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.