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.