the FRED computer. Ideally the same archivist will process the born-digital and paper portions of the collection concurrently. NYPL has only recently begun processing email accounts and this is done in ePADD. All archival description is entered into ArchivesSpace. The end result of arrangement and descriptions is one or more AIPs consisting of the arranged files, descriptive metadata, and technical metadata that is then passed to the Data Management and Storage phases.

Data Management

New York Public Library manages archival descriptions in an ArchivesSpace instance maintained by the Metadata Archivist. Patrons do not have direct access to this instance. Instead, descriptive records are exported from ArchivesSpace as XML finding aids and MARC catalog records and published on NYPL's Archives Portal, Catalog, and Digital Collections platforms. These platforms reference the extent and content of any born-digital archival material in the collection, but access must be requested in-person in a reading room.

Storage

After processing and packaging, the archival materials are uploaded via an Archivematica pipeline managed by Digital Archives and Digital Preservation staff to a library-owned storage environment. The 3PB Isilon system is also used to store materials from all other digital curation streams, although still images, audio, and moving images materials are not transferred using Archivematica, but custom processes developed according to their package specifications.

Access

In the past, access to materials in published finding aids was provided on patron requests using air-gapped terminals. To support this, the Digital Archivist delivered copies of digital collections on external hard drives and maintained instructions for how reading room staff could copy materials to terminals on request. Quickview Plus and a number of emulators were installed on these terminals to facilitate access. However, this pilot project has been unable to scale with the increasing level of born-digital collecting. New mechanisms are being evaluated.

Digital Infrastructure Support
Organizationally, the digital infrastructure used in digital curation activities is maintained by three programs. First, the Information Technology Group is responsible for the installation and maintenance of capital infrastructure, including network storage, servers for running programs such as Archivematica, and networking to transport materials. Second, the Digital Department is responsible for the creation and maintenance of applications such as the library's discovery interface and digital media platform. Finally, individual programs are responsible for software with particularly small user bases or use cases.

This activity may be outsourced (e.g., a support contract for the FRED) or kept in-house (e.g., custom shell scripts to automate the packaging of files during processing). Programs are also responsible for documenting their systems. Most do so in an open manner using the systems supported by ITG, including Google Documents (Digital Archives), Github wikis (Preservation of Audio and Moving Images), Confluence (Digital Imaging Unit).

An ongoing effort is underway to improve the sustainability of programs by sharing resources where possible. One example of this is moving the primary storage for digital archives to the same network storage as other materials. This consolidation should decrease maintenance overhead and better prepare the Library for terabyte-scale acquisitions.

GOALS FOR DIGITAL CURATION

The central challenge facing NYPL's digital curation activities is coordination. The organic growth of each of its streams of digital collections allowed each stream to flourish; however, differing access to support has affected the Library's ability to successfully scale to accommodate increasingly larger and more complex born-digital acquisitions and robust digitization initiatives.

With that in mind, NYPL has the following digital curation goals:

1. Access platforms capable of providing acccess to all digital collections.
    - A pathway for access has to be developed for all streams listed in the Digital Curation at NYPL section above.
    - Access platforms have to accommodate multiple levels of access from Digital Collections on the Library's website, which are freely available to the general public, to managed collections that may only be viewed in a specific reading room after consulting with a Library staff member.
    - A wide variety of digital material will need to be accommodated from common formats to material that can only be rendered in an emulation environment.
2. Network and storage infrastructure to support the continued growth of digital collections.
3. Improved digital literacy among staff to promote the use of digital collections.

OSSArcFlow

Digital Dossier

By Rebecca Russell
Woodson Research Center, Rice University

OVERVIEW

Woodson Research Center (WRC) is the Special Collections and University Archives for Rice University. Rice University is a private research university with an undergraduate focus located in Houston, Texas.

There are 5 professional Archivists in the department. 4 staff members work directly with digital preservation efforts, and have SAA-DAS certification. As a small department, we do not have dedicated curatorial foci, but do 'a bit of everything' including digital preservation. We typically spend about 20% of our time working on digital curation. A staff member could easily to do this full-time.

Our institutional culture shapes our response to digital preservation. At Rice, it is very DIY and therefore we tend to use open-source tools and there is tolerance for staff time to learn and work with the DIY tools. Administration were supportive of staff taking time to become DAS trained, and there is enough staff experience to learn and document software steps, including command line interfaces.

Digital Scholarship Services (DSS) provide advisory services, guidance and review of the curatorial activities of the Woodson Research Center (including review of metadata, review of the ingest of collections, and monitoring collections within the preservation system.)

EDUCOPIA
INSTITUTE

Fondren IT provides technical support by consulting with the WRC and DSS in acquiring hardware, software, and cloud storage space to manage digital collections. Our IT staff are for desktop support, not technical help with our digital preservation.

Digital Infrastructure

We have robust local and cloud storage environments, DSpace is our primary public discovery system. WRC uses an iterative process for worfklows, using Google Docs as location for the initial centralized access/editing of our workflows and policies.

DIGITAL CURATION ACTIVITIES

The purpose of our Digital Preservation Policy is to establish a long-term digital preservation solution for our institution that will assure accessibility to special collections and unique resources.

- In accordance with our primary mission to support the institutional, research, and public service programs of the University, WRC plans to assure the long term access of our collections by continuing the digital preservation program which was developed in 2014.
- https://digitalriceprojects.pbworks.com/w/page/44763477/Digital%20Preservatio n%20Support%20at%20Fondren%20Library

Tools/Environment

- ArchivesSpace (collection management system - previously Archivists' Toolkit - full migration to AS in fall 2014)
- DSpace - to manage publicly available digital objects, Duracloud backup
- Google spreadsheet, tracking all AIPs (where they are stored, their size, their basic formats, where their hash value logs are)
- BitCurator (legacy media)

- DROID and Exiftool to gather key technical and possibly descriptive info
  - Droid reports give fixity, MIMEtype, file format name, file format version, PUIDsoftwareModified date (but not Date created or Accessed), file name and size
  - Exiftool (Command Line interface) reports give MAC times (context) - sometimes include created, includes Modified date, plus any embedded descriptive metadata
- Quickhash - to create and compare hash values over time (nearline) - investigating Fixity as a replacement for this tool
- BWF metadedit for audio files
- Handbrake for video files
- CERP (obsolete) investigating ePADD software 2018 for email preservation

Storage Environments

- DSpace (Institutional archive, publicly available digital objects) managed by 2 programmers
- DDN2 -- Nearline server storage of AIPS, local storage
- Amazon Glacier -- off-site storage of AIPs
- Offline -- not stored online (large files-such as uncompressed video, not stored in DSpace or DDN2) AIPs stored on 2 local external hard drives

Digital Curation Lifecycle (Major Phases)

- Digital objects → Ingest/Appraisal/Analysis (ArchivesSpace → Data Management (Digital object spreadsheet, ArchivesSpace - assign administrative, descriptive, technical, structural and preservation metadata) → Storage (either in public system or nearline system, external hard drive farm) → Finding Aid → Access (either in public system or nearline system)

Categories of Digital Content

- Publicly available materials
  - No copyright or privacy problems
  - Item level, generally
  - Workflow on public wiki
  - Least hands-on for our staff, least complicated, least documented
- Nearline/non-public materials
  - Has copyright or privacy concerns
  - Groups of files, or could be item level
  - Most hands-on, most complicated, most documented workflows

The category of materials determine the workflow. We use the same principles for each category:

- Rule of 3 for storage locations (3 copies, 2 formats, 1 offsite)
- Storage format must meet standards for AIP
- Creation, access, tracking - reasonable workflow for our staff and well documented
- Cost - must fit our budget of $ and time

Recent Digital Preservation Activities at Rice

- We set DP goals annually and update our Digital Preservation Policy online

GOALS FOR DIGITAL CURATION

- Audio and video digitization, packaging and storage: KTRU (Rice's on-campus, student run radio station), Shepherd Music School digitization, oral histories
- At end of October, we hired a 6-month temporary DP staff member to help us with our legacy media processing backlog and contribute to the OSSArcFlow project.
- Darchive- Investigating a "dark" DSpace instance to manage our nearline materials.

OSSArcFlow

Digital Dossier

By Michael Olson and Glynn Edwards
Stanford University Libraries

OVERVIEW

Stanford University hosts twenty-three libraries of which nineteen are under the direction of the University Librarian. For the purposes of this document the following libraries are not included in this dossier: Hoover Institution Library and Archives, Lane Medical Library, Crown Law Library, and the SLAC National Linear Accelerator Laboratory Research Library.

All of the nineteen libraries included in this dossier acquire digital resources. These digital acquisitions consist of purchased bibliographic / serial content and born-digital content. The majority of the born-digital content is acquired by through our Department of Special Collections, the Archive of Recorded Sound, and the David Rumsey Map Center. Selection of digital resources for acquisition is accomplished by over twenty bibliographers / curators that act as subject selectors for specific study areas. The total size of our holdings for the Department of Special Collections (includes University Archives), the Archive of Recorded Sounds and the David Rumsey Map Center are difficult to estimate.

EDUCOPIA INSTITUTE

DIGITAL CURATION ACTIVITIES

Our digital curation activities follow two distinct workflows:

- Stanford faculty, research groups, librarians, and staff members of the acquisition department are able to self-deposit to the Stanford Digital Repository through our online self-deposit service at https://sdr.stanford.edu/. Note that all data deposited to the Stanford Digital Repository must be classified as either Low or Medium risk based on data classification guidelines at https://uit.stanford.edu/guide/riskclassifications.
- Special Collections libraries in the Stanford Library System (Department of Special Collections, Archive of Recorded Sound, David Rumsey Map Center) acquire born-digital resources that require specialize archival handling, description, and discovery conditions that fall outside of our workflows for traditional analogue library materials.

Staff

- Approximately 20 curators / bibliographers spend a percentage of their time selecting born- digital resources for acquisition. This includes working with donors and library leadership on deeds of gift / purchase agreements, communicating with donors on access rights and delivery methods.
- 1 Full-time digital archivist (divided equally between project management for ePADD in Special Collections and doing day to day technical lab tasks for Digital Library Systems and Services). Responsible for conducting survey of donors of born-digital content, creating disk images from born-digital media, acquiring born-digital resources from cloud based services, training processing archivists how to arrange and describe born-digital collections, transfer of content to born-digital servers.

One instance hosts the HBCU Library Alliance Digital Collection, a collection of primary resources from 23 HBCU libraries and archives that is comprised of over 16,000 images. The other CONTENTdm instance houses 60,000 digitized images from the Morehouse College Martin Luther King, Jr. Collection, that are available only in the reading room of the Archives Research Center.

Our digital exhibits also include digitized archival materials, and these are hosted on an instance of Omeka. Currently, we have four digital exhibits drawing from multiple archival collections. We are also in the process of adding an exhibit based on digitized materials from the Spreading the Word grant to be published in Spring 2018.

In addition to these systems, AUC Woodruff Library uses ArchivesSpace to create and maintain all finding aids for archival collections. ArchivesSpace serves as the backend; the front end is provided by XTF, and many digital objects are linked within the finding aids.   AUC Woodruff is currently exploring adopting the ArchivesSpace public interface as updates make that more feasible.

Currently, our digital collections material exceeds 40 TB, including both master and derivative files. Large, digitized AV files and grant funded content are backed up in the cloud via an Amazon Snowball into Amazon Glacier. Our main categories of digital content are digitized archival materials, born-digital archival materials, digitized scholarly communication from member institutions, born-digital scholarly content from member institutions, and born-digital institutional records and photographs. The AUC Woodruff Library has been involved in digital preservation activities since 2010 when it joined the MetaArchive Cooperative on behalf of the HBCU Library Alliance.

- A search for a new 100 % FTE digital archivist is underway and this new staff member will work primarily on born-digital materials for the Department of Special Collections.
- 1 metadata librarian, ~ 15% FTE.
- 1 technical service manager, 40 % FTE. Responsible for lab hardware and software purchases and steering of technical aspects of our Born-Digital / Forensics lab.
- 1 manuscript librarian, 20 % FTE. Provides program oversight and guidance for all born- digital acquisitions, overseas all born-digital processing staff.
- 1 System Administrator, 20% FTE. Responsible for specifying and maintaining 75 TB server for high risk data (also referred to as our BDFL server), increasing disk resources for server, patching and security monitoring of server. This system administrator is also responsible for maintaining a dedicated workstation for born-digital content that is only available on workstations in the reading room(s).
- x number of processing archivists. Processing archivists are trained to process analogue and born-digital materials. Note that this staffing is usually soft funded and collection specific and most born-digital collections remain minimally processed pending additional funding.

Technical Support

- Stanford Libraries maintains two born-digital / forensics labs (BDFLs) for the processing of born- digital content. The BDFL labs are one of the three digitization services that are run by Stanford Libraries. The other service arms are the Digital Production Group (responsible for digital photography and 3-D imaging) and the Stanford Media Preservation Lab (responsible for digitization of analogue audio and video). There is some overlap between the services offered by the Born-Digital / Forensics Lab and the Stanford Media Preservation Lab with the growing number of born-digital audio/video acquisitions.

- The Born-Digital / Forensics Labs maintain FREDs, custom built capture stations, commodity workstations, suites of portable write-blockers, and a large and growing collection of legacy computers, computer hardware and software.
- Unrestricted born-digital materials can be published via our online discovery environment SearchWorks. When possible PURLs of born-digital collection materials are available online using Spotlight, our online digital exhibits platform. For material that cannot be viewable online, access is only available via computer workstations in the reading room(s). This content is stored on a dedicated server that is only available from locked down workstations in the reading room(s).

Digital Curation Lifecycle

All of our born-digital acquisitions undergo minimal processing that includes forensic and or logical disk imaging using either FTK imager or BitCurator tools, scanning for PII using either Forensic Toolkit or BitCurator Bulk Extractor, transfer of files to an encrypted and server for high risk data and fixity verification. PII reports for each collection are maintained in ArchivesSpace and on with the collection files. Collections that are allocated resources to undergo additional processing are processed using Forensic Toolkit. There are however a few areas in our workflows that are worth describing as they may differ from our peer institutions.

The first is that that Stanford Libraries has twenty or more subject selectors or curators that actively acquire collections. This provides unique challenges in the volume of acquired materials, in prioritization and resource allocation. A second unique challenge is presented by the presence of high risk data in many of our born-digital acquisitions. All of our born-digital collections are screened for PII using either Forensic Toolkit or Bulk Extractor and all collections are treated as high risk as defined by Stanford's Internet Security Office.

This has necessitated the creation of a storage solution that is separate from our institutional repository. All collections are considered to contain PII and are assumed as high risk until evaluated by an archivist. Our institutional repository (SDR) is only used for processed collections that have undergone human evaluation of PII reports.

GOALS FOR DIGITAL CURATION

- Script and deploy Bulk Extractor to run against all born-digital collections that are stored on our server for high risk data. This is currently initiated on a collection by collection basis by our digital archivist. This should be an automated scan that initiated as soon as born-digital content is transferred to our encrypted server.
- We are currently undergoing a pilot to use BitCurator to generate machine actionable reports. Our goal is to use these reports for collections that can be ingested into our digital repository to populate descriptive and technical metadata streams.